บทความวิจัย

การเปรียบเทียบการประมาณค่าสูญหายในแผนแบบบล็อกสุ่มสมบูรณ์

ชญาดา แก้วชัยเจริญกิจ* บุญอ้อม โฉมที และ วันดี วณิชย์ศักดิ์พงศ์
ภาควิชาสถิติ คณะวิทยาศาสตร์ มหาวิทยาลัยเกษตรศาสตร์

* ผู้นิพนธ์ประสานงาน โทรศัพท์ 09 4498 6258 อีเมล: k.chayadaa@gmail.com DOI: 10.14416/j.kmutnb.2021.07.006
รับเมื่อ 28 พฤษภาคม 2563 แก้ไขเมื่อ 14 กรกฎาคม 2563 ตอบรับเมื่อ 17 สิงหาคม 2563 เผยแพร่ออนไลน์ 29 กรกฎาคม 2564

บทคัดย่อ

การวิจัยนี้เพื่อศึกษาเปรียบเทียบวิธีการประมาณค่าสูญหาย 3 วิธี คือ วิธี OLS วิธี MI และวิธี GA ในแผนแบบบล็อกสุ่มสมบูรณ์โดยใช้ข้อมูลจริง และข้อมูลจำลอง กำหนดให้ข้อมูลสูญหาย (m) อย่างสุ่มไม่เกิน 5% สำหรับข้อมูลจริงจำนวน 2 ชุด คือ t=5 b=3, t=8 b=4 และสำหรับข้อมูลจำลองมี 4 กรณี คือ t=3 b=3, t=5 b=3, t=8 b=4, t=8 b=5 กำหนดให้ C.V. เท่ากับ 10%, 30% และ 50% ให้ค่าคงที่ h เท่ากับ 1, 2 และ 3 และกระทำซ้ำแต่ละสถานการณ์ 12,000 รอบ โดยใช้โปรแกรม R และเมื่อพิจารณาจากค่า MSE และค่า MAPE ผลการศึกษาพบว่า เมื่อจำนวนข้อมูลสูญหายมีค่ามากขึ้น ทำให้ค่า MSE และ MAPE มีค่ามากขึ้น สำหรับข้อมูลจริงที่มี m เท่ากับ 1 และ 2 ค่า วิธี MI ให้ค่า MSE และ MAPE ต่ำที่สุด สำหรับข้อมูลจำลองที่มี m เท่ากับ 1 ส่วนใหญ่วิธี OLS ให้ค่า AVG MSE และ AVG MAPE ต่ำที่สุด ยกเว้นในกรณีที่ C.V. เท่ากับ 10% และ h เท่ากับ 1 ซึ่งพบว่า วิธี GA ให้ค่า AVG MSE และ AVG MAPE ต่ำที่สุด สำหรับข้อมูลจำลองที่ m เท่ากับ 2 เมื่อ C.V. เท่ากับ 10% และ 30% พบว่า วิธี MI และวิธี GA ให้ค่า AVG MSE และ AVG MAPE ต่ำที่สุด ตามลำดับ และเมื่อ C.V. เท่ากับ 50% ส่วนใหญ่วิธี OLS ให้ค่า AVG MSE และ AVG MAPE ต่ำที่สุด

คำสำคัญ: แผนแบบบล็อกสุ่มสมบูรณ์ ค่าสูญหาย วิธีกำลังสองน้อยที่สุด วิธี Multiple Imputation วิธี Genetic Algorithm

การอ้างอิงบทความ: ชญาดา แก้วชัยเจริญกิจ, บุญอ้อม โฉมที และ วันดี วณิชย์ศักดิ์พงศ์, “การเปรียบเทียบการประมาณค่าสูญหายในแผนแบบบล็อกสุ่มสมบูรณ์,” *วารสารวิชาการพระจอมเกล้าพระนครเหนือ*, ปีที่ 32, ฉบับที่ 2, หน้า 415–425, เม.ย.–มิ.ย. 2565.

Research Article

A Comparison of Missing Value Estimations in Randomized Complete Block Design

Chayada Kaewchaicharoenkit*, Boonorm Chomtee and Wandee Wanishsakpong
Department of Statistics, Faculty of Science, Kasetsart University, Bangkok, Thailand

* Corresponding Author, Tel. 09 4498 6258, E–mail: k.chayadaa@gmail.com
DOI: 10.14416/j.kmutnb.2021.07.006
Received 28 May 2020; Revised 14 July 2020; Accepted 17 August 2020; Published online: 29 July 2021

Abstract

The study aims to compare the 3 missing value estimations: OLS, MI, and GA in randomized complete block design using both real and simulated data. In the study, there are at least 5% of randomly missing value (m). Two real data sets : t=5 b=3 and t=8 b=4 are used in this research. Also, there are four cases of simulated data : t=3 b=3, t=5 b=3, t=8 b=4 and t=8 b=5. There are 10%, 30% and 50% of C.V., the constant h = 1, 2 and 3. R Program is used to simulate data with 12,000 times for each situation. MSE and MAPE are used as criteria for determination. For the result, when m increases, MSE and MAPE values tend to increase. For the real data, when m=1 and 2, MI has the lowest MSE and MAPE values. For the simulated data, when m=1, OLS has the lowest AVG MSE and AVG MAPE values for almost cases, except when C.V.=10%, h=1, GA has the lowest AVG MSE and AVG MAPE values. When m=2 C.V. = 10% and 30%, it is found that MI and GA have the lowest AVG MSE and AVG MAPE values respectively. When C.V. = 50%, for almost cases, OLS has the lowest AVG MSE and AVG MAPE values.

Keywords: Randomized Complete Block Design, Missing Value, Ordinary Least Square, Multiple Imputation, Genetic Algorithm

Please cite this article as: C. Kaewchaicharoenkit, B. Chomtee, and W. Wanishsakpong, "A comparison of missing value estimations in randomized complete block design," *The Journal of KMUTNB*, vol. 32, no. 2, pp. 415–425, Apr.–Jun. 2022 (in Thai).

1. บทนำ

ในปัจจุบันเทคโนโลยีมีความก้าวไกลทำให้การส่งข้อมูลข่าวสารมีความรวดเร็วมากขึ้น ส่งผลให้จำนวนข้อมูลก็มีมากขึ้น ความผิดพลาดจากการเก็บข้อมูลก็จะมากขึ้นด้วย ซึ่งมาจากหลายสาเหตุ เช่น การไม่ตอบคำถาม การตอบผิดจากความจริงทำให้เกิดข้อมูลสูญหาย

ข้อมูลสูญหาย คือ ค่าสังเกตที่ต้องการทราบค่า แต่ไม่สามารถทราบค่าได้สามารถพบได้บ่อยในงานวิจัย ผู้วิจัยจำเป็นต้องหาวิธีที่เหมาะสมสำหรับประมาณค่าสูญหาย หากเลือกใช้วิธีจัดการกับข้อมูลสูญหายที่ไม่เหมาะสมย่อมส่งผลทำให้เกิดการบิดเบือนผลการวิเคราะห์

การวางแผนการทดลองเกี่ยวข้องกับการวิเคราะห์ข้อมูลที่ได้จากการทดลอง โดยใช้วิธีทางสถิติให้เหมาะสม และสอดคล้องกับข้อมูล ซึ่งการวางแผนการทดลองได้นำมาใช้อย่างกว้างขวาง ในด้านการเกษตร ด้านอุตสาหกรรม และด้านการแพทย์ แผนการทดลองหนึ่งที่นิยมใช้ คือ แผนแบบบล็อกสุ่มสมบูรณ์ ซึ่งเป็นแผนการทดลองที่ศึกษาปัจจัย 1 ปัจจัยที่มีอิทธิพลต่อตัวแปรตอบสนอง (y) โดยจัดหน่วยทดลองเป็นกลุ่มๆ หรือบล็อก (b) และมีเงื่อนไขว่าหน่วยทดลองที่อยู่ในบล็อกเดียวกันจะต้องมีลักษณะเหมือนกัน ส่วนหน่วยทดลองที่อยู่ต่างบล็อกกันต้องมีลักษณะแตกต่างกัน และในแต่ละบล็อกมีหน่วยทดลองที่ได้รับทรีตเมนต์ครบทุกทรีตเมนต์ (t) ซึ่งจะมีผลทำให้ขนาดของแต่ละบล็อกเท่ากับจำนวนทรีตเมนต์ที่ศึกษา ทั้งนี้ถ้ามีข้อมูลสูญหายในแผนแบบบล็อกสุ่มสมบูรณ์ที่ไม่มีการทำซ้ำจะทำให้ไม่สามารถวิเคราะห์ข้อมูลได้ จึงต้องทำการประมาณค่าข้อมูลที่สูญหาย [1], [2]

การประมาณค่าข้อมูลสูญหายนั้น ทำได้หลายวิธีโดยวิธีที่นิยม คือ วิธีค่าเฉลี่ย หรือวิธีการวิเคราะห์การถดถอยเชิงพหุ และจากการศึกษางานวิจัยที่เกี่ยวข้องกับการประมาณค่าสูญหายพบว่า ประพจน์ [3] ศึกษาและเปรียบเทียบวิธีการประมาณค่าสูญหาย 3 วิธี ในแผนแบบบล็อกสุ่มสมบูรณ์ที่ไม่มีการทำซ้ำ ได้แก่ วิธีกำลังสองน้อยที่สุด (OLS) วิธีค่าคาดหวังสูงสุด และวิธี Multiple Imputation (MI) ผลการศึกษาพบว่า วิธี MI ให้ความคลาดเคลื่อนกำลังสองเฉลี่ย (Mean Square Error; *MSE*) ต่ำกว่าวิธีค่าคาดหวังสูงสุด และวิธี OLS ในทุกสถานการณ์ที่ศึกษา ศุภลักษณ์ [4] ศึกษาและเปรียบเทียบวิธีการประมาณค่าสูญหาย 3 วิธี ในแผนแบบจัตุรัสละติน ได้แก่ วิธี OLS วิธีค่าคาดหวังสูงสุด และวิธี MI ผลการศึกษาพบว่า วิธี MI ให้ค่า *MSE* ต่ำกว่าวิธีค่าคาดหวังสูงสุด และวิธี OLS ในทุกสถานการณ์ที่ทำศึกษา Azadeh และคณะ [5] ศึกษาการประมาณค่าสูญหายในแผนแบบบล็อกสุ่มสมบูรณ์ 3 วิธี คือ วิธี Genetic Algorithm (GA) วิธี Particle Swarm Optimization (PSO) และวิธี Artificial Neural Network (ANN) ผลการศึกษาพบว่า วิธี GA เป็นวิธีการประมาณค่าสูญหายที่ดีกว่าวิธีอื่นๆ เนื่องจากให้ค่าเฉลี่ยของเปอร์เซ็นต์ความคลาดเคลื่อนสัมบูรณ์ (Mean Absolute Percentage Error; *MAPE*) ต่ำที่สุด

จากที่ได้กล่าวมา ผู้วิจัยจึงสนใจศึกษาเปรียบเทียบวิธีการประมาณค่าข้อมูลสูญหายในแผนแบบบล็อกสุ่มสมบูรณ์ที่ไม่มีการทำซ้ำ โดยใช้วิธี OLS วิธี MI และวิธี GA โดยพิจารณาจากค่า *MSE* และ *MAPE* โดยวิธีที่ให้ค่า *MSE* และ *MAPE* ต่ำที่สุด จะเป็นวิธีที่ดีที่สุด

2. วัสดุ อุปกรณ์และวิธีการวิจัย

ในงานวิจัยนี้ใช้ข้อมูลจริง และข้อมูลจำลองของแผนแบบบล็อกสุ่มสมบูรณ์ที่ไม่มีการทำซ้ำ กำหนดให้จำนวนข้อมูลสูญหาย (m) อย่างสุ่มไม่เกิน 5% ของจำนวนหน่วยทดลองทั้งหมด (n) (ถ้า $m < 1$ จะกำหนดให้ $m = 1$) โดยใช้ข้อมูลจริงชุดที่ 1 จาก [2] เมื่อ t=5 b=3 และ m=1 แสดงดังตารางที่ 1 ข้อมูลชุดที่ 2 จาก [6] เมื่อ t=8 b=4 และ m=2 แสดงดังตารางที่ 2 สำหรับข้อมูลจำลองในที่นี้จำลองแต่ละสถานการณ์ (s) 12,000 รอบ โดยกำหนดให้ t=3 b=3, t=5 b=3, t=8 b=4, t=8 b=5 ซึ่งเมื่อกำหนดให้ m สูญหายอย่างสุ่มไม่เกิน 5% จึงทำให้ในการศึกษาครั้งนี้ m=1, 2 ในการเปรียบเทียบวิธีการประมาณค่าสูญหายสำหรับข้อมูลจริงจะพิจารณาจากค่า *MSE* และ *MAPE* และสำหรับข้อมูลจำลองจะพิจารณาจากค่าเฉลี่ยของ *MSE* (*AVG MSE*) และค่าเฉลี่ยของ *MAPE* (*AVG MAPE*)

ตัวแบบเชิงเส้นของแผนแบบบล็อกสุ่มสมบูรณ์ คือ

$$y_{ij} = \mu + \tau_i + \beta_j + \varepsilon_{ij} \; ; i = 1,2,...,t; j = 1,2,...,b$$

โดยที่ y_{ij} คือ ค่าสังเกตของทรีตเมนต์ที่ i และบล็อกที่ j

μ คือ ค่าเฉลี่ยรวม (Grand Mean)

τ_i คือ อิทธิพลของทรีตเมนต์ที่ i

β_j คือ อิทธิพลของบล็อกที่ j

ε_{ij} คือ ความคลาดเคลื่อนสุ่ม (Random Error) โดยที่ $\varepsilon_{ij} \sim NID(0,\sigma_\varepsilon^2)$

ในที่นี้พิจารณาให้อิทธิพลของทรีตเมนต์และบล็อกเป็นอิทธิพลแบบสุ่ม ดังนั้น

$y_{ij} \sim NID(\mu_y,\sigma_y^2)$

โดยที่ μ_y คือ ค่าเฉลี่ยของ y_{ij}

σ_y^2 คือ ความแปรปรวนของ y_{ij} และในการศึกษานี้กำหนดให้ $\sigma_\tau^2 = \sigma_\beta^2 = h\sigma_\varepsilon^2$ และ $h = 1,2,3$

ข้อมูลจริงที่ใช้สำหรับศึกษาการประมาณค่าสูญหายในแผนแบบบล็อกสุ่มสมบูรณ์

ข้อมูลชุดที่ 1 แสดงข้อมูลสำหรับแผนแบบบล็อกสุ่มสมบูรณ์ซึ่งประกอบด้วย 5 ทรีตเมนต์ 3 บล็อก แสดงข้อมูลดังตารางที่ 1

ตารางที่ 1 ค่าของตัวแปรตอบสนองข้อมูลชุดที่ 1

Block		1	2	3
Treatment	1	7.62	8.00	7.93
	2	8.14	8.15	7.87
	3	7.76	7.73	7.74
	4	7.17	7.57	7.80
	5	7.46	7.68	7.21

ข้อมูลชุดที่ 2 งานวิจัยนี้ทดลองผลผลิตข้าวสาลีโดยศึกษาน้ำหนักของข้าวสาลี (กิโลกรัมต่อแปลง) ที่ศูนย์วิจัยทางการเกษตร ประเทศปากีสถาน โดยใช้ปุ๋ยทั้งหมด 8 ชนิด ทำซ้ำจำนวน 4 ซ้ำ แสดงข้อมูลดังตารางที่ 2 ความแปรปรวนของความคลาดเคลื่อน

ความคลาดเคลื่อนของการทดลอง [2] คือ ความผันแปรที่เกิดขึ้นระหว่างหน่วงทดลองที่ได้รับทรีตเมนต์เดียวกัน งานวิจัยนี้กำหนดให้ข้อมูลจำลองมีค่าสัมประสิทธิ์ความผันแปร [Coefficient of Variance; $C.V.(y_{ij})$] ในระดับต่างๆ กัน คือ 10%, 30% และ 50% และ μ = 60 ดังนั้น

ตารางที่ 2 ข้อมูลน้ำหนักข้าวสาลี

Block		1	2	3	4
Treatment	1	2.67	3.52	2.10	2.61
	2	3.32	4.00	2.89	2.70
	3	4.37	3.15	3.77	2.16
	4	2.89	3.97	2.89	3.01
	5	2.78	2.55	2.64	4.17
	6	4.31	2.70	2.72	2.04
	7	3.12	4.82	3.43	3.41
	8	3.46	3.18	4.23	2.98

จากสมการที่ (1)

$$C.V.(y_{ij}) = \frac{S.D.(y_{ij})}{\mu} = \frac{\sqrt{\sigma_\tau^2 + \sigma_\beta^2 + \sigma_\varepsilon^2}}{60} \qquad (1)$$

และเมื่อกำหนดให้ $\sigma_\tau^2 = \sigma_\beta^2 = h\sigma_\varepsilon^2$ แทนลงในสมการที่ (1) โดยที่ h เป็นค่าคงที่ซึ่งเป็นจำนวนเต็มบวกเท่ากับ 1, 2 และ 3 ดังนั้น

$$C.V.(y_{ij}) = \frac{\sqrt{h\sigma_\varepsilon^2 + h\sigma_\varepsilon^2 + \sigma_\varepsilon^2}}{60} = \frac{\sigma_\varepsilon\sqrt{2h+1}}{60} \qquad (2)$$

จากสมการที่ (2) จะได้ว่า

$$\sigma_\varepsilon^2 = \frac{[C.V.(y_{ij})(60)]^2}{2h+1} \qquad (3)$$

2.1 ขั้นตอนการวิจัย

2.1.1 กำหนดค่า $C.V.(y_{ij})$ เท่ากับ 10%, 30% และ 50% μ = 60 และค่าคงที่ h เท่ากับ 1, 2 และ 3

2.1.2 กำหนดให้ σ_ε^2 เท่ากับ สมการที่ (3)

2.1.3 จำลองให้ $\tau_i \sim N(0,h\sigma_\varepsilon^2)$ $\beta_j \sim N(0,h\sigma_\varepsilon^2)$ และ $\varepsilon_{ij} \sim N(0,\sigma_\varepsilon^2)$ ภายใต้สมมติฐานว่างเป็นจริง

2.1.4 จำลองข้อมูลตามตัวแบบเชิงเส้นของแผนแบบบล็อกสุ่มสมบูรณ์

2.1.5 นำข้อมูลมาทำให้สูญหาย m ค่าอย่างสุ่ม โดยใช้วิธีการเลือกตัวอย่างสุ่มแบบง่าย [7]

2.1.6 ประมาณค่าสูญหายโดยใช้วิธีการประมาณค่า 3 วิธีที่ศึกษา คือ วิธี OLS วิธี MI และวิธี GA

2.1.7 คำนวณค่าในการเปรียบเทียบ โดยข้อมูลจริง คำนวณค่า MSE และ $MAPE$ และข้อมูลจำลอง คำนวณหาค่า $AVG\ MSE$ และ $AVG\ MAPE$

2.2 วิธี OLS

อลันและวิสชาร์ท (Allan และ Wishart: 1930) เป็นผู้ริเริ่มวิธีประมาณค่าข้อมูลสูญหาย และต่อมาเยทส์ (Yate: 1937) ได้แสดงว่าสูตรดังกล่าวได้มาจากวิธีกำลังสองน้อยที่สุด ซึ่งเป็นการประมาณค่าสูญหายโดยทำให้ผลรวมของกำลังสองของความคลาดเคลื่อน (SSE) มีค่าน้อยที่สุด [2], [3] สามารถคำนวณได้ดังสมการที่ (4)

$$\hat{y}_{ij} = \frac{ty'_{i.} + by'_{.j} - y'_{..}}{(t-1)(b-1)} \quad (4)$$

เมื่อ $\hat{y}_{ij}$ คือ ค่าประมาณใหม่จากทรีตเมนต์ที่ i และบล็อกที่ j

t คือ จำนวนทรีตเมนท์

b คือ จำนวนบล็อก

$y'_{i.}$ คือ ผลรวมของข้อมูลในทรีตเมนต์ที่มีค่าสูญหาย

$y'_{.j}$ คือ ผลรวมของข้อมูลในบล็อกที่มีค่าสูญหาย

$y'_{..}$ คือ ผลรวมของข้อมูลทั้งหมดที่มีค่าสูญหาย

2.3 วิธี MI

รูบิน (Rubin: 1978) ได้เสนอวิธี MI [4] เป็นครั้งแรกเพื่อนำมาใช้แก้ปัญหากรณีที่ข้อมูลมีการสูญหายซึ่งข้อมูลที่สูญหายนั้นเป็นไปอย่างสุ่ม (Missing at Random; MAR) โดยถ้าวิธีใดให้ค่า MSE และค่า $MAPE$ ต่ำกว่าวิธีนั้นจะถูกใช้เป็นตัวประมาณของวิธี MI ประกอบด้วย 2 วิธี ดังนี้

2.3.1 วิธีค่าเฉลี่ย (Mean Imputation)

วิธีค่าเฉลี่ย [8] เป็นวิธีแทนค่าสูญหายด้วยค่าเฉลี่ยของข้อมูลที่มีอยู่ โดยคำนวณค่าเฉลี่ยของข้อมูลจากสมการที่ (5)

$$\bar{y}_{ij} = \frac{\sum_{i=1}^{n-m}\sum_{j=1}^{n-m} y_{ij}}{n-m} \quad (5)$$

เมื่อ $\bar{y}_{ij}$ คือ ค่าเฉลี่ยของข้อมูลที่ไม่สูญหาย

y_{ij} คือ ข้อมูลจากทรีตเมนต์ที่ i และบล็อกที่ j โดยที่ $i = 1, 2, ..., t$; $j = 1, 2, ..., b$

n คือ จำนวนหน่วยทดลองทั้งหมด

m คือ จำนวนข้อมูลสูญหาย

2.3.2 วิธี EM (Expectation Maximization)

วิธี EM [3], [9] เป็นกระบวนการหาค่าประมาณพารามิเตอร์ซึ่งในที่นี้คือค่าข้อมูลสูญหาย แบบภาวะน่าจะเป็นสูงสุด (Maximum Likelihood) โดยมีการวนซ้ำ สมมติให้ข้อมูลทั้งหมดจำนวน n ค่า ค่าสูญหายจำนวน m ค่า ดังนั้นข้อมูลที่มีอยู่ $n-m$ ค่า โดยการกำหนด $\mu^{(t)} = \mu^{(t+1)} = \hat{\mu}$ และ $\sigma^{(t)} = \sigma^{(t+1)} = \hat{\sigma}$ ในสมการที่ (8)–(11)

$$\hat{\mu} = \frac{\sum_{i=1}^{n-m}\sum_{j=1}^{n-m} y_{ij}}{n-m} \quad (6)$$

$$\hat{\sigma}^2 = \frac{\sum_{i=1}^{n-m}\sum_{j=1}^{n-m} y_{ij}^2}{n-m} - \hat{\mu}^2 \quad (7)$$

ซึ่งวิธี EM สามารถแบ่งออกเป็น 2 ขั้นตอน คือ

ขั้นตอนที่ 1 E-step เป็นขั้นตอนในการหาค่าคาดหวังของค่าข้อมูลสูญหายภายใต้เงื่อนไขของชุดข้อมูลที่ไม่มีการสูญหาย และพารามิเตอร์ตัวปัจจุบัน ($\theta^{(t)}$) เพื่อนำค่าคาดหวังที่ได้นี้ไปประมาณค่าข้อมูลที่สูญหาย

$$E\left(\sum_{i=1}^{n}\sum_{j=1}^{n} y_{ij} \mid \theta^{(t)}, Y_{obs}\right) = \sum_{i=1}^{n-m}\sum_{j=1}^{n-m} y_{ij} + (m)\mu^{(t)} \quad (8)$$

$$E\left(\sum_{i=1}^{n}\sum_{j=1}^{n} y_{ij}^2 \mid \theta^{(t)}, Y_{obs}\right) = \sum_{i=1}^{n-m}\sum_{j=1}^{n-m} y_{ij}^2 + (m)\left[\left(\mu^{(t)}\right)^2 + \left(\sigma^{(t)}\right)^2\right] \quad (9)$$

ขั้นตอนที่ 2 M-step เป็นขั้นตอนการประมาณค่าพารามิเตอร์ภาวะน่าจะเป็นสูงสุดจากข้อมูลที่มีอยู่ ซึ่งเป็นการแทนค่าสูญหายที่ได้จากขั้นตอน E-step

$$\mu^{(t+1)} = \frac{E\left(\sum_{i=1}^{n}\sum_{j=1}^{n} y_{ij} \mid \theta^{(t)}, Y_{obs}\right)}{n} \quad (10)$$

$$(\sigma^{(t+1)})^2 = \frac{E\left(\sum_{i=1}^{n}\sum_{j=1}^{n} y_{ij}^2 \mid \theta^{(t)}, Y_{obs}\right)}{n} - (\mu^{(t+1)})^2 \quad (11)$$

โดยที่ Y_{obs} คือ หน่วยทดลองที่เก็บรวบรวมได้

$\sum_{i=1}^{n-m}\sum_{j=1}^{n-m} y_{ij}$ คือ ผลรวมของข้อมูลที่มีอยู่

$\sum_{i=1}^{n-m}\sum_{j=1}^{n-m} y_{ij}^2$ คือ ผลรวมของข้อมูลที่มีอยู่กำลังสอง

$\mu^{(t)}$ คือ ค่าเฉลี่ยในรอบที่ t

$\mu^{(t+1)}$ คือ ค่าเฉลี่ยในรอบที่ $t+1$

$(\sigma^{(t)})^2$ คือ ความแปรปรวนในรอบที่ t

$(\sigma^{(t+1)})^2$ คือ ความแปรปรวนในรอบที่ $t+1$

ทำซ้ำในสมการที่ (10) และ (11) จนกระทั่งได้ค่าประมาณเข้าใกล้ค่าในสมการที่ (6) และ (7) ตามลำดับ จากนั้นทำการสุ่มค่าประมาณข้อมูลสูญหาย (y_{miss}) คือ $y_{miss} \sim NID(\mu^{(t+1)}, (\sigma^{(t+1)})^2)$ ตามจำนวนข้อมูลที่สูญหาย

2.4 วิธี GA

วิธี GA [5], [10] คิดค้นโดย John Holland เป็นเทคนิคการหาค่าประมาณของคำตอบของปัญหา ในที่นี้คือค่าข้อมูลสูญหาย โดยอาศัยหลักการจากทฤษฎีวิวัฒนาการทางชีววิทยา และการคัดเลือกธรรมชาติ

วิธี GA เป็นการจำลองทางคอมพิวเตอร์เพื่อหาค่าคำตอบที่เหมาะสมที่สุดโดยการแทนข้อมูลที่มีอยู่ให้อยู่ในสัญลักษณ์โครโมโซม ตามขั้นตอนดังนี้

2.4.1 การกำหนดค่าเริ่มต้น (Initialization)

เป็นการสร้างประชากรก่อนจะเข้ากระบวนการวิธี GA ค่าที่ได้จะอยู่ในขอบเขตของคำตอบ ตามจำนวนประชากรที่กำหนดไว้

2.4.2 การคัดเลือก (Selection)

เป็นวิธีการประเมินค่าความเหมาะสมที่เป็นไปได้ของโครโมโซมว่าโครโมโซมไหนเหมาะสมที่จะสืบทอดต่อไปโดยเกณฑ์ที่ใช้ในการคัดเลือกข้อมูล คือ ค่า MSE ถ้าข้อมูลชุดไหนให้ MSE ต่ำที่สุด จะถูกเลือกให้เป็น Elitism ซึ่งคือการเก็บโครโมโซมที่ดีที่สุดจากโครโมโซมพ่อแม่ เพื่อเป็นการยืนยันว่าจะได้โครโมโซมที่เหมาะสมที่สุดในการประเมินสำหรับรุ่นต่อไป

2.4.3 การผลิตรุ่นต่อไป (New Population)

เป็นการปรับเปลี่ยนตามองค์ประกอบของวิธี GA โดยวิธีที่นิยมใช้มีดังนี้

1) การผสม (Blending) คือการผสมระหว่างโครโมโซมพ่อและแม่ สมมติให้ A_a และ B_b เป็นค่าที่มาจากวิธีการคัดเลือกว่าให้โครโมโซมไหนเป็น Elitism ถ้าความน่าจะเป็นของการทดสอบผ่าน ค่า A_a และ B_b จะผสมกัน โดยในรุ่นลูกหลานจะได้ค่าใหม่ตามสมการที่ (12) และ (13) ดังนี้

$$A_a^* = \beta A_a + (1-\beta) B_b \quad (12)$$

$$B_b^* = (1-\beta) A_a + \beta B_b \quad (13)$$

โดย A_a^* และ B_b^* เป็นโครโมโซมลูกหลาน

2) การสลับสายพันธุ์ (Crossover) เป็นกระบวนการที่สำคัญของวิธี GA เมื่อเกิดการสลับสายพันธุ์ขึ้นในทางพันธุกรรมจะทำให้เกิดการเปลี่ยนแปลงซึ่งในงานวิจัยนี้กำหนดเกณฑ์ในการผสม และการสลับสายพันธุ์ โดยที่ให้ β คือ ความน่าจะเป็นของเกณฑ์การผสมและการสลับสายพันธุ์ดังนี้

ถ้า $\beta \geq 20$ ไม่ต้องทำขั้นตอนนี้

$\beta < 20$ ต้องทำขั้นตอนนี้

2.4.4 การจบการทำงาน (End)

ทำการวนซ้ำในขั้นตอนการคัดเลือก โดยกระบวนการจะจบสิ้นเมื่อขั้นตอนการเลือก Elitism ของรอบใหม่ได้ประชากรเริ่มต้นเหมือนประชากรเริ่มต้นของรอบก่อนหน้านี้ ซึ่งในที่นี้คือการได้ค่าประมาณของข้อมูลสูญหายคงที่

2.5 เกณฑ์ในการเปรียบเทียบ

ในการศึกษานี้จะพิจารณาเกณฑ์การเปรียบเทียบประสิทธิภาพของวิธีการประมาณค่าสูญหาย 2 วิธี คือ MSE และ $MAPE$ โดยที่

1) ค่า MSE เป็นวิธีที่พิจารณาความแตกต่างระหว่างค่าจริงและค่าประมาณโดยวิธียกกำลังสอง สามารถคำนวณดังสมการที่ (14)

$$MSE = \frac{\sum_{i=1}^{t}\sum_{j=1}^{b}\left(y_{ij} - \hat{y}_{ij}\right)^2}{n} \quad (14)$$

สำหรับข้อมูลจำลอง เนื่องจากกระทำซ้ำ 12,000 รอบในแต่ละสถานการณ์ จึงจะใช้ค่า *AVG MSE* ซึ่งคำนวณดังนี้

2) ค่า *MAPE* เป็นวิธีการคำนวณเปอร์เซ็นต์ความผิดพลาดในการประมาณค่าสูญหาย สามารถคำนวณดังสมการที่ (15)

$$MAPE = \frac{100}{n}\sum_{i=1}^{t}\sum_{i=1}^{b}\left|\frac{y_{ij} - \hat{y}_{ij}}{y_{ij}}\right| \quad (15)$$

สำหรับข้อมูลจำลอง เนื่องจากกระทำซ้ำ 12,000 รอบในแต่ละสถานการณ์ จึงจะใช้ค่า *AVG MAPE* ซึ่งคำนวณดังสมการที่ (16)

$$AVG\ \ MAPE = \frac{\sum_{s=1}^{12000} MAPE}{12000} \quad (16)$$

โดยที่ y_{ij} คือ ค่าจริงของข้อมูลจากทรีตเมนต์ที่ i และบล็อกที่ j

$\hat{y}_{ij}$ คือ ค่าประมาณของข้อมูลที่สูญหายจากทรีตเมนต์ที่ i และบล็อกที่ j

n คือ จำนวนหน่วยทดลอง

ในการเปรียบเทียบวิธีการประมาณค่าข้อมูลสูญหายสำหรับข้อมูลจริงจะพิจารณาจากค่าเกณฑ์ *MSE* และ *MAPE* และสำหรับข้อมูลจำลองจะพิจารณาจากค่าเกณฑ์ *AVG MSE* และ *AVG MAPE* โดยวิธีใดให้ค่าเกณฑ์ดังกล่าวต่ำกว่า วิธีนั้นจะเป็นวิธีที่ดีกว่าดังสมการที่ (17)

$$AVG\ MSE = \frac{\sum_{s=1}^{12000} MSE}{12000} \quad (17)$$

3. ผลการทดลอง

ผลการศึกษาเปรียบเทียบวิธีการประมาณค่าสูญหาย 3 วิธี คือ วิธี OLS วิธี MI และวิธี GA สำหรับข้อมูลจริงจะพิจารณาจากค่า *MSE* และ *MAPE* แสดงผลในตารางที่ 3 และสำหรับข้อมูลจำลองจะพิจารณาจากค่า *AVG MSE* และ *AVG MAPE* แสดงผลในตารางที่ 4 และ 5 สำหรับ m เท่ากับ 1 และ 2 ตามลำดับ

สำหรับข้อมูลจริงชุดที่ 1 ที่มี m เท่ากับ 1 พบว่า วิธี MI และวิธี GA ให้ค่า *MSE* และ *MAPE* เท่ากัน และข้อมูลจริงชุดที่ 2 ที่มี m เท่ากับ 2 วิธี MI ให้ค่า *MSE* และ *MAPE* ต่ำที่สุด ดังแสดงผลในตาราง 3

สำหรับข้อมูลจำลองที่มี m เท่ากับ 1 ที่ *C.V.* เท่ากับ 10% สำหรับค่า h เท่ากับ 1 พบว่า วิธี GA ให้ค่า *AVG MSE* และ *AVG MAPE* ต่ำที่สุด สำหรับค่า h เท่ากับ 2 วิธี OLS ให้ค่า *AVG MSE* ต่ำที่สุด สำหรับแผนแบบ t=3 b=3 วิธี GA ให้ค่า *AVG MAPE* ต่ำที่สุด และแผนแบบ t=5 b=3 วิธี OLS ให้ค่า *AVG MAPE* ต่ำที่สุด และสำหรับค่า h เท่ากับ 3 วิธี OLS ให้ค่า *AVG MSE* และ *AVG MAPE* ต่ำที่สุด ที่ *C.V.* เท่ากับ 30% สำหรับค่า h เท่ากับ 1 พบว่า วิธี GA ให้ค่า *AVG MSE* ต่ำที่สุด สำหรับแผนแบบ t=3 b=3 วิธี GA ให้ค่า *AVG MAPE* ต่ำที่สุด และแผนแบบ t=5 b=3 วิธี OLS ให้ค่า *AVG MAPE* ต่ำที่สุด และในกรณี *C.V.* เท่ากับ 50% สำหรับค่า h เท่ากับ 1 สำหรับแผนแบบ t=3 b=3 วิธี GA ให้ค่า *AVG MSE* ต่ำที่สุด ในขณะที่แผนแบบ t=5 b=3 วิธี OLS ให้ค่า *AVG MSE* และ *AVG MAPE* ต่ำที่สุดและสำหรับ *C.V.* เท่ากับ 30% และ 50% ค่า h เท่ากับ 2 และ 3 วิธี OLS ให้ค่า *AVG MSE* และ *AVG MAPE* ต่ำที่สุด ดังแสดงผลในตารางที่ 4

สำหรับข้อมูลจำลองที่มี m เท่ากับ 2 ในกรณีที่ *C.V.* เท่ากับ 10% วิธี MI ให้ค่า *AVG MSE* และ *AVG MAPE* ต่ำที่สุด ในกรณีที่ *C.V.* เท่ากับ 30% วิธี GA ให้ค่า *AVG MSE* และ *AVG MAPE* ต่ำที่สุด และในกรณีที่ *C.V.* เท่ากับ 50% สำหรับแผนแบบ t=8 b=4 ที่ค่า h เท่ากับ 1, 2 และ 3 และแผนแบบ t=8 b=5 ที่ค่า h เท่ากับ 1 วิธี GA ให้ค่า *AVG MSE* ที่สุดในขณะที่แผนแบบ t=8 b=5 ค่า h เท่ากับ 2 และ 3 วิธี OLS ให้ค่า *AVG MSE* ต่ำที่สุด ส่วนวิธี OLS ให้ค่า *AVG MAPE* ต่ำที่สุดที่ค่า h เท่ากับ 1, 2 และ 3 ดังแสดงผลในตารางที่ 5

ตารางที่ 3 ค่า MSE และ MAPE ของข้อมูลจริงชุดที่ 1 (t=5 b=3 m=1) และชุดที่ 2 (t=8 b=4 m=2)

Data set	Method	MSE	MAPE
1	OLS	0.0032	0.1897
	MI	**0.0001**	**0.0074**
	GA	**0.0001**	**0.0074**
2	OLS	0.3464	3.5831
	MI	**0.0394**	**1.1395**
	GA	0.0449	1.1892

ตารางที่ 4 ค่า AVG MSE และ AVG MAPE ของข้อมูลจำลอง เมื่อ m = 1

No. of Treatment			3		5	
No. of Block			3		3	
Criteria			AVG MSE	AVG MAPE	AVG MSE	AVG MAPE
C.V. = 10%	h = 1	OLS	2.9499	0.7697	1.4980	0.4265
		MI	3.6927	0.8682	2.1651	0.5181
		GA	**2.3857**	**0.6472**	**1.3168**	**0.3731**
	h = 2	OLS	**1.8314**	**0.6059**	**0.8954**	**0.3290**
		MI	3.6122	0.8589	2.1082	0.5060
		GA	1.9845	**0.5778**	1.1597	0.3370
	h = 3	OLS	**1.2669**	**0.5050**	**0.6155**	**0.2732**
		MI	3.4644	0.8400	2.0796	0.5038
		GA	1.7431	0.5350	1.0088	0.3083
C.V. = 30%	h = 1	OLS	27.0145	2.8487	13.7388	**1.5889**
		MI	33.9416	3.6675	20.0823	2.1791
		GA	**23.3647**	**2.6982**	**13.0797**	1.6045
	h = 2	OLS	**16.1989**	**2.1607**	**8.2499**	**1.1828**
		MI	32.2529	3.6004	19.3432	2.1011
		GA	18.9538	2.5047	10.9529	1.3438
	h = 3	OLS	**11.6781**	**1.7837**	**5.9152**	**1.0236**
		MI	31.8143	3.3510	18.6718	2.0562
		GA	17.4289	2.0969	9.9883	1.2884
C.V. = 50%	h = 1	OLS	73.2364	**11.3866**	**36.8992**	**6.5438**
		MI	92.4485	22.8183	54.5018	11.3594
		GA	**65.1333**	13.1341	37.0431	7.7200
	h = 2	OLS	44.1003	8.6920	22.4228	3.7489
		MI	91.5892	16.5454	52.7621	9.1029
		GA	54.6176	11.5285	32.2879	5.5669
	h = 3	OLS	**31.4528**	**6.0751**	**16.3742**	**3.3222**
		MI	87.0173	16.8383	50.9546	8.8223
		GA	47.7520	10.4722	29.3623	5.6829

ตารางที่ 5 ค่า AVG MSE และ AVG MAPE ของข้อมูลจำลอง เมื่อ m = 2

No. of Treatment			8		8	
No. of Block			4		5	
Criteria			AVG MSE	AVG MAPE	AVG MSE	AVG MAPE
C.V. = 10%	h = 1	OLS	17.9901	1.4332	8.5915	0.9084
		MI	**2.1285**	**0.4934**	**1.6897**	**0.3958**
		GA	2.3318	0.5138	1.7914	0.4060
	h = 2	OLS	17.6890	1.4337	8.4671	0.9095
		MI	**2.0477**	**0.4869**	**1.6555**	**0.3902**
		GA	2.2407	0.5062	1.7682	0.4014
	h = 3	OLS	17.7753	1.4352	8.4103	0.9098
		MI	**2.0321**	**0.4805**	**1.6348**	**0.3876**
		GA	2.1978	0.4974	1.7487	0.3994
C.V. = 30%	h = 1	OLS	30.9845	2.1879	18.9316	1.7119
		MI	18.6711	1.9995	15.2740	2.0652
		GA	**17.5258**	**1.8243**	**13.8692**	**1.6491**
	h = 2	OLS	29.6513	2.2167	17.3980	7.1264
		MI	18.8008	1.9677	14.9961	7.2041
		GA	**17.7293**	**1.8053**	**13.5413**	**7.0938**
	h = 3	OLS	28.4037	2.3614	16.9985	1.6222
		MI	18.2484	2.0816	15.0480	1.7129
		GA	**17.4202**	**1.9454**	**13.5638**	**1.4654**
C.V. = 50%	h = 1	OLS	57.5540	14.5769	39.2671	7.2421
		MI	52.5626	20.0213	41.8006	10.2743
		GA	**48.2743**	15.8406	**36.7868**	8.4697
	h = 2	OLS	51.4079	**9.3169**	**34.9731**	**6.7770**
		MI	51.0285	12.0374	41.1198	9.9115
		GA	**45.6167**	9.9236	35.6739	7.6372
	h = 3	OLS	49.7538	**22.4715**	**33.1593**	**6.0574**
		MI	50.5838	28.1397	40.7775	9.0661
		GA	**45.2017**	24.3254	35.3846	7.5572

4. อภิปรายผลและสรุป

งานวิจัยนี้ศึกษา และเปรียบเทียบการประมาณค่าสูญหายในแผนแบบบล็อกสุ่มสมบูรณ์ 3 วิธี คือ วิธี OLS วิธี MI และวิธี GA ข้อมูลที่ใช้ในการทดสอบเป็นข้อมูลจริงและข้อมูลจำลอง โดยมีค่า m เท่ากับ 1 และ 2 เกณฑ์ที่ใช้ในการเปรียบเทียบ คือ ค่า *MSE* และ *MAPE* สำหรับข้อมูลจริงและค่า *AVG MSE* และ *AVG MAPE* สำหรับข้อมูลจำลอง ผลการศึกษาพบว่า ถ้าจำนวนหน่วยทดลอง ค่า *C.V.* และค่า h มีค่ามากขึ้น ความแปรปรวนของข้อมูลก็จะมีค่ามากขึ้นด้วย แต่ในทางกลับกัน ถ้าจำนวนหน่วยทดลองในแต่ละกรณีของจำนวนข้อมูลสูญหาย (m=1,2) เพิ่มขึ้นและให้ค่า *C.V.* คงที่ ค่า h เพิ่มขึ้น ความแปรปรวนของข้อมูลก็จะมีค่าลดลง สำหรับข้อมูลจริงที่มีค่า m เท่ากับ 1 วิธี MI และวิธี GA ให้ค่า *MSE* และ *MAPE* ต่ำที่สุดและเมื่อมีค่า m เท่ากับ 2 วิธี MI ให้ค่า *MSE* และ *MAPE* ต่ำที่สุด กรณีข้อมูลจำลองที่มีค่า m เท่ากับ 1 เมื่อค่าคงที่ h มีค่าน้อย (h=1) วิธี GA ให้ค่า *AVG MSE*

และ *AVG MAPE* ต่ำที่สุด และเมื่อค่าคงที่ h มีค่ามากขึ้น (h=2, 3) วิธี OLS ให้ค่า *AVG MSE* และ *AVG MAPE* ต่ำที่สุดทุกสถานการณ์ และเมื่อมีค่า m เท่ากับ 2 ที่ *C.V.* มีค่าน้อย (*C.V.* = 10%) วิธี MI ให้ค่า *AVG MSE* และ *AVG MAPE* ต่ำที่สุด และเมื่อ *C.V.* มีค่ามากขึ้น (*C.V.* = 30%, 50%) วิธี GA และวิธี OLS มีแนวโน้มให้ค่า *AVG MSE* และ *AVG MAPE* ต่ำที่สุดทุกสถานการณ์

สำหรับผลการศึกษาข้อมูลจริงและข้อมูลจำลองเมื่อพิจารณาจากเกณฑ์ *MSE*, *MAPE*, *AVG MSE* และ *AVG MAPE* พบว่า วิธีการประมาณค่าแต่ละวิธีให้ผลการศึกษาที่แตกต่างกันในบางสถานการณ์ โดยเมื่อ m เท่ากับ 1 แผนแบบ t=5 b=3 สำหรับข้อมูลจริงและข้อมูลจำลองในกรณี *C.V.* เท่ากับ 10% ค่า h เท่ากับ 1 วิธี GA ให้ค่าความคลาดเคลื่อนของข้อมูลต่ำที่สุดแต่เมื่อค่า *C.V.* และค่า h มีค่ามากขึ้น วิธี OLS ให้ค่าความคลาดเคลื่อนของข้อมูลต่ำที่สุด และเมื่อ m เท่ากับ 2 แผนแบบ t=8 b=4 สำหรับข้อมูลจริง และข้อมูลจำลองในกรณี *C.V.* เท่ากับ 10% ค่า h เท่ากับ 1, 2 และ 3 วิธี MI ให้ค่าความคลาดเคลื่อนของข้อมูลต่ำที่สุดแต่เมื่อค่า *C.V.* มากขึ้น ส่วนใหญ่วิธี GA และวิธี OLS จะให้ค่าความคลาดเคลื่อนของข้อมูลต่ำที่สุด

จากขอบเขตในการศึกษาครั้งนี้ บางสถานการณ์ที่ผลการศึกษาจากการพิจารณาค่า *MSE* และ *MAPE* หรือค่า *AVG MSE* และ *AVG MAPE* ไม่สอดคล้องกันนั้น ผู้วิจัยเห็นว่าเกณฑ์ที่เหมาะสมที่จะสรุปว่าวิธีไหนดีที่สุดสำหรับการประมาณค่าสูญหาย คือ ค่า *MAPE* เนื่องจากเป็นวิธีที่วัดความแม่นยำโดยคำนวณเปอร์เซ็นความผิดพลาดในการประมาณค่าสูญหายโดยไม่คำนึงถึงเครื่องหมาย เช่นถ้า *MAPE* เท่ากับ 4% แสดงว่าวิธีนั้นมีความคลาดเคลื่อนในการประมาณค่าสูญหายอยู่ที่ร้อยละ 4 [11]

เอกสารอ้างอิง

[1] P. Prasitwattanaseree and S. Prasitwattanaseree, "Missing data and management," *Data Management and Biostatistics Journal*, vol. 4, no. 3, 2009 (in Thai).

[2] B. Chomtee, *Statistics for Experimental design : Theory and Analysis with SAS*. Bangkok: Department of Statistics , Faculty of Science, Kasetsart University, 2013 (in Thai).

[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, Bangkok, 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, Bangkok, 2006 (in Thai).

[5] A. Azadeh, S.M. Asadzadeh, R. Jafari-Marandi, S. Nazari-Shirkouhi, G.Baharian Khoshkhou, S. Talebi, and A. Naghavi, "Optimum estimation of missing values in randomized complete block design by genetic algorithm," *Journal of Knowledge-Based Systems*, vol. 37, pp. 37–47, 2012.

[6] S. A. A. Shah, Alamgir, and M. Khan, "Comparative efficiency of randomized complete block design Vs. latin square design in wheat yield trial," *Journal of Natural Sciences Research*, vol. 7, no. 1, pp. 22–25, 2017.

[7] H. Pakcharoen. (2012, March). *Sampling Techniques*. [Online]. Available: http://service.nso.go.th/nso/nsopublish/Toneminute/files/55/0203-5.pdf

[8] S. Wongsriya, "A comparison of treatment sum of squares for Greco Latin square design with one missing observation," M.E. thesis, Department of Engineering in Industrial Engineering, Faculty

of Engineering, King Mongkut's Institute of Technology Ladkrabang, Bangkok, 2018 (in Thai).

[9] R. Lamjaisue, A. Thongteeraparp, and J. Sinsomboonthong, "Comparison of missing data estimation methods for the multiple regression analysis with missing at random dependent variable," *Journal of Science and Technology*, vol. 25, no. 5, 2017 (in Thai).

[10] W. Limmun, John J. Borkowski, and B. Chomtee, "Using a genetic algorithm to generate d-optimal designs for mixture experiments," *Journal of Quality and Reliability Engineering International*, vol. 35, pp. 1055–1068, 2012.

[11] R. Sriprakho, "Reduction of product shortages by using Forecasting techniques: A case study Aicello (Thailand) Co., Ltd." M.E. thesis, Department of Engineering Management, Faculty of Engineering Dhurakij Pundit University, 2014 (in Thai).