วารสารวิชาการพระจอมเกล้าพระนครเหนือ ปีที่ 32, ฉบับที่ 2 เม.ย.–มิ.ย. 2565
The Journal of KMUTNB., Vol. 32, No. 2, Apr.–Jun. 2022

บทความวิจัย

การเปรียบเทียบประสิทธิภาพวิธีการประมาณค่าสูญหายของตัวแปรตอบสนองสำหรับแผนแบบแฟกทอเรียล 3 ปัจจัยในบล็อกสมบูรณ์เชิงสุ่ม

ธัญรดา ชัยขจรวัฒน์* จุฑาภรณ์ สินสมบูรณ์ทอง และ ธิดาพร ศุภภากร
ภาควิชาสถิติ คณะวิทยาศาสตร์ มหาวิทยาลัยเกษตรศาสตร์

* ผู้นิพนธ์ประสานงาน โทรศัพท์ 08 6476 6646 อีเมล: rada.hua@gmail.com DOI: 10.14416/j.kmutnb.2021.05.040
รับเมื่อ 24 เมษายน 2563 แก้ไขเมื่อ 22 มิถุนายน 2563 ตอบรับเมื่อ 10 สิงหาคม 2563 เผยแพร่ออนไลน์ 27 พฤษภาคม 2564

บทคัดย่อ

การวิจัยครั้งนี้มีวัตถุประสงค์เพื่อเปรียบเทียบประสิทธิภาพวิธีการประมาณค่าสูญหายของตัวแปรตอบสนองสำหรับแผนแบบแฟกทอเรียล 3 ปัจจัยในบล็อกสมบูรณ์เชิงสุ่ม 4 วิธี คือ วิธีค่าคาดหวังสูงสุด (Expectation Maximization) วิธีค่าทดแทนพหุ 1 (Multiple Imputation 1) วิธีค่าทดแทนพหุ 2 (Multiple Imputation 2) และวิธีเคเนียร์เรสเนเบอร์อิมพิวเทชัน (K-Nearest Neighbor Imputation) ซึ่งวิธีค่าทดแทนพหุ 1 และวิธีค่าทดแทนพหุ 2 จะแตกต่างกันที่วิธีการที่นำมาใช้ในการคำนวณ ทั้งนี้จำลองข้อมูลด้วยเทคนิคมอนติคาร์โล จำนวน 108 สถานการณ์ และทำการทดลองซ้ำในแต่ละสถานการณ์ 2,000 รอบ กำหนดให้แต่ละปัจจัยมีจำนวน 3, 4 และ 5 ระดับ และมีจำนวนบล็อกเท่ากับ 3 บล็อก ข้อมูลมีการสูญหายแบบสุ่ม ร้อยละการสูญหายของข้อมูลเท่ากับ 5 และ 10 และความแปรปรวนของค่าสังเกตเท่ากับ 25 และ 625 โดยเกณฑ์ที่ใช้ในการเปรียบเทียบประสิทธิภาพ คือ ค่าประมาณความคลาดเคลื่อนกำลังสองเฉลี่ย จากการศึกษาพบว่า วิธีเคเนียร์เรสเนเบอร์อิมพิวเทชันให้ค่าประมาณความคลาดเคลื่อนกำลังสองเฉลี่ยต่ำที่สุดในทุกสถานการณ์ที่ทำการศึกษา ดังนั้นวิธีเคเนียร์เรสเนเบอร์อิมพิวเทชันให้ประสิทธิภาพสูงสุดในทุกสถานการณ์ที่ทำการศึกษา

คำสำคัญ: แผนแบบแฟกทอเรียล 3 ปัจจัยในบล็อกสมบูรณ์เชิงสุ่ม ค่าสูญหาย วิธีค่าคาดหวังสูงสุด วิธีค่าทดแทนพหุ วิธีเคเนียร์เรสเนเบอร์อิมพิวเทชัน

การอ้างอิงบทความ: ธัญรดา ชัยขจรวัฒน์ จุฑาภรณ์ สินสมบูรณ์ทอง และ ธิดาพร ศุภภากร, "การเปรียบเทียบประสิทธิภาพวิธีการประมาณค่าสูญหายของตัวแปรตอบสนองสำหรับแผนแบบแฟกทอเรียล 3 ปัจจัยในบล็อกสมบูรณ์เชิงสุ่ม," *วารสารวิชาการพระจอมเกล้าพระนครเหนือ*, ปีที่ 32, ฉบับที่ 2, หน้า 434–444, เม.ย.–มิ.ย. 2565.

Research Article

Efficiency Comparison of Missing Value Estimation Methods of Response Variable for Three Factor Factorial Experiment in Randomized Complete Block Design

Thanrada Chaikajonwat*, Juthaphorn Sinsomboonthong and Thidaporn Supapakorn
Department of Statistics, Faculty of Science, Kasetsart University, Bangkok, Thailand

* Corresponding Author, Tel. 08 6476 6646, E–mail: rada.hua@gmail.com
DOI: 10.14416/j.kmutnb.2021.05.040
Received 24 April 2020; Revised 22 June 2020; Accepted 10 August 2020; Published online: 27 May 2021

Abstract

The objective of this research is to compare the efficiency of four missing value estimation methods; i.e. Expectation Maximization, Multiple Imputation 1, Multiple Imputation 2, and K-Nearest Neighbor Imputation. The response variables of three factor factorial experiment were tested in randomized complete block design. The difference between Multiple Imputation 1 and Multiple Imputation 2 is the distance calculation methods of observations. A simulation study is conducted by Monte Carlo technique for 108 situations and 2,000 replications for each situation. The studied points are as follows : the numbers of each factor are 3, 4 and 5, the number of block is 3 with the percentages of missing values at 5 and 10, and the studied variances of observation are 25 and 625. In addition, the efficiency comparison criterion is the estimated mean squared error. The result shows that K-Nearest Neighbor Imputation has the lowest estimated mean squared error for all situations. Therefore, K-Nearest Neighbor Imputation is the most efficient estimator for all situations.

Keywords: Three Factor Factorial Experiment in Randomized Complete Block Design, Missing Value, Expectation Maximization, Multiple Imputation, K-Nearest Neighbor Imputation

Please cite this article as: T. Chaikajonwat, J. Sinsomboonthong, and T. Supapakorn, "Efficiency comparison of missing value estimation methods of response variable for three factor factorial experiment in randomized complete block design," *The Journal of KMUTNB*, vol. 32, no. 2, pp. 434–444, Apr.–Jun. 2022 (in Thai).

1. บทนำ

แผนการทดลองในทางสถิตินั้นมีหลายแบบ เช่น แผนแบบสุ่มสมบูรณ์ (Completely Randomized Design; CRD) แผนแบบบล็อกสมบูรณ์เชิงสุ่ม (Randomized Complete Block Design; RCBD) แผนแบบจัตุรัสละติน (Latin Square Design; LSD) แผนแบบแฟกทอเรียล (Factorial Design) แผนแบบซ้อนใน (Nested Design) และแผนแบบสปลิตพล็อต (Split-Plot Design) แต่ละแผนการทดลองจะเหมาะสมกับข้อมูลที่แตกต่างกันไป ซึ่งการวางแผนการทดลองเป็นสิ่งที่สำคัญเป็นอย่างยิ่งในการทำการทดลอง โดยการวางแผนการทดลองถูกนำมาใช้อย่างกว้างขวางในด้านต่างๆ เช่น ด้านเกษตรกรรม ด้านสังคมศาสตร์ ด้านการแพทย์ ด้านอุตสาหกรรม ในงานวิจัยนี้ผู้วิจัยสนใจศึกษาแผนแบบแฟกทอเรียล 3 ปัจจัยในบล็อกสมบูรณ์เชิงสุ่ม เนื่องจากเป็นแผนแบบที่มีการวิเคราะห์ข้อมูลแบบสองทาง ทำให้สามารถสรุปผลการทดสอบสมมติฐานเกี่ยวกับปัจจัยทั้ง 3 ปัจจัยที่สนใจศึกษาพร้อมๆ กันได้ และเนื่องจากแผนแบบแฟกทอเรียล 3 ปัจจัยที่มีการทดลองแบบสุ่มสมบูรณ์ต้องใช้หน่วยทดลองที่มีลักษณะเหมือนกันมาทำการทดลองให้ได้ครบทุกๆ หมู่ทรีตเมนต์ (Treatment Combination) แต่ในบางครั้งในทางปฏิบัติหน่วยทดลองที่จะนำมาเป็นหน่วยทดลองนั้นเป็นสิ่งที่หาได้ยาก หรือไม่สามารถหาหน่วยทดลองที่มีลักษณะเหมือนกันมาทำการทดลองได้ครบทุกๆ หมู่ทรีตเมนต์ ดังนั้นแผนแบบแฟกทอเรียล 3 ปัจจัยในบล็อกสมบูรณ์เชิงสุ่มจึงเป็นอีกหนึ่งทางเลือกในการทำการทดลอง เพราะหน่วยทดลองจะถูกจำแนกตามลักษณะใดลักษณะหนึ่งออกเป็นกลุ่มที่เรียกว่า “บล็อก” [1] และในขั้นตอนการเก็บรวบรวมข้อมูลที่ได้จากการทดลองบางครั้งอาจไม่ได้ข้อมูลครบตามจำนวนที่ผู้ทดลองต้องการศึกษาหรือข้อมูลเกิดการสูญหาย อาจเนื่องมาจากหน่วยทดลองเสียชีวิตหรือตายระหว่างที่ทำการทดลอง ทำให้เก็บรวบรวมข้อมูลได้ไม่ครบถ้วนตามแผนการทดลองที่ออกแบบไว้ ถ้าหากใช้ข้อมูลเท่าที่เก็บรวบรวมมาได้อาจไม่เพียงพอต่อการวิเคราะห์ ซึ่งจะส่งผลทำให้การสรุปผลคลาดเคลื่อนได้และเนื่องจากในแต่ละการทดลองจะพบลักษณะของข้อมูลสูญหายต่างกัน Little และ Rubin [2] จึงแบ่งกลไลการสูญหายของข้อมูลออกเป็น 3 แบบ ได้แก่ การสูญหายแบบสุ่ม (Missing at Random; MAR) การสูญหายแบบสุ่มอย่างสมบูรณ์ (Missing Completely at Random; MCAR) และการสูญหายแบบไม่สุ่ม (Not Missing at Random; NMAR) โดยจะเรียกค่าของข้อมูลที่สูญหายไปว่า “ค่าสูญหาย” ทั้งนี้การสูญหายของข้อมูลอาจเกิดได้จากหลายสาเหตุ สำหรับบางการทดลองไม่สามารถทำการทดลองใหม่ได้ เนื่องจากมีข้อจำกัดบางอย่าง เช่น ระยะเวลาที่มีอยู่อย่างจำกัด ทรัพยากรที่มีอยู่อย่างจำกัด และทุนที่มีอยู่อย่างจำกัด เพราะฉะนั้นจึงต้องดำเนินการประมาณค่าสูญหาย เพื่อนำค่าประมาณที่ได้มาใช้ทดแทนค่าของข้อมูลที่สูญหายไป จากงานวิจัยที่ผ่านมามีผู้ศึกษาและเปรียบเทียบวิธีการประมาณค่าสูญหายในแผนแบบการทดลองต่างๆ กันดังรายละเอียดต่อไปนี้

ประพจน์ [3] ศึกษาและเปรียบเทียบประสิทธิภาพวิธีการประมาณค่าสูญหายในการวางแผนแบบบล็อกสมบูรณ์เชิงสุ่ม 3 วิธี ได้แก่ วิธีกำลังสองน้อยสุด (Least Squares Method) วิธีค่าคาดหวังสูงสุด (EM) และวิธีการทดแทน (Imputation) พบว่า วิธีการทดแทนให้ค่าเฉลี่ยกำลังสองของความคลาดเคลื่อน (MSE) ต่ำกว่าวิธีค่าคาดหวังสูงสุด และวิธีกำลังสองน้อยสุด ในทุกสถานการณ์ของการทดลองที่ทำการศึกษา

ศุภลักษณ์ [4] ศึกษาและเปรียบเทียบประสิทธิภาพวิธีการประมาณค่าสูญหายในการวางแผนแบบจัตุรัสละติน 3 วิธี คือ วิธีกำลังสองน้อยสุด วิธีค่าคาดหวังสูงสุด และวิธีค่าทดแทนพหุ (MI) ซึ่งพบว่ากรณีที่ร้อยละของข้อมูลสูญหายและสัมประสิทธิ์การแปรผันมีค่ามากควรเลือกใช้วิธีค่าทดแทนพหุ ในการประมาณค่าสูญหาย แต่สำหรับกรณีที่ร้อยละของข้อมูลสูญหายและสัมประสิทธิ์การแปรผันมีค่าน้อยพบว่า ค่าความคลาดเคลื่อนสัมบูรณ์สูงสุดของทั้ง 3 วิธี มีค่าใกล้เคียงกันมาก ดังนั้นจึงควรเลือกใช้วิธีกำลังสองน้อยสุดในการประมาณค่าสูญหาย เนื่องจากสะดวกและรวดเร็วกว่า

วิสุทธิดา [5] ศึกษาและเปรียบเทียบประสิทธิภาพวิธีการประมาณค่าสูญหาย 3 วิธี คือ วิธีการวนซ้ำ (Iterative) วิธีของ

วิลคินซัล (Wilkinson) และวิธีเคเนียร์เรสเนเบอร์อิมพิวเทชัน (K-Nearest Neighbor Imputation; KNN) ในแผนแบบแฟกทอเรียล 2 ปัจจัยในบล็อกสมบูรณ์เชิงสุ่ม ซึ่งจากการศึกษาพบว่า ควรเลือกใช้การประมาณค่าสูญหายด้วยวิธีเคเนียร์เรสเนเบอร์อิมพิวเทชัน

Lovelyn [6] ศึกษา และเปรียบเทียบประสิทธิภาพวิธีการประมาณค่าสูญหายด้วยวิธีการตัดข้อมูลแบบแพร์ไวส์ (PD) วิธีค่าทดแทนพหุและวิธีค่าคาดหวังสูงสุด ในแผนแบบแฟกทอเรียลในบล็อกสมบูรณ์เชิงสุ่มที่มีการทำซ้ำ ซึ่งจากผลการศึกษาพบว่า วิธีค่าคาดหวังสูงสุดให้ประสิทธิภาพสูงสุด เพราะค่าเฉลี่ย (Mean) ค่าคลาดเคลื่อนมาตรฐาน (Standard Error) และค่าพี (*p*-value) มีค่าใกล้เคียงค่าจริงมากที่สุด เมื่อเทียบกับวิธีการตัดข้อมูลแบบแพร์ไวส์ และวิธีค่าทดแทนพหุ

จากงานวิจัยที่กล่าวมาข้างต้นพบว่า ยังไม่มีงานวิจัยใดศึกษาและเปรียบเทียบวิธีการประมาณค่าสูญหายในแผนแบบแฟกทอเรียล 3 ปัจจัยในบล็อกสมบูรณ์เชิงสุ่ม จึงทำให้ผู้วิจัยสนใจศึกษาหัวข้อดังกล่าวนี้ และจากการทบทวนวรรณกรรมพบว่า วิธีค่าคาดหวังสูงสุด วิธีค่าทดแทนพหุ และวิธีเคเนียร์เรสเนเบอร์อิมพิวเทชัน ให้ประสิทธิภาพสูงสุด ทั้งนี้ยังพบว่า มาตรวัดระยะห่างซิตีบล็อก (City Block Distance) ที่ใช้ในการคำนวณหาระยะห่างระหว่างข้อมูลเป็นมาตรวัดระยะทางที่มีแนวโน้มให้ประสิทธิภาพสูงสุดสำหรับการประมาณค่าสูญหายด้วยวิธีเคเนียร์เรสเนเบอร์อิมพิวเทชัน ผู้วิจัยจึงมีความสนใจที่จะศึกษา และเปรียบเทียบประสิทธิภาพวิธีการประมาณค่าสูญหายสำหรับแผนแบบแฟกทอเรียล 3 ปัจจัยในบล็อกสมบูรณ์เชิงสุ่ม โดยใช้วิธีที่มีแนวโน้มให้ประสิทธิภาพสูงของงานวิจัยที่กล่าวมาข้างต้นคือวิธีค่าคาดหวังสูงสุด วิธีค่าทดแทนพหุ 1 (MI1) วิธีค่าทดแทนพหุ 2 (MI2) และวิธีเคเนียร์เรส เนเบอร์อิมพิวเทชัน (KNN) ซึ่งความแตกต่างของวิธีค่าทดแทนพหุ 1 และวิธีค่าทดแทนพหุ 2 ในการวิจัยครั้งนี้คือวิธีค่าทดแทนพหุ 1 ใช้วิธีซิตีบล็อกในการคำนวณหาระยะห่างระหว่างข้อมูลในการประมาณค่าสูญหายด้วยวิธีเคเนียร์เรสเนเบอร์อิมพิวเทชัน ส่วนวิธีค่าทดแทนพหุ 2 ใช้วิธียุคลิดในการคำนวณหาระยะห่างระหว่างข้อมูลในการประมาณค่าสูญหายด้วยวิธีเคเนียร์เรสเนเบอร์อิมพิวเทชัน โดยศึกษาภายใต้สถานการณ์การจำลองข้อมูลที่แตกต่างกัน 108 สถานการณ์ และเกณฑ์ที่ใช้ในการพิจารณาเปรียบเทียบประสิทธิภาพของวิธีการประมาณค่าสูญหาย คือ พิจารณาจากค่าประมาณความคลาดเคลื่อนกำลังสองเฉลี่ย (EMSE) โดยวิธีใดมีค่าประมาณความคลาดเคลื่อนกำลังสองเฉลี่ย (EMSE) ต่ำที่สุด วิธีนั้นจะเป็นวิธีที่ให้ค่าประมาณใกล้เคียงกับค่าจริงมากที่สุด หรืออาจกล่าวได้ว่าวิธีนั้นเป็นวิธีที่มีประสิทธิภาพมากที่สุด

2. วัสดุ อุปกรณ์และวิธีการวิจัย

การประมาณค่าสูญหายสำหรับแผนแบบแฟกทอเรียล 3 ปัจจัยในบล็อกสมบูรณ์เชิงสุ่มครั้งนี้จำลองข้อมูลทั้งหมด 108 สถานการณ์ โดยมีวิธีดำเนินการวิจัยดังนี้

2.1 กำหนดขนาดประชากรและขนาดตัวอย่าง

โดยการวิจัยครั้งนี้กำหนดให้ประชากรที่ใช้ในการวิจัยนี้เป็นข้อมูลที่ได้จากการจำลองสถานการณ์การทดลองในแผนแบบแฟกทอเรียล 3 ปัจจัยในบล็อกสมบูรณ์เชิงสุ่ม คือ ปัจจัย A ปัจจัย B และปัจจัย C โดยที่แต่ละปัจจัยมี 3, 4 และ 5 ระดับ และกำหนดให้มีจำนวนบล็อกเท่ากับ 3 บล็อก ทั้งนี้ขนาดประชากรที่ใช้ในการศึกษาเท่ากับ 160,000 และขนาดตัวอย่าง (n) เท่ากับ 81, 108, 135, 144, 180, 192, 225, 240, 300 และ 375 ซึ่งสามารถเขียนตัวแบบได้ดังสมการที่ (1)

$$y_{ijlq} = \mu + B_q + \alpha_i + \beta_j + \gamma_l + (\alpha\beta)_{ij} + (\alpha\gamma)_{il} + (\beta\gamma)_{jl} + (\alpha\beta\gamma)_{ijl} + \varepsilon_{ijlq} \quad (1)$$

เมื่อ y_{ijlq} คือ ค่าสังเกตจากหน่วยตัวอย่างของปัจจัย A ระดับ i ปัจจัย B ระดับที่ j ปัจจัย C ระดับที่ l และบล็อกที่ q โดยที่ $Y_{ijlq} \sim N(\mu_{ijlq}, \sigma^2)$

μ คือ ค่าเฉลี่ยของประชากร

B_q คือ อิทธิพลของบล็อกที่ q

α_i คือ อิทธิพลของปัจจัย A ระดับที่ i

B_j คือ อิทธิพลของปัจจัย B ระดับที่ j

γ_l คือ อิทธิพลของปัจจัย C ระดับที่ l

$(\alpha\beta)_{ij}$ คือ อิทธิพลร่วมของปัจจัย A ระดับที่ i และปัจจัย B ระดับที่ j

$(\alpha\gamma)_{il}$ คือ อิทธิพลร่วมของปัจจัย A ระดับที่ i และปัจจัย C ระดับที่ l

$(\beta\gamma)_{jl}$ คือ อิทธิพลร่วมของปัจจัย B ระดับที่ j และปัจจัย C ระดับที่ l

$(\alpha\beta\gamma)_{ijl}$ คือ อิทธิพลร่วมของปัจจัย A ระดับที่ i ปัจจัย B ระดับที่ j และปัจจัย C ระดับที่ l

ε_{ijlq} คือ ความคลาดเคลื่อนของค่าสังเกตจากหน่วยตัวอย่างของปัจจัย A ระดับที่ i ปัจจัย B ระดับที่ j ปัจจัย C ระดับที่ l และบล็อกที่ q โดยที่ $\varepsilon_{ijlq} \sim N(0,\sigma_\varepsilon^2)$

2.2 จำลองข้อมูลด้วยเทคนิคมอนติคาร์โล

โดยใช้โปรแกรม R กำหนดให้ข้อมูลมีการแจกแจงปรกติที่มีค่าเฉลี่ยประชากร เท่ากับ 50 และค่าสัมประสิทธิ์การแปรผัน (C.V.) เท่ากับ 10% และ 50% และเพื่อทำให้การจำลองข้อมูลมีหลักเกณฑ์มากขึ้น จึงกำหนดให้ $\sigma_B^2=\sigma_\alpha^2=\sigma_\beta^2=\sigma_\gamma^2=\sigma_{\alpha\beta}^2=\sigma_{\alpha\gamma}^2=\sigma_{\beta\gamma}^2=\sigma_{\alpha\beta\gamma}^2=h\sigma_\varepsilon^2$ โดยค่าคงที่ h เท่ากับ 3 เนื่องจากในการวิจัยครั้งนี้พบว่า เมื่อกำหนดให้ค่าคงที่ h มีค่าแตกต่างกัน กล่าวคือกำหนดให้ h = 1, 2 และ 3 พบว่าค่าความแปรปรวนของค่าสังเกตมีค่าไม่แตกต่างกัน ดังนั้นในบทความงานวิจัยที่นำเสนอในครั้งนี้ผู้วิจัยจึงเลือกค่า h = 3 มานำเสนอเท่านั้น

เมื่อ σ_B^2 คือ ความแปรปรวนอิทธิพลของบล็อก

σ_α^2 คือ ความแปรปรวนของอิทธิพลของปัจจัย A

σ_β^2 คือ ความแปรปรวนของอิทธิพลของปัจจัย B

σ_γ^2 คือ ความแปรปรวนของอิทธิพลของปัจจัย C

$\sigma_{\alpha\beta}^2$ คือ ความแปรปรวนของอิทธิพลร่วมของปัจจัย A และปัจจัย B

$\sigma_{\alpha\gamma}^2$ คือ ความแปรปรวนของอิทธิพลร่วมของปัจจัย A และปัจจัย C

$\sigma_{\beta\gamma}^2$ คือ ความแปรปรวนของอิทธิพลร่วมของปัจจัย B และปัจจัย C

$\sigma_{\alpha\beta\gamma}^2$ คือ ความแปรปรวนของอิทธิพลร่วมของปัจจัย A ปัจจัย B และปัจจัย C

σ_ε^2 คือ ความแปรปรวนของความคลาดเคลื่อนของค่าสังเกต

เนื่องจาก

$$\sigma_Y^2=\sigma_B^2+\sigma_\alpha^2+\sigma_\beta^2+\sigma_\gamma^2+\sigma_{\alpha\beta}^2+\sigma_{\alpha\gamma}^2+\sigma_{\beta\gamma}^2+\sigma_{\alpha\beta\gamma}^2+\sigma_\varepsilon^2$$

และ $C.V.(Y_{ijlq})=\dfrac{S.D.(Y_{ijlq})}{\mu}=\dfrac{\sigma_\varepsilon\sqrt{8h+1}}{\mu}$

เมื่อ $C.V.(Y_{ijlq})$ คือ ค่าสัมประสิทธิ์การแปรผันของค่าสังเกตจากหน่วยตัวอย่างของปัจจัย A ระดับ i ปัจจัย B ระดับที่ j ปัจจัย C ระดับที่ l และบล็อกที่ q

$S.D.(Y_{ijlq})$ คือ ส่วนเบี่ยงเบนมาตรฐานของค่าสังเกตจากหน่วยตัวอย่างของปัจจัย A ระดับ i ปัจจัย B ระดับที่ j ปัจจัย C ระดับที่ l และ บล็อกที่ q

จึงได้ว่า $\sigma_\varepsilon^2=\dfrac{\left(C.V.(Y_{ijlq})\cdot\mu\right)^2}{8h+1}$

และเนื่องจาก $Y_{ijlq} \sim N(\mu, \sigma_Y^2)$

ดังนั้นความแปรปรวนของค่าสังเกต (σ_Y^2) เท่ากับ 25 และ 625 ตามลำดับ

จากนั้นสุ่มข้อมูลตัวอย่างขนาด 81, 108, 135, 144, 180, 192, 225, 240, 300 และ 375 จากข้อมูลประชากร

2.3 สุ่มตัดข้อมูล

สำหรับตัวแปรตอบสนองให้มีการสูญหายแบบสุ่ม (MAR) และตรวจสอบรูปแบบการสูญหายของข้อมูล ถ้าหากข้อมูลสูญหายในแถวเดียวกันหมด จะต้องดำเนินการสุ่มตัดให้ข้อมูลสูญหายอีกครั้ง ทั้งนี้กำหนดให้ร้อยละของข้อมูลสูญหายมี 2 ระดับ คือ ร้อยละ 5 และ 10 ของขนาดตัวอย่าง

2.4 ประมาณค่าสูญหายด้วยวิธีต่างๆ

2.4.1 วิธีค่าคาดหวังสูงสุด (EM) เป็นกระบวนการวนซ้ำเพื่อใช้สำหรับหาค่าประมาณภาวะน่าจะเป็นสูงสุด (Maximum Likelihood) ของพารามิเตอร์ เมื่อมีข้อมูลบางส่วนเกิดการสูญหาย ซึ่งถูกเสนอโดย Dempster และคณะ [7] และในการประมาณค่าสูญหายด้วยวิธีนี้ต้องทำการพิจารณาก่อนว่า

ค่าสังเกตที่เหลืออยู่นั้นมีการแจกแจงแบบใด ถ้ามีการแจกแจงปรกติสามารถดำเนินการต่อไปได้ [2] แต่ถ้าข้อมูลไม่มีการแจกแจงปรกติจะดำเนินการจำลองข้อมูลชุดใหม่ โดยงานวิจัยนี้สมมติให้ค่าสังเกต Y หรือ $Y = (Y_{mis}, Y_{obs})$ มี n ค่า ซึ่งประกอบด้วย 2 ส่วน คือ Y_{mis} และ Y_{obs} เมื่อ Y_{mis} แทน ค่าสังเกตที่สูญหาย มีจำนวน m ค่า และ Y_{obs} แทน ค่าสังเกตที่เก็บรวบรวมมาได้ มีจำนวน $n–m$ ค่า ในที่นี้กำหนดให้ปัจจัย A มี a ระดับ ปัจจัย B มี b ระดับ ปัจจัย C มี c ระดับ และบล็อกมี ω บล็อก และเนื่องจากการวิจัยครั้งนี้กำหนดให้มี 3 บล็อก เพราะฉะนั้น $\omega = 3$ โดยวิธีค่าคาดหวังสูงสุด (EM) มีขั้นตอนการดำเนินการดังนี้

ขั้นตอนที่ 1 การประมาณค่าคาดหวัง (E-step) ซึ่งเป็นขั้นตอนในการหาค่าคาดหวังของค่าสูญหายภายใต้เงื่อนไขของชุดข้อมูลที่ไม่มีการสูญหายและพารามิเตอร์ตัวปัจจุบัน เพื่อนำค่าคาดหวังที่ได้นี้ไปประมาณค่าสูญหาย โดยคำนวณได้จาก

$$E\left(\sum_{a=1}^{n} y_a \middle| \theta^{(t)}, Y_{obs}\right) = \sum_{a=1}^{n-m} y_{obs,a} + (n-m)\mu^{(t)} \qquad (2)$$

$$E\left(\sum_{a=1}^{n} y_a^2 \middle| \theta^{(t)}, Y_{obs}\right) = \sum_{a=1}^{n-m} y_{obs,a}^2 + (n-m)\left(\left(\mu^{(t)}\right)^2 + \left(\sigma^{(t)}\right)^2\right) \qquad (3)$$

เมื่อ $t = 0,1,2,...$ สำหรับพารามิเตอร์ตัวปัจจุบันคือ $\theta^{(y)} = (\mu^{(t)}, \sigma^{(t)})$ และกำหนดให้พารามิเตอร์เริ่มต้นคือ $\mu^{(o)}$ และ $(\sigma^{(o)})^2$ ดังนี้

$$\mu^{(0)} = \frac{\sum_{a=1}^{n} y_{obs,a}}{n} \text{ และ } \left(\sigma^{(0)}\right)^2 = \frac{\sum_{a=1}^{n-m} y_{obs,a}^2}{n} - \left(\mu^{(0)}\right)^2$$

ขั้นตอนที่ 2 การใช้ค่าสูงสุด (M-step) ซึ่งเป็นการประมาณค่าโดยวิธีภาวะน่าจะเป็นสูงสุดของพารามิเตอร์จากข้อมูลที่ไม่มีการสูญหาย และทดแทนค่าสูญหายด้วยค่าที่ได้จากขั้นตอนการประมาณค่าคาดหวัง (E-step) และทำการประมาณค่าคาดหวังซ้ำ เพื่อเปรียบเทียบจนได้ค่าที่เปลี่ยนแปลงน้อยมาก และใช้ค่านั้นสร้างค่าทดแทนค่าข้อมูลสูญหาย โดยหาได้จาก

$$\mu^{(t+1)} = \frac{E\left(\sum_{a=1}^{n} y_a \middle| \theta^{(t)}, Y_{obs}\right)}{n} \qquad (4)$$

$$\left(\sigma^{(t+1)}\right)^2 = \frac{E\left(\sum_{a=1}^{n} y_a^2 \middle| \theta^{(t)}, Y_{obs}\right)}{n} - \left(\mu^{(t+1)}\right)^2 \qquad (5)$$

เมื่อ $t = 0,1,2,...$ และทำการวนซ้ำในสมการที่ (2)–(5) จนกระทั่ง $\mu^{(t+1)}$ และ ลู่เข้าสู่ค่าคงที่ $\left(\sigma^{(t+1)}\right)^2$

เมื่อ $\sum_{a=1}^{n-m} y_{obs,a}$ คือ ผลรวมของค่าสังเกตที่เก็บรวบรวมมาได้

$\sum_{a=1}^{n-m} y_{obs,a}^2$ คือ ผลรวมกำลังสองของค่าสังเกตที่เก็บรวบรวมมาได้

$E\left(\sum_{a=1}^{n} y_a \middle| \theta^{(t)}, Y_{obs}\right)$ คือ ค่าคาดหวังของผลรวมของค่าสังเกตที่เก็บรวบรวมมาได้

$E\left(\sum_{a=1}^{n} y_a^2 \middle| \theta^{(t)}, Y_{obs}\right)$ คือ ค่าคาดหวังของผลรวมกำลังสองของค่าสังเกตที่เก็บรวบรวมมาได้

$\mu^{(t)}$ คือ ค่าเฉลี่ยที่ได้จากการประมาณค่ารอบที่ t

$\alpha^{(t)}$ คือ ส่วนเบี่ยงเบนมาตรฐานที่ได้จากการประมาณค่ารอบที่ t

$\hat{\mu}$ คือ ตัวประมาณของค่าเฉลี่ย

$\hat{\sigma}$ คือ ตัวประมาณของส่วนเบี่ยงเบนมาตรฐาน

2.4.2 วิธีค่าทดแทนพหุ (MI) ได้แนวคิดมาจากวิธี Simple Imputation ซึ่งถูกเสนอโดย Rubin [8] เพื่อนำมาใช้ในการแก้ปัญหาเมื่อข้อมูลมีการสูญหายแบบสุ่ม โดยแต่ละข้อมูลสูญหายจะถูกแทนที่ด้วยชุดของข้อมูลตั้งแต่ 2 ชุดขึ้นไป เพื่อสร้างข้อมูลที่สมบูรณ์จำนวน x ชุด โดยที่ $x = 0,1,2,...$ ทั้งนี้ค่า x ในช่วงดังกล่าวต้องเพียงพอที่จะให้ผลลัพธ์ที่มีประสิทธิภาพ โดยสามารถคำนวณหาประสิทธิภาพของการประมาณค่าได้จาก [9]

ประสิทธิภาพของการประมาณค่า $= \left(1 + \frac{\gamma}{m}\right)^{-1}$

เมื่อ γ แทน ร้อยละของข้อมูลสูญหาย

วิธีค่าทดแทนพหุ ประกอบด้วย 3 ขั้นตอน ดังนี้

ขั้นตอนที่ 1 ขั้นตอนการสร้างข้อมูลสูญหาย และนำค่าที่ได้ไปทดแทนข้อมูลสูญหายได้เป็นชุดข้อมูลที่สมบูรณ์ แต่ทำซ้ำเพื่อให้ได้ชุดข้อมูลหลายๆ ชุด โดยในการวิจัยครั้งนี้

ดำเนินการสร้างข้อมูลที่สมบูรณ์ 4 ชุด ($x = 4$) ซึ่งมาจากการประมาณค่าด้วยวิธีค่าเฉลี่ย (Mean Imputation) วิธีค่ามัธยฐาน (Median Imputation) วิธีค่าคาดหวังสูงสุด และวิธีเคเนียร์เรสเนเบอร์อิมพิวเทชัน และเนื่องจากในการวิจัยครั้งนี้มีร้อยละการสูญหายของข้อมูลเท่ากับร้อยละ 5 และ 10 เมื่อนำไปคำนวณหาประสิทธิภาพของการประมาณค่าจะได้ว่าการประมาณค่าจะมีประสิทธิภาพร้อยละ 98.77 และ 97.56 ตามลำดับ เมื่อนำไปเทียบกับการสร้างข้อมูลที่สมบูรณ์ 5 ชุด ($x = 5$) การประมาณค่าจะมีประสิทธิภาพเพิ่มขึ้นเล็กน้อยเป็นร้อยละ 99.01 และ 98.04 ตามลำดับ ดังนั้นการสร้างข้อมูลที่สมบูรณ์ 4 ชุด จึงเพียงพอที่จะให้ผลลัพธ์ที่มีประสิทธิภาพ

ขั้นตอนที่ 2 การวิเคราะห์ข้อมูลแต่ละชุดแยกกัน เพื่อประมาณค่าพารามิเตอร์จากข้อมูลแต่ละชุด ในขั้นตอนนี้จะมีค่าพารามิเตอร์ที่สนใจเท่ากับจำนวนชุดข้อมูลที่สร้างขึ้นในขั้นตอนแรก

ขั้นตอนที่ 3 การรวบรวมผลที่ได้มาสรุปค่าที่จะใช้แทนค่าสูญหายทั้งหมด โดยการนำค่าประมาณพารามิเตอร์ที่คำนวณได้จากแต่ละชุดข้อมูลมารวมกันด้วยการเฉลี่ย เพื่อนำค่าที่ได้ไปสร้างข้อมูลสูญหาย

ในงานวิจัยนี้วิธีค่าทดแทนพหุ 1 และวิธีค่าทดแทนพหุ 2 แตกต่างกันที่ขั้นตอนที่ 1 โดยวิธีค่าทดแทนพหุ 1 ใช้วิธีซิตีบล็อก (City Block Distance) ดังสมการที่ (6) ในการคำนวณหาระยะห่างระหว่างข้อมูลในการประมาณค่าสูญหายด้วยวิธี KNN ในขณะที่วิธีค่าทดแทนพหุ 2 ใช้วิธียุคลิด (Euclidean Distance) ซึ่งนิยมใช้กันเป็นส่วนใหญ่ [10] ในการคำนวณหาระยะห่างระหว่างข้อมูลในการประมาณค่าสูญหายด้วยวิธี KNN ดังนั้นการสร้างข้อมูลสูญหายที่ได้จึงมีค่าต่างกัน

2.4.3 วิธีเคเนียร์เรสเนเบอร์อิมพิวเทชัน (KNN) เป็นวิธีการคำนวณหาข้อมูลที่อยู่ใกล้กับข้อมูลสูญหายมากที่สุด k ตัว ซึ่งก็คือมีระยะห่างระหว่างข้อมูลสูญหายที่ต้องการพิจารณากับข้อมูลที่เก็บรวบรวมมาได้ต่ำที่สุด k ตัว [11] โดยมีขั้นตอนดังนี้

ขั้นตอนที่ 1 กำหนดค่าคงที่ k เพื่อใช้ในการพิจารณาจำนวนสมาชิกที่ใกล้ที่สุด โดยที่ Duba และ Hart [12] ได้เสนอให้ $k \approx \sqrt{n-m}$ เมื่อ n–m คือ จำนวนค่าสังเกตที่เก็บรวบรวมมาได้

ขั้นตอนที่ 2 คำนวณหาระยะห่างระหว่างค่าสังเกตของข้อมูลที่สูญหายที่ต้องการพิจารณากับค่าสังเกตของข้อมูลที่เก็บรวบรวมมาได้ โดยใช้ข้อมูลที่เก็บรวบรวมข้อมูลมาได้ในแถวเดียวกันกับข้อมูลที่สูญหาย และคำนวณด้วยวิธีซิตีบล็อก [13] ซึ่งมีรูปแบบการคำนวณดังนี้

$$d(y_b, y_c) = \sum_{q=1}^{p-1} \left| y_{b,q} - y_{c,q} \right| \qquad (6)$$

เมื่อ $d(y_b, y_c)$ คือ ระยะห่างระหว่างข้อมูลแถวที่ b และข้อมูลแถวที่ c

p คือ จำนวนคอลัมน์ในแถวที่มีข้อมูลสูญหาย

$y_{b,q}$ คือ ค่าสังเกตที่สูญหาย แถวที่ b คอลัมน์ที่ q

$y_{c,q}$ คือ ค่าสังเกตที่เก็บมาได้ แถวที่ c คอลัมน์ที่ q

ตัวอย่างการคำนวณหาระยะห่างระหว่างข้อมูลด้วยวิธีซิตีบล็อกในการประมาณค่าสูญหายด้วยวิธีเคเนียร์เรสเนเบอร์อิมพิวเทชัน ดังข้อมูลแสดงในตารางที่ 1

ตารางที่ 1 ข้อมูลจากแผนแบบแฟกทอเรียล 3 ปัจจัยในบล็อกสมบูรณ์เชิงสุ่ม

	C1	C2	C3
R1	46.7641	60.68464	49.87981
R2	49.21649	NA	53.37298
R3	49.18564	50.658	48.03401

ระยะห่างระหว่างข้อมูลด้วยวิธีซิตีบล็อก กรณีมีข้อมูลสูญหาย 1 ค่า คำนวณได้ดังนี้

$$d(y_2, y_1) = |49.21649 - 46.7641| + |53.37298 - 49.87981| = 5.94556$$

$$d(y_2, y_3) = |49.21649 - 49.18564| + |53.37298 - 48.03401| = 5.36982$$

ขั้นตอนที่ 3 เรียงลำดับระยะห่างระหว่างข้อมูล แล้วพิจารณาเลือกข้อมูลที่ใกล้กับข้อมูลสูญหายมากที่สุด k จำนวน

ขั้นตอนที่ 4 ประมาณค่าสูญหายจากค่าเฉลี่ยของข้อมูลที่ใกล้กับข้อมูลสูญหายมากที่สุด ดังสมการที่ (7)

$$\hat{y}_r = \frac{\sum_{g=1}^{k} y_r}{k} \quad (7)$$

เมื่อ $\hat{y}_r$ คือ ค่าสังเกตที่สูญหายที่ได้จากการประมาณค่าใหม่ค่าที่ r

y_r คือ ค่าสังเกตที่เก็บมาได้ที่ตรงกับค่าสังเกตที่สูญหายค่าที่ r

k คือ ค่าคงที่ ซึ่ง $k \approx \sqrt{n-m}$

n–m คือ จำนวนค่าสังเกตที่เก็บรวบรวมมาได้

หลังจากนั้นแทนค่าสังเกตที่สูญหายด้วยค่าเฉลี่ยของข้อมูลที่อยู่ใกล้ที่สุด

2.5 คำนวณหาค่าประมาณความคลาดเคลื่อนกำลังสองเฉลี่ย (EMSE)

เพื่อใช้เป็นเกณฑ์ในการเปรียบเทียบประสิทธิภาพของวิธีการประมาณค่าสูญหาย โดยมีวิธีการคำนวณดังนี้

1) คำนวณหาค่าประมาณความคลาดเคลื่อนกำลังสองเฉลี่ย (MSE) รอบที่ l ของค่าสังเกตจากสูตรสมการที่ (8)

$$MSE_l(Y) = \frac{\sum_{h=1}^{m} \left(y_{h,l} - \hat{y}_{h,l}\right)^2}{m} \quad (8)$$

เมื่อ $y_{h,l}$ คือ ค่าจริงที่ h ที่ได้จากการจำลองซ้ำรอบที่ l

$\hat{y}_{h,l}$ คือ ค่าประมาณค่าที่ h ที่ได้จากการประมาณค่าซ้ำรอบที่ l

m คือ จำนวนข้อมูลสูญหาย

2) คำนวณหาค่าประมาณความคลาดเคลื่อนกำลังสองเฉลี่ย (EMSE) ของค่าสังเกตดังสมการที่ (9)

$$EMSE(Y) = \frac{\sum_{l=1}^{2000} MSE_l(Y)}{2000} \quad (9)$$

เมื่อ $MSE_l(Y)$ คือ ค่าประมาณความคลาดเคลื่อนกำลังสองเฉลี่ย (MSE) รอบที่ l ของค่าสังเกต

ในงานวิจัยนี้จะทำการทดลองซ้ำในแต่ละสถานการณ์ 2,000 รอบ โดยจะพิจารณาว่าถ้าวิธีการประมาณค่าสูญหายวิธีใดมีค่าประมาณความคลาดเคลื่อนกำลังสองเฉลี่ย (EMSE) ต่ำที่สุด วิธีนั้นจะเป็นวิธีที่ให้ค่าประมาณใกล้เคียงกับค่าจริงมากที่สุด หรืออาจกล่าวได้ว่าวิธีนั้นเป็นวิธีที่มีประสิทธิภาพสูงสุด

3. ผลการทดลอง

จากการจำลองข้อมูลด้วยเทคนิคมอนติคาร์โล จำนวน 108 สถานการณ์ และทำการทดลองซ้ำในแต่ละสถานการณ์ 2,000 รอบ เพื่อประมาณค่าสูญหายสำหรับแผนแบบแฟกทอเรียล 3 ปัจจัยในบล็อกสมบูรณ์เชิงสุ่มด้วยวิธีค่าคาดหวังสูงสุด (EM) วิธีค่าทดแทนพหุ 1 (MI1) วิธีค่าทดแทนพหุ 2 (MI2) และวิธีเคเนียร์เรสเนเบอร์อิมพิวเทชัน (KNN) ได้ผลการวิจัยดังแสดงในตารางที่ 2 ถึงตารางที่ 4 ดังนี้

ผลการวิจัยการประมาณค่าสูญหายสำหรับแผนแบบแฟกทอเรียล 3 ปัจจัยในบล็อกสมบูรณ์เชิงสุ่มที่มีจำนวนบล็อกเท่ากับ 3 ดังตารางที่ 1, 2 และ 3 สรุปได้ดังนี้

1) เมื่อพิจารณาจำนวนปัจจัยจะพบว่า ทั้งในกรณีที่จำนวนปัจจัยเท่ากัน และกรณีที่จำนวนปัจจัยไม่เท่ากัน วิธีเคเนียร์เรสเนเบอร์อิมพิวเทชัน มีค่าประมาณความคลาดเคลื่อนกำลังสองเฉลี่ยต่ำสุด ซึ่งพบว่า จำนวนปัจจัยที่เท่ากันหรือไม่เท่ากันไม่มีผลต่อประสิทธิภาพของวิธีการประมาณค่าสูญหาย

2) เนื่องจากในงานวิจัยนี้กำหนดให้ค่าสัมประสิทธิ์การแปรผัน (C.V.) เท่ากับ 10% และ 50% เมื่อค่าคงที่ h เท่ากับ 3 จึงได้ว่าความแปรปรวนของค่าสังเกต (σ_Y^2) เท่ากับ 25 และ 625 ตามลำดับ ดังนั้นเมื่อพิจารณาค่าประมาณความคลาดเคลื่อนกำลังสองเฉลี่ย พบว่า เมื่อค่าสัมประสิทธิ์การแปรผัน (C.V.) มีค่าเพิ่มขึ้น ความแปรปรวนของค่าสังเกตจะเพิ่มขึ้น และเมื่อความแปรปรวนของค่าสังเกตเพิ่มขึ้น มีแนวโน้มทำให้ค่าประมาณความคลาดเคลื่อนกำลังสองเฉลี่ยมีค่าเพิ่มขึ้นตามไปด้วยเช่นกัน

3) ในทุกสถานการณ์ของการศึกษาพบว่า วิธีเคเนียร์เรสเนเบอร์อิมพิวเทชัน จะให้ค่าประมาณความคลาดเคลื่อนกำลังสองเฉลี่ยต่ำสุด จึงกล่าวได้ว่าวิธีเคเนียร์เรสเนเบอร์อิมพิวเทชัน มีแนวโน้มให้ประสิทธิภาพ

ตารางที่ 2 ค่าประมาณความคลาดเคลื่อนกำลังสองเฉลี่ยของการประมาณค่าสูญหายทั้ง 4 วิธี เมื่อปัจจัย A มี 3 ระดับ

จำนวนปัจจัย A			3								
จำนวนปัจจัย B			3			4			5		
จำนวนปัจจัย C			3	4	5	3	4	5	3	4	5
ข้อมูลสูญหายร้อยละ 5	$\sigma_Y^2 = 25$	EM	54.83618	52.22104	52.7111	56.16509	53.01584	53.66945	56.85091	56.73382	52.82985
		MI1	55.05089	52.21956	53.15863	56.70702	53.1872	54.44403	55.02762	52.07811	50.61329
		MI2	3317.75	1623.799	1777.378	700.7083	2488.512	1533.963	3620.616	631.0163	526.6475
		KNN	**30.51706**	**28.53236**	**28.55324**	**31.42628**	**28.55173**	**28.83089**	**30.26576**	**30.34731**	**28.18913**
	$\sigma_Y^2 = 625$	EM	22912.24	23843.79	23535.4	22455.48	23599.25	22849.13	22090.8	22456.89	22288.33
		MI1	23458.17	22395.46	22635.15	22566.95	22563.96	21868.03	22717.84	23556.68	21733.12
		MI2	260036.9	440369.8	156278.9	680509.3	42782.07	76258.94	49071.09	79288.21	51746.92
		KNN	**12618.42**	**13052.63**	**12496.43**	**12494.12**	**12446.09**	**12176.04**	**12476.34**	**12579.65**	**11698.62**
ข้อมูลสูญหายร้อยละ 10	$\sigma_Y^2 = 25$	EM	53.82513	54.69676	53.76136	54.67204	54.00654	54.65059	53.6196	54.10663	54.41685
		MI1	55.22548	53.67534	52.39187	53.78817	52.30702	51.78341	52.77193	54.10197	51.53134
		MI2	1403.511	924.8652	912.4614	1034.673	528.0239	1416.758	6951.634	865.7354	1429.488
		KNN	**30.42651**	**31.03443**	**29.10014**	**29.74221**	**29.13676**	**29.3601**	**30.20818**	**29.67245**	**29.54671**
	$\sigma_Y^2 = 625$	EM	22621.66	22448.07	22834.6	23415.05	22152.22	22601.06	23067.28	23704.94	22366.92
		MI1	22184.63	23194.65	22089.18	22625.65	21440.83	21722.94	21652.18	21621.19	21794.48
		MI2	186390.8	373317.7	73365.64	39260.37	185846.8	32779.8	164563.1	155882.1	60939.71
		KNN	**12512.85**	**13052.73**	**12389.96**	**13224.52**	**12212.98**	**12122.54**	**12341.03**	**12665.98**	**12172.97**

หมายเหตุ: ตัวหนา หมายถึง ค่าประมาณความคลาดเคลื่อนกำลังสองเฉลี่ยต่ำที่สุดในสถานการณ์นั้น

ตารางที่ 3 ค่าประมาณความคลาดเคลื่อนกำลังสองเฉลี่ยของการประมาณค่าสูญหายทั้ง 4 วิธี เมื่อปัจจัย A มี 4 ระดับ

จำนวนปัจจัย A			4								
จำนวนปัจจัย B			3			4			5		
จำนวนปัจจัย C			3	4	5	3	4	5	3	4	5
ข้อมูลสูญหายร้อยละ 5	$\sigma_Y^2 = 25$	EM	54.07452	55.33826	54.7892	57.74635	54.63667	56.39958	55.59649	54.18959	56.3274
		MI1	55.61977	52.71921	54.31887	56.79804	53.36156	55.53407	53.24105	53.44824	53.42784
		MI2	1575.271	3032.425	1726.38	661.015	1480.208	3084.428	617.189	1143.699	1061.527
		KNN	**29.14396**	**29.86641**	**29.10879**	**31.49534**	**28.7483**	**29.55774**	**29.09804**	**28.87255**	**29.17042**
	$\sigma_Y^2 = 625$	EM	23172.96	22540.9	22349.02	23719.7	23179.48	23287.78	22881.98	23574.22	23002.47
		MI1	22495.33	22148.83	22227.54	22545.98	22535.66	22959.58	22570.82	22335.02	22706.38
		MI2	206098.2	285247	195002.6	42337.29	208715.6	200833.1	610476.8	93946.98	44272.87
		KNN	**13021.15**	**12232.79**	**11978.3**	**12586.38**	**12211.46**	**12235.69**	**12499.58**	**12302.36**	**12209.64**
ข้อมูลสูญหายร้อยละ 10	$\sigma_Y^2 = 25$	EM	54.57817	55.33655	54.27769	53.03452	56.29976	54.53241	54.33059	55.0475	56.72438
		MI1	53.42512	51.80532	52.62701	52.62846	52.02882	53.53306	53.94203	51.85477	54.43124
		MI2	3858.245	892.2146	513.0816	1500.048	1061.912	2220.711	1476.236	1097.603	3687.785
		KNN	**30.39533**	**29.42599**	**29.12093**	**29.15183**	**29.33359**	**29.14413**	**29.99602**	**29.23972**	**29.57197**
	$\sigma_Y^2 = 625$	EM	23130.68	22030.41	22848.7	23095.12	23174.45	23304.74	22287.27	23145.53	22516.01
		MI1	23031.17	22711.16	21604.89	22181.91	21628.54	22592.92	21536.51	21978.17	22193.28
		MI2	443982.2	46621.89	67166.71	140526.2	85359.62	92150.68	26620.03	50012.53	191554.6
		KNN	**12876.07**	**12631.48**	**12375.73**	**12511.11**	**12204.14**	**12449.55**	**12377.24**	**12036.21**	**12149.21**

หมายเหตุ: ตัวหนา หมายถึง ค่าประมาณความคลาดเคลื่อนกำลังสองเฉลี่ยต่ำที่สุดในสถานการณ์นั้น

ตารางที่ 4 ค่าประมาณความคลาดเคลื่อนกำลังสองเฉลี่ยของการประมาณค่าสูญหายทั้ง 4 วิธี เมื่อปัจจัย A มี 5 ระดับ

จำนวนปัจจัย A			5								
จำนวนปัจจัย B			3			4			5		
จำนวนปัจจัย C			3	4	5	3	4	5	3	4	5
ข้อมูลสูญหายร้อยละ 5	$\sigma_Y^2 = 25$	EM	54.83618	52.22104	52.7111	56.16509	53.01584	53.66945	56.85091	56.73382	52.82985
		MI1	55.05089	52.21956	53.15863	56.70702	53.1872	54.44403	55.02762	52.07811	50.61329
		MI2	3317.75	1623.799	1777.378	700.7083	2488.512	1533.963	3620.616	631.0163	526.6475
		KNN	**30.51706**	**28.53236**	**28.55324**	**31.42628**	**28.55173**	**28.83089**	**30.26576**	**30.34731**	**28.18913**
	$\sigma_Y^2 = 625$	EM	22912.24	23843.79	23535.4	22455.48	23599.25	22849.13	22090.8	22456.89	22288.33
		MI1	23458.17	22395.46	22635.15	22566.95	22563.96	21868.03	22717.84	23556.68	21733.12
		MI2	260036.9	440369.8	156278.9	680509.3	42782.07	76258.94	49071.09	79288.21	51746.92
		KNN	**12618.42**	**13052.63**	**12496.43**	**12494.12**	**12446.09**	**12176.04**	**12476.34**	**12579.65**	**11698.62**
ข้อมูลสูญหายร้อยละ 10	$\sigma_Y^2 = 25$	EM	53.82513	54.69676	53.76136	54.67204	54.00654	54.65059	53.6196	54.10663	54.41685
		MI1	55.22548	53.67534	52.39187	53.78817	52.30702	51.78341	52.77193	54.10197	51.53134
		MI2	1403.511	924.8652	912.4614	1034.673	528.0239	1416.758	6951.634	865.7354	1429.488
		KNN	**30.42651**	**31.03443**	**29.10014**	**29.74221**	**29.13676**	**29.3601**	**30.20818**	**29.67245**	**29.54671**
	$\sigma_Y^2 = 625$	EM	22621.66	22448.07	22834.6	23415.05	22152.22	22601.06	23067.28	23704.94	22366.92
		MI1	22184.63	23194.65	22089.18	22625.65	21440.83	21722.94	21652.18	21621.19	21794.48
		MI2	186390.8	373317.7	73365.64	39260.37	185846.8	32779.8	164563.1	155882.1	60939.71
		KNN	**12512.85**	**13052.73**	**12389.96**	**13224.52**	**12212.98**	**12122.54**	**12341.03**	**12665.98**	**12172.97**

หมายเหตุ: ตัวหนา หมายถึง ค่าประมาณความคลาดเคลื่อนกำลังสองเฉลี่ยต่ำที่สุดในสถานการณ์นั้น

สูงสุด รองลงมาคือวิธีค่าทดแทนพหุ 1 และลำดับถัดมาคือวิธีค่าคาดหวังสูงสุด นอกจากนี้ยังได้อีกว่าวิธีที่มีแนวโน้มให้ประสิทธิภาพต่ำสุด คือ วิธีค่าทดแทนพหุ 2 เนื่องจากให้ค่าประมาณความคลาดเคลื่อนกำลังสองเฉลี่ยสูงสุดในทุกสถานการณ์

4. อภิปรายผลและสรุป

1) จากการจำลองข้อมูลพบว่า เมื่อกำหนดให้แต่ละตัวแบบมีจำนวนบล็อก 3, 4 และ 5 บล็อก พบว่า จำนวนบล็อกที่แตกต่างกันไม่มีผลต่อค่าประมาณความคลาดเคลื่อนกำลังสองเฉลี่ย จึงกล่าวได้ว่าจำนวนบล็อกมีแนวโน้มที่ไม่มีผลต่อค่าประมาณความคลาดเคลื่อนกำลังสองเฉลี่ย ในงานวิจัยนี้จึงนำเสนอตารางผลการวิจัยเพียงแค่จำนวนบล็อกเท่ากับ 3 บล็อกเท่านั้น กล่าวคือไม่ว่าจะใช้จำนวนบล็อกเป็น 3, 4 หรือ 5 พบว่า ผลการวิจัยยังคงเหมือนเดิม นั่นคือวิธีเคเนียร์เรสเนเบอร์อิมพิวเทชันให้ประสิทธิภาพสูงสุด ในทุกสถานการณ์ของการศึกษา

2) ค่าคงที่ h เป็นค่าที่กำหนดขึ้นเพื่อใช้ในการศึกษาขนาดความแปรปรวนของความคลาดเคลื่อนของการทดลอง เมื่อกำหนดให้ค่าสัมประสิทธิ์การแปรผัน (C.V.) เท่ากัน และค่าคงที่ h มีค่าเพิ่มขึ้น จะทำให้ความแปรปรวนของความคลาดเคลื่อนของค่าสังเกตลดลง แต่ค่าคงที่ h ไม่มีผลต่อความแปรปรวนของค่าสังเกต เนื่องจากค่าคงที่ h ในระดับต่างๆ ที่ค่าสัมประสิทธิ์การแปรผัน (C.V.) เท่ากัน ความแปรปรวนของค่าสังเกตจะไม่ต่างกัน

3) ในการประมาณค่าสูญหายจะต้องตรวจสอบรูปแบบการสูญหายของข้อมูลก่อนเสมอ กล่าวคือ เมื่อใดก็ตามที่ข้อมูลเกิดการสูญหายในแถวเดียวกันหมด จะไม่สามารถประมาณค่าสูญหายด้วยวิธีเคเนียร์เรสเนเบอร์อิมพิวเทชันได้ เนื่องจากไม่สามารถคำนวณหาระยะห่างระหว่างข้อมูล เพื่อนำมาพิจารณาเลือกข้อมูลที่ใกล้กับข้อมูลสูญหายมากที่สุด k จำนวนได้

4) วิธีเคเนียร์เรสเนเบอร์อิมพิวเทชัน มีแนวโน้มให้ประสิทธิภาพสูงสุด ซึ่งสอดคล้องกับผลการวิจัยของวิสุทธิดา [5]

5) ควรศึกษาวิธีการประมาณค่าสูญหายวิธีอื่นเพิ่มเติม เช่น วิธีการประมาณค่าสูญหายที่ได้จากการรวมวิธีเคเนียร์เรสเนเบอร์อิมพิวเทชันกับวิธีค่าคาดหวังสูงสุด ด้วยฟังก์ชันถ่วงน้ำหนัก และวิธีการประมาณค่าสูญหายที่ได้จากการรวมวิธีเคเนียร์เรสเนเบอร์อิมพิวเทชันกับวิธีค่าทดแทนพหุด้วยฟังก์ชันถ่วงน้ำหนัก

5. กิตติกรรมประกาศ

งานวิจัยนี้ได้รับการสนับสนุนทุนการทำวิจัยจากโครงการพัฒนากำลังคนด้านวิทยาศาสตร์ (ทุนเรียนดีวิทยาศาสตร์แห่งประเทศไทย)

เอกสารอ้างอิง

[1] B. Chomtee, “Factorial designs,” in *Statistical Experimental Design: Theory and Analysis by Using SAS Software*. Bangkok: Department of Statistics, Faculty of Science, Kasetsart University, 2013 (in Thai).

[2] R. J. A. Little and D. B. Rubin, *Statistical Analysis with Missing Data*, 2nd ed. New York: John Wiley and Sons, 2002.

[3] P. Damrongsuttipong, “A comparison of missing value estimation methods for randomized complete block design,” M.S. thesis, Department of Statistics, Faculty of Commerce and Accountancy, Chulalongkorn University, 2003 (in Thai).

[4] S. Kannika, “A comparison of missing value estimation methods for latin square design,” M.S. thesis, Department of Statistics, Faculty of Commerce and Accountancy, Chulalongkorn University, 2006 (in Thai).

[5] W. Sriduangchot, “A study on the efficiency of missing data estimation methods for two factors factorial experiment in randomized complete block design,” M.S. thesis, Department of Educational Research and Statistics, Faculty of Education, Srinakharinwirot University, 2013 (in Thai).

[6] E. C. Lovelyn, “Estimation of missing values in replicated factorial experiment,” M.S. thesis, Department of Mathematics, Faculty of Physical Sciences, Ahmadu Bello University, 2014.

[7] A. P. Dempster, N. M. Laird, and D. B. Rubin, “Maximum likelihood from incomplete data via the EM algorithm,” *Journal of the Royal Statistical Society: Series B (Methodological)*, vol. 39, no. 1, pp. 1–38, 1977.

[8] D. B. Rubin, *Multiple Imputation for Nonresponse in Surveys*. New York: John Wiley and Sons, 1987.

[9] S. Sinharay, H. S. Stern, and D. Russell, “The use of multiple imputation for the analysis of missing data,” *Psychological Methods*, vol. 6, no. 4, pp. 317–329, 2001.

[10] M. M. Deza and E. Deza, *Encyclopedia of Distances*. New York: Springer, 2009.

[11] P. Jonsson and C. Wohlin, “An evaluation of k-nearest neighbur imputation using likert data,” in *Proceedings of the Software Metrics*, 2004, pp. 108–118.

[12] R. O. Duba and P. E. Hart, *Pattern Classification and Scene Analysis*. New York: John Wiley and Sons, 1987.

[13] K. Chomboon, P. Chujai, P. Teerarassamee, K. Kerdprasop, and N. Kerdprasop, “An empirical study of distance metrics for k-nearest neighbor algorithmz,” in *Proceedings of the 3rd International Conference on Industrial Applicatio*n Engineering, 2015, pp. 280–285.