

การเปรียบเทียบประสิทธิภาพการประมาณค่าเฉลี่ยจากการประมาณค่า สุญหายด้วยตัวแปรเสริมภายใต้กลไกการสุญหายหลายแบบ

ณัทย สระกบแก้ว¹ คณิศา โชติจันทิก^{2*} และ คณารักษ์ ศรีสมบูรณ์³

¹กองวิชาคณิตศาสตร์และคอมพิวเตอร์ ส่วนการศึกษา โรงเรียนนายร้อยพระจุลจอมเกล้า

²ภาควิชาคณิตศาสตร์ สถิติ และคอมพิวเตอร์ คณะวิทยาศาสตร์ มหาวิทยาลัยอุบลราชธานี

³หมวดวิทยาศาสตร์และเทคโนโลยี คณะเทคโนโลยีและนวัตกรรมดิจิทัล มหาวิทยาลัยนอร์ทกรุงเทพ

¹99 หมู่ 1 ตำบลพรหมณี อำเภอเมือง จังหวัดนครนายก 26001

²85 ถนนสกลมารค์ อำเภอวารินชำราบ จังหวัดอุบลราชธานี 34190

³6/999 พหลโยธิน 52 แขวงคลองถนน เขตสายไหม กรุงเทพมหานคร 10220

รับบทความ 30 มิถุนายน 2566 แก้ไขบทความ 15 กรกฎาคม 2567 ตอบรับบทความ 24 กรกฎาคม 2567

บทคัดย่อ

งานวิจัยนี้มีวัตถุประสงค์เพื่อเปรียบเทียบประสิทธิภาพการประมาณค่าเฉลี่ยเมื่อข้อมูลสุญหาย ตัวประมาณค่าเฉลี่ยมาจากการประมาณค่าสุญหาย ได้แก่ วิธีค่าเฉลี่ย วิธีแฟคเตอร์ และวิธีเลขชี้กำลัง ภายใต้กลไกการสุญหาย 3 รูปแบบ คือ การสุญหายแบบสุ่ม การสุญหายแบบสุ่มอย่างสมบูรณ์ และการสุญหายแบบไม่สุ่ม งานวิจัยนี้ใช้ข้อมูลฝุ่นละอองขนาดเล็กจากแหล่งข้อมูลของกรมควบคุมมลพิษในการศึกษาประสิทธิภาพการประมาณค่าเฉลี่ยด้วยวิธีการประมาณค่าสุญหายในแต่ละสถานการณ์ สำหรับเกณฑ์ที่ใช้ในการเปรียบเทียบประสิทธิภาพ คือ ค่าความคลาดเคลื่อนกำลังสองเฉลี่ย โดยกำหนดขนาดตัวอย่าง คือ 30 100 และ 500 ที่ระดับการสุญหาย 10% 20% และ 30% ผลการวิจัยพบว่า ภายใต้กลไกการสุญหายทุกรูปแบบประสิทธิภาพของการประมาณค่าเฉลี่ยจากวิธีที่ใช้ข้อมูลตัวแปรเสริมร่วมกับตัวแปรที่สนใจศึกษา คือ วิธีแฟคเตอร์ และวิธีเลขชี้กำลัง ดีกว่า วิธีค่าเฉลี่ยที่ใช้ข้อมูลตัวแปรที่สนใจศึกษาเพียงอย่างเดียว นอกจากนี้ขนาดตัวอย่างและระดับการสุญหายของข้อมูลมีผลต่อประสิทธิภาพของวิธีการ เมื่อขนาดตัวอย่างเพิ่มขึ้นประสิทธิภาพของวิธีการเพิ่มสูงขึ้น แต่ถ้ระดับการสุญหายเพิ่มขึ้นประสิทธิภาพของวิธีการลดลง

คำสำคัญ : การประมาณค่าเฉลี่ย; ข้อมูลสุญหาย; ตัวแปรเสริม

The Performance Comparison of Mean Estimation by Imputation with Auxiliary Variable under Various Missing Data Mechanisms

Nahathai Srakobkaew¹ Kanisa Chodjuntug^{2*} and Kanarak Srisomboon³

¹Department of Mathematics and Computer Science, Academic Division, Chulachomklao Royal Military Academy

²Department of Mathematics Statistics and Computer, Faculty of Science, Ubon Ratchathani University

³Science and Technology Program, Faculty of Information Technology and Digital Innovation, North Bangkok University

¹99 Moo 1, Prommanee, Mueang, Nakhon Nayok 26001

²85 Sathonlamark Road, Warin Chamrap District, Ubon Ratchathani Province 34190

³6/999 Phahonyothin 52, Khlong Thanon Subdistrict, Sai Mai District, Bangkok 10220

Received 30 June 2023; Revised 15 July 2024; Accepted 24 July 2024

Abstract

The objective of this research is to compare the performance of mean estimation when data is missing. Mean estimators are derived from imputation methods, including the mean method, factor method, and exponential method. Under missing data mechanisms, there are 3 types, including missing completely at random, missing at random, and missing not at random. This research uses the small dust particles from the Pollution Control Department's sources to study the efficiency of mean estimation using the imputation method for each situation. The criterion used to compare efficiency is the mean square error. The sample sizes were set to be 30, 100, and 500 at missing data levels of 10%, 20%, and 30%. The results showed that under all types of missing data mechanisms, the efficiency of mean estimation from the imputation method that uses information from an auxiliary variable together with the study variable as the factor method and the exponent method was better than the mean method that uses the study variable only. Moreover, the sample size and the missing data levels affect the performance of the method. As the sample size increases, the efficiency of the method increases. But if the missing data levels increase, the efficiency of the method decreases.

Keywords : Mean Estimation; Missing data; Auxiliary Variable

** Corresponding Author. Tel.: +66 4535 3401, E-mail Address: kanisa.c@ubu.ac.th*

1. บทนำ

ข้อมูลสูญหายเป็นปัญหาที่เกิดขึ้นบ่อยครั้งในงานวิจัยด้านต่าง ๆ เช่น สิ่งแวดล้อม มลพิษทางอากาศ เศรษฐกิจ การแพทย์ การเกษตร เป็นต้น ซึ่งเป็นปัญหาสำคัญของการทำวิจัยเนื่องจากปัญหาดังกล่าวไม่สามารถคาดการณ์ล่วงหน้าได้ ข้อมูลที่รวบรวมได้จึงเป็นข้อมูลที่ไม่สมบูรณ์หรือมีจำนวนข้อมูลไม่ครบถ้วนตามแผนงานวิจัยที่วางไว้ มีผลทำให้เมื่อนำข้อมูลที่ไม่มีความสมบูรณ์ไปทำการวิเคราะห์ส่งผลให้ผลลัพธ์ที่ได้มีความคลาดเคลื่อนจากความจริงที่เกิดขึ้น และด้วยข้อจำกัดเกี่ยวกับระยะเวลาและงบประมาณในการทำวิจัย หากแก้ปัญหาโดยการเก็บรวบรวมข้อมูลใหม่อีกครั้งเพื่อให้ได้จำนวนข้อมูลครบถ้วนตามแผนการดำเนินงานจึงค่อนข้างเป็นไปได้ยาก ดังนั้นการแก้ไขปัญหาดังกล่าวด้วยวิธีทางสถิติโดยใช้การประมาณค่าสูญหาย (Imputation) จึงมีความเหมาะสม R. Little and D. Rubin [1] ได้จำแนกกลไกการสูญหายของข้อมูล 3 รูปแบบ ได้แก่ การสูญหายแบบสุ่มอย่างสมบูรณ์ (Missing Completely at Random; MCAR) คือความน่าจะเป็นที่ข้อมูลสูญหายไม่ขึ้นกับค่าใด ๆ หรือไม่มีความสัมพันธ์กับข้อมูลอื่นเลย การสูญหายแบบสุ่ม (Missing at Random; MAR) คือ ความน่าจะเป็นที่ข้อมูลสูญหายขึ้นอยู่กับข้อมูลอื่น ๆ หรือมีความสัมพันธ์กับข้อมูลอื่นที่ทราบค่า แต่ไม่ขึ้นกับค่าของข้อมูลที่สูญหาย และการสูญหายแบบไม่สุ่ม (Missing Not at Random; MNAR) คือ ความน่าจะเป็นที่ข้อมูลสูญหายที่ตำแหน่งใด ๆ โดยขึ้นกับค่าของตำแหน่งนั้น ๆ

สำหรับวิธีการประมาณค่าสูญหายที่นิยมใช้กันอย่างแพร่หลายคือ วิธีค่าเฉลี่ย มีการดำเนินงานโดยแทนค่าข้อมูลที่หายไปหรือข้อมูลสูญหายด้วยค่าเฉลี่ยของข้อมูลที่สนใจศึกษาที่สามารถเก็บรวบรวมได้ นอกจากนี้ยังสามารถนำวิธีการดังกล่าวมาพัฒนาต่อยอดสู่ตัวประมาณค่าเฉลี่ยเมื่อข้อมูลมีปัญหาการสูญหายเกิดขึ้นเพื่อทำให้เกิดความสะดวกต่อการนำไปใช้ในการประมาณค่าเฉลี่ยซึ่งเป็นการวิเคราะห์ข้อมูลพื้นฐานที่

ต้องใช้ในงานวิจัยทั่วไปและยังสามารถทำให้ข้อมูลกลับมาเป็นจำนวนครบสมบูรณ์ตามแผนงานวิจัยเพื่อนำไปสู่การวิเคราะห์ข้อมูลขั้นสูงต่อไป อีกทั้งได้มีการนำวิธีการประมาณค่าสูญหายด้วยวิธีค่าเฉลี่ยมาประยุกต์ใช้กับข้อมูลทางมลพิษทางอากาศ เช่น M. N. Norazian et al. [2], N. A. Zainuri et al. [3], K. Chodjuntug and N. Lawson [4], [5] ซึ่งปัญหามลพิษทางอากาศเป็นปัญหาหนึ่งที่สำคัญในขณะนี้ โดยมีค่าระดับมลพิษทางอากาศเกินค่ามาตรฐานในหลายเขตพื้นที่ นอกจากนี้ยังมีผู้เสนอการประมาณค่าสูญหายอีกหลายวิธีเพื่อนำไปสู่ตัวประมาณค่าเฉลี่ยเมื่อข้อมูลสูญหาย ยกตัวอย่างคือ H. Lee et al. [6] ได้เสนอตัวประมาณค่าเฉลี่ยที่ได้จากการประมาณค่าสูญหายด้วยวิธีอัตราส่วน ซึ่งได้แนวคิดมาจาก W. G. Cochran [7] โดยวิธีดังกล่าวเป็นการใช้ข้อมูลของตัวแปรเสริม (Auxiliary Information) หรือตัวแปรที่มีความเกี่ยวข้องสัมพันธ์กับตัวแปรที่สนใจศึกษาเพื่อเพิ่มประสิทธิภาพการประมาณค่าสูญหาย ซึ่งลักษณะชุดข้อมูลของตัวแปรเสริมจะต้องไม่มีข้อมูลสูญหายหรือให้มีข้อมูลสูญหายเกิดขึ้นเฉพาะตัวแปรที่สนใจศึกษาเท่านั้น และจากการศึกษาพบว่า วิธีอัตราส่วนซึ่งเป็นวิธีที่ใช้ข้อมูลตัวแปรเสริมมีประสิทธิภาพดีกว่าวิธีค่าเฉลี่ยซึ่งเป็นวิธีที่ใช้เฉพาะข้อมูลตัวแปรที่สนใจศึกษาโดยไม่มีการใช้ข้อมูลตัวแปรเสริม ในกรณีนี้ตัวแปรทั้งสองมีความสัมพันธ์ระดับสูงในทิศทางเดียวกัน

S. Singh and S. Horn [8] ได้เสนอวิธีการประมาณค่าสูญหายที่ปรับปรุงจากวิธีอัตราส่วนและมีการใช้ค่าคงที่เพื่อปรับการถ่วงน้ำหนักระหว่างกลุ่มข้อมูลที่เก็บรวบรวมได้และกลุ่มข้อมูลสูญหาย โดยวิธีดังกล่าวมีประสิทธิภาพดีกว่าวิธีอัตราส่วน และวิธีค่าเฉลี่ย อย่างไรก็ตามประสิทธิภาพของวิธีการขึ้นอยู่กับ การเลือกค่าคงที่ที่เหมาะสม นอกจากนี้ยังพบว่า เมื่อค่าคงที่มีค่าเท่ากับ 1 วิธีดังกล่าวจะเป็นวิธีอัตราส่วน

ต่อมา D. Shukla and N. Thakur [9] ได้มีการประยุกต์ใช้วิธีของ V. K. Singh and D. Shukla [10] หรือที่เรียกว่า วิธีแฟคเตอร์ โดยถูกปรับปรุงจากวิธี

อัตราส่วน สำหรับการประมาณค่าข้อมูลสูญหาย ผลการศึกษาพบว่าวิธีแฟคเตอร์มีประสิทธิภาพดีกว่าวิธีอัตราส่วน และวิธีค่าเฉลี่ย นอกจากนี้ D. Shukla et al. [11] ได้ปรับปรุงวิธีแฟคเตอร์โดยใช้แนวคิดของ S. Singh and S. Horn [8] ซึ่งใช้ค่าคงที่ในการเพิ่มประสิทธิภาพของวิธีการประมาณค่าสูญหาย โดยเลือกค่าคงที่ที่ทำให้ค่าความคลาดเคลื่อนกำลังสองเฉลี่ยมีค่าต่ำสุด แต่อย่างไรก็ตามวิธีการดังกล่าวศึกษาเฉพาะข้อมูลที่มีการสูญหายในรูปแบบ MCAR ซึ่งรูปแบบการสูญหายที่เกิดขึ้นในการสำรวจตัวอย่างนั้นยังมีอีกสองรูปแบบคือ MAR และ MNAR ดังที่กล่าวมาแล้วข้างต้น

A. K. Singh et al. [12] ได้เสนอวิธีการประมาณค่าสูญหายด้วยตัวประมาณค่าเลขชี้กำลัง หรือเรียกว่าวิธีเลขชี้กำลัง โดยวิธีการดังกล่าวเป็นการใช้ข้อมูลตัวแปรเสริมซึ่งจัดอยู่ในรูปแบบเลขชี้กำลัง จากการศึกษาผลการวิจัยพบว่าวิธีการดังกล่าวมีประสิทธิภาพดีกว่าวิธีการที่ตัวแปรเสริมไม่ได้จัดอยู่ในรูปของเลขชี้กำลัง โดยลักษณะความสัมพันธ์ระหว่างตัวแปรที่สนใจศึกษาและตัวแปรเสริมมีทิศทางเดียวกัน และลักษณะข้อมูลสูญหายอยู่ในรูปแบบ MCAR

สำหรับการประมาณค่าเฉลี่ยจากวิธีค่าเฉลี่ยที่ใช้เฉพาะข้อมูลที่สนใจศึกษาหรือวิธีที่ไม่ใช้ตัวแปรเสริมอาจเกิดความคลาดเคลื่อนจากค่าจริงที่เกิดขึ้นหากข้อมูลมีค่าผิดปกติปะปนอยู่ในชุดข้อมูลจึงถือได้ว่าเป็นข้อจำกัดของวิธีนี้ [13] ปัจจุบันจึงมีนักวิจัยสนใจศึกษาพัฒนาวิธีการประมาณค่าสูญหายเพื่อนำไปประมาณค่าเฉลี่ยโดยใช้ตัวแปรเสริม (Auxiliary Variable) ร่วมกับตัวแปรที่สนใจศึกษาให้มีประสิทธิภาพมากขึ้นหรือดีกว่าวิธีที่ใช้เฉพาะข้อมูลของตัวแปรที่สนใจศึกษา ซึ่งมีผลทำให้ความคลาดเคลื่อนลดลงจากตัวประมาณด้วยวิธีการดังกล่าว [8]-[12]

ดังนั้นงานวิจัยนี้จึงสนใจศึกษาการเปรียบเทียบประสิทธิภาพการประมาณค่าเฉลี่ยด้วยวิธีการประมาณ

ค่าสูญหายด้วยตัวแปรเสริมได้แก่ วิธีแฟคเตอร์ และวิธีเลขชี้กำลัง กับวิธีค่าเฉลี่ยซึ่งเป็นวิธีเบื้องต้นที่ไม่ใช้ตัวแปรเสริมภายใต้รูปแบบการสูญหาย ได้แก่ MCAR, MAR และ MNAR โดยใช้ข้อมูลเกี่ยวกับผู้ลงคะแนนขนาดเล็ก จังหวัดอุบลราชธานี ซึ่งเป็นพื้นที่หนึ่งที่ประสบปัญหาเกี่ยวกับผู้ลงคะแนนขนาดเล็กเกินค่ามาตรฐานและส่งผลกระทบต่อสุขภาพของประชาชนในจังหวัดอุบลราชธานี [14] ข้อมูลที่ใช้เป็นข้อมูลทุติยภูมิจากกรมควบคุมมลพิษ โดยใช้เกณฑ์สำหรับการเปรียบเทียบประสิทธิภาพวิธีการภายใต้สถานการณ์ต่าง ๆ คือ ความคลาดเคลื่อนกำลังสองเฉลี่ย (Mean Square Error; MSE)

2. ระเบียบวิธีวิจัย

เพื่อศึกษาการเปรียบเทียบประสิทธิภาพการประมาณค่าเฉลี่ยเมื่อข้อมูลสูญหายภายใต้รูปแบบการสูญหายที่แตกต่างกัน ในงานวิจัยนี้ใช้ข้อมูลจากกรมควบคุมมลพิษเกี่ยวกับผู้ลงคะแนนขนาดเล็ก จังหวัดอุบลราชธานี ซึ่งใช้เกณฑ์การเปรียบเทียบประสิทธิภาพวิธีการภายใต้สถานการณ์ต่าง ๆ คือ MSE

2.1 สัญลักษณ์และข้อตกลงเบื้องต้น

งานวิจัยนี้ศึกษาการประมาณค่าเฉลี่ยจากวิธีการประมาณค่าสูญหาย ได้แก่ วิธีค่าเฉลี่ย วิธีแฟคเตอร์ และวิธีเลขชี้กำลัง โดยกำหนด Y คือ ตัวแปรที่สนใจศึกษา X คือ ตัวแปรเสริม N คือ ขนาดประชากร n คือ ขนาดตัวอย่าง r คือ จำนวนข้อมูลตอบสนอง $n-r$ คือ จำนวนข้อมูลไม่ตอบสนอง y_i คือ ค่าตัวแปรที่สนใจศึกษาลำดับที่ i x_i คือ ค่าตัวแปรเสริมลำดับที่ i R คือ ชุดของข้อมูลที่ตอบสนอง R' คือ ชุดของข้อมูลที่ไม่ตอบสนอง $\bar{Y}$ คือ ค่าเฉลี่ยของประชากร $\bar{Y}_r$ คือ ค่าเฉลี่ยของประชากร X $\bar{y}_r$ คือ ค่าเฉลี่ยของตัวอย่าง Y จำนวน r หน่วย $\bar{x}_r$ คือ ค่าเฉลี่ยของตัวอย่าง X จำนวน n หน่วย และ $\bar{x}_r$ คือ ค่าเฉลี่ยของตัวอย่าง X จำนวน r หน่วย

2.2 วิธีที่ใช้ในงานวิจัย

สำหรับตัวประมาณค่าเฉลี่ยเมื่อข้อมูลสูญหายจากการประมาณค่าสูญหาย ได้แก่ วิธีค่าเฉลี่ย วิธีแฟคเตอร์ และวิธีเลขชี้กำลัง แล้วจึงทำการหาค่าเฉลี่ยอยู่ในรูปตัวประมาณค่าแบบจุด โดยมีรูปแบบทั่วไปของการหาตัวประมาณดังนี้ [15]

$$y_{.i} = \begin{cases} y_i ; i \in R \\ \hat{y}_i ; i \in R' \end{cases} \quad (1)$$

จากสมการที่ 1 ตัวประมาณค่าแบบจุดรูปแบบทั่วไป คือ

$$\bar{y} = \frac{1}{n} \left[\sum_{i \in R} y_i + \sum_{i \in R'} \hat{y}_i \right] \quad (2)$$

2.2.1 วิธีค่าเฉลี่ย

การประมาณค่าสูญหายโดยวิธีการประมาณค่าเฉลี่ยเป็นวิธีการที่นิยมนำมาใช้กันอย่างแพร่หลายเนื่องจากเป็นวิธีที่ไม่ซับซ้อนและง่ายต่อการนำไปใช้ในงานวิจัยด้านต่าง ๆ โดยมีรายละเอียดของวิธีการดังนี้ [16]

$$y_{.i} = \begin{cases} y_i ; i \in R \\ \bar{y}_r ; i \in R' \end{cases} \quad (3)$$

จากสมการที่ 1 ได้ตัวประมาณค่าแบบจุด คือ

$$\bar{y}_M = \bar{y}_r \quad (4)$$

2.2.2 วิธีแฟคเตอร์

วิธีแฟคเตอร์เพื่อใช้ในการประมาณค่าสูญหายถูกเสนอโดย D. Shukla and N. Thakur [9] โดยมีการใช้ข้อมูลของตัวแปรเสริมในการประมาณค่าสูญหายต่อมา D. Shukla et al. [11] ได้ปรับปรุงวิธีแฟคเตอร์โดยใช้

แนวคิดของ S. Singh and S. Horn [8] มีรายละเอียดของวิธีการดังนี้

$$(y_{.i})_j = \begin{cases} \frac{kn}{r} y_i + (1-k)\phi_j ; i \in R \\ \frac{1}{(n-r)} \bar{y}_r [\phi_j - r] ; i \in R' \end{cases} \quad (5)$$

เมื่อ $\phi_j ; j = 1, 2, 3$

โดยที่

$$\phi_1 = \frac{n \{ (A+C) \bar{X} + (n/N) B \bar{X}_n \}}{\{ (A + (n/N) B) \bar{X} + C \bar{X}_n \}}$$

$$\phi_2 = \frac{n \{ (A+C) \bar{X}_n + (n/N) B \bar{X}_r \}}{\{ (A + (n/N) B) \bar{X}_n + C \bar{X}_r \}}$$

$$\phi_3 = \frac{n \{ (A+C) \bar{X} + (n/N) B \bar{X}_r \}}{\{ (A + (n/N) B) \bar{X} + C \bar{X}_r \}}$$

$$A = (k-1)(k-2); 0 < k < \infty$$

$$B = (k-1)(k-4)$$

$$C = (k-2)(k-3)(k-4)$$

จากสมการที่ 5 ตัวประมาณค่าแบบจุดคือ

$$(\bar{y}_{Fj})_k = k\bar{y}_r + (1-k)\phi_j \quad (6)$$

นอกจากนี้ D. Shukla et al. [11] ได้แสดงให้เห็นว่าสามารถแทนค่า $k = 1, 2, 3, 4$ ได้ และหากแทนค่า k ด้วย 1 และ 4 ตัวประมาณค่า $\bar{y}_{Fj}$ เท่ากับ $\bar{y}_r$ หรือ $\bar{y}_{Fj} = \bar{y}_M = \bar{y}_r$ ดังนั้นในงานวิจัยนี้จะแทนค่า k ด้วย 2 และ 3 เนื่องจากต้องการให้ตัวประมาณค่าที่ได้แตกต่างจาก $\bar{y}_M = \bar{y}_r$

2.2.3 วิธีเลขชี้กำลัง

A. K. Singh et al. [12] ได้เสนอวิธีเลขชี้กำลังซึ่งได้แนวคิดจาก S. Bahl and R. K. Tuteja [17] และ

A. K. Singh et al. [18] ในการปรับปรุงวิธีการประมาณค่าสูญหาย มีรายละเอียดของวิธีการดังนี้

$$(y_i)_j = \begin{cases} y_i & ; i \in R \\ \frac{1}{(n-r)} \bar{y}_r [nT_j - r] ; i \in R' \end{cases} \quad (7)$$

เมื่อ T_j ; $j = 1, 2, 3$

โดยที่

$$T_1 = \exp \left[\alpha \left(\frac{\bar{X} - \bar{x}_n}{\bar{X} + \bar{x}_n} \right) \right]$$

$$T_2 = \exp \left[\alpha \left(\frac{\bar{x}_n - \bar{x}_r}{\bar{x}_n + \bar{x}_r} \right) \right]$$

$$T_3 = \exp \left[\alpha \left(\frac{\bar{X} - \bar{x}_r}{\bar{X} + \bar{x}_r} \right) \right]$$

จากสมการที่ 7 ตัวประมาณค่าแบบจุดคือ

$$\bar{y}_{Ej} = \bar{y}_r T_j \quad (8)$$

โดยเสนอให้ $\alpha = 2\rho \frac{C_Y}{C_X}$

เมื่อ

ρ คือ ค่าสัมประสิทธิ์สหสัมพันธ์ระหว่างตัวแปร

(X, Y) C_X คือ ค่าสัมประสิทธิ์การแปรผันของ X

C_Y คือ ค่าสัมประสิทธิ์การแปรผันของ Y

2.3 วิธีการศึกษา

2.3.1 ข้อมูล

งานวิจัยนี้ใช้ข้อมูลของฝุ่นละอองขนาดเล็กจังหวัดอุบลราชธานี จากแหล่งข้อมูลของกรมควบคุมมลพิษ โดยกำหนดตัวแปรที่สนใจศึกษาและตัวแปรเสริม (Y, X) คือ $PM_{2.5}$ และ PM_{10} ตามลำดับ มีขนาดประชากร (N) เท่ากับ 1,126 จำนวน โดยข้อมูลมีค่า $\bar{X}$ เท่ากับ 52.83 $\bar{Y}$ เท่ากับ 76.35 ρ เท่ากับ 0.97

C_X เท่ากับ 0.53 และ C_Y เท่ากับ 0.45 ทำการสุ่มตัวอย่างแบบง่ายโดยไม่ใส่คืน กำหนดขนาดตัวอย่าง 30 100 และ 500 ในแต่ละขนาดตัวอย่างกำหนดให้มีร้อยละข้อมูลสูญหาย คือ 10 20 และ 30 ซึ่งลักษณะการสูญหายเป็นแบบ MCAR MAR และ MNAR สำหรับการสูญหายแบบ MCAR คือ การสูญหายของค่าสังเกตในตัวแปร Y จะไม่มีความสัมพันธ์กับตัวแปรใดเลยหรือข้อมูลอื่น ๆ ส่วนการสูญหายแบบ MAR คือ การสูญหายของค่าสังเกตในตัวแปร Y จะมีความสัมพันธ์กับข้อมูลในตัวแปรเกี่ยวข้องที่ทราบค่า และการสูญหายแบบ MNAR คือ การสูญหายของค่าสังเกตในตัวแปร Y จะไม่เหมือนรูปแบบที่ได้กล่าวไปแล้วข้างต้น ซึ่งการสูญหายจะเกิดขึ้นที่ตำแหน่งใด ๆ โดยไม่เป็นไปอย่างสุ่ม ในแต่ละเหตุการณ์จะมีการทำซ้ำทั้งหมด 20,000 รอบ และในงานวิจัยนี้ใช้โปรแกรม R ในการวิเคราะห์ข้อมูล

2.3.2 การเปรียบเทียบประสิทธิภาพ

ในแต่ละสถานการณ์ ใช้ MSE เป็นเกณฑ์เปรียบเทียบประสิทธิภาพการประมาณค่าเฉลี่ยที่ได้จากวิธีต่าง ๆ มีรายละเอียดการคำนวณดังนี้

$$MSE(\bar{y}) = \sum_{l=1}^{20,000} \frac{(\bar{y}_l - \bar{Y})^2}{20,000} \quad (9)$$

โดยที่

$\bar{y}$ คือ ตัวประมาณค่าเฉลี่ยจากวิธีการต่าง ๆ

$\bar{Y}$ คือ ค่าเฉลี่ยของประชากร

$MSE(\bar{y})$ คือ ค่าเฉลี่ยความคลาดเคลื่อน

กำลังสองของ $\bar{y}$

3. ผลการศึกษาและอภิปรายผล

การเปรียบเทียบประสิทธิภาพการประมาณค่าเฉลี่ยที่ได้จากวิธีการประมาณค่าสูญหายจะพิจารณาจากค่า MSE โดยเปรียบเทียบระหว่างวิธีเบื้องต้นที่ใช้เฉพาะตัวแปรที่สนใจศึกษาคือ $\bar{y}_M$ กับวิธีที่ใช้ตัวแปรเสริมร่วมด้วย ได้แก่ $(\bar{y}_{F1})_2$ $(\bar{y}_{F2})_2$ $(\bar{y}_{F3})_2$

$(\bar{y}_{F1})_3$ $(\bar{y}_{F2})_3$ $(\bar{y}_{F3})_3$ $\bar{y}_{E1}$ $\bar{y}_{E2}$ และ $\bar{y}_{E3}$ สำหรับผลการศึกษาได้จำแนกตามลักษณะรูปแบบการสูญหายของข้อมูลได้แก่ MCAR MAR และ MNAR ดังตารางที่ 1 ถึง 3 ตามลำดับ ต่อมาจะพิจารณาเฉพาะ $\bar{y}_M$ ซึ่งเป็นวิธีเบื้องต้น และวิธีที่มีค่า MSE ต่ำกว่า $\bar{y}_M$ โดยแสดงผลดังรูปที่ 1 ถึง 3 ตามลำดับ จำแนกตามลักษณะรูปแบบการสูญหาย เพื่อศึกษาอิทธิพลของระดับการสูญหายที่มีต่อวิธีการดังกล่าว

3.1 ผลการศึกษา

3.1.1 MCAR

จากผลการศึกษาดังตารางที่ 1 พบว่า $(\bar{y}_{F1})_2$ $(\bar{y}_{F2})_2$ $(\bar{y}_{F3})_2$ $\bar{y}_{E1}$ $\bar{y}_{E2}$ และ $\bar{y}_{E3}$ มีค่า MSE ต่ำกว่า $\bar{y}_M$ เมื่อพิจารณาค่า MSE โดยจำแนกตามขนาดตัวอย่างและร้อยละการสูญหาย พบว่า

ขนาดตัวอย่าง เท่ากับ 30 เมื่อร้อยละข้อมูลสูญหาย เท่ากับ 10 สามารถเรียงลำดับ ค่า MSE ต่ำสุดได้ดังนี้ $(\bar{y}_{F3})_2$ $\bar{y}_{E3}$ และ $(\bar{y}_{F1})_2$ ตามลำดับ

ขนาดตัวอย่าง เท่ากับ 30 เมื่อร้อยละข้อมูลสูญหาย เท่ากับ 20 สามารถเรียงลำดับ ค่า MSE ต่ำสุดได้ดังนี้ $(\bar{y}_{F3})_2$ $\bar{y}_{E3}$ โดย MSE ของ $(\bar{y}_{F3})_2$ เท่ากับ $\bar{y}_{E3}$ ถัดมา คือ $(\bar{y}_{F1})_2$ และ $\bar{y}_{E1}$ ตามลำดับ

ขนาดตัวอย่าง เท่ากับ 30 เมื่อร้อยละข้อมูลสูญหาย เท่ากับ 30 สามารถเรียงลำดับ ค่า MSE ต่ำสุดได้ดังนี้ $\bar{y}_{E3}$ $(\bar{y}_{F3})_2$ ตามลำดับ สุดท้าย คือ $(\bar{y}_{F1})_2$ กับ $\bar{y}_{E1}$ โดย MSE ของ $(\bar{y}_{F1})_2$ เท่ากับ $\bar{y}_{E1}$

ขนาดตัวอย่าง เท่ากับ 100 เมื่อร้อยละข้อมูลสูญหาย เท่ากับ 10 สามารถเรียงลำดับ ค่า MSE ต่ำสุดได้ดังนี้ $(\bar{y}_{F3})_2$ $\bar{y}_{E3}$ และ $(\bar{y}_{F1})_2$ ตามลำดับ

ขนาดตัวอย่าง เท่ากับ 100 เมื่อร้อยละข้อมูลสูญหาย เท่ากับ 20 สามารถเรียงลำดับ ค่า MSE ต่ำสุดได้ดังนี้ $(\bar{y}_{F3})_2$ $\bar{y}_{E3}$ และ $(\bar{y}_{F1})_2$ ตามลำดับ

ขนาดตัวอย่าง เท่ากับ 100 เมื่อร้อยละข้อมูลสูญหาย เท่ากับ 30 สามารถเรียงลำดับ ค่า MSE ต่ำสุดได้ดังนี้ $(\bar{y}_{F3})_2$ $\bar{y}_{E3}$ และ $(\bar{y}_{F1})_2$ ตามลำดับ

ขนาดตัวอย่าง เท่ากับ 500 เมื่อร้อยละข้อมูลสูญหาย เท่ากับ 10 สามารถเรียงลำดับ ค่า MSE ต่ำสุดได้ดังนี้ $(\bar{y}_{F3})_2$ กับ $\bar{y}_{E3}$ ถัดมา $(\bar{y}_{F1})_2$ กับ $\bar{y}_{E1}$ และ $(\bar{y}_{F2})_2$ กับ $\bar{y}_{E2}$ ตามลำดับ โดยลำดับเดียวกันนี้ ค่า MSE เท่ากัน

ขนาดตัวอย่าง เท่ากับ 500 เมื่อร้อยละข้อมูลสูญหาย เท่ากับ 20 สามารถเรียงลำดับ ค่า MSE ต่ำสุดได้ดังนี้ $(\bar{y}_{F3})_2$ $\bar{y}_{E3}$ ตามลำดับ สุดท้ายคือ $(\bar{y}_{F1})_2$ $(\bar{y}_{F2})_2$ $\bar{y}_{E1}$ และ $\bar{y}_{E2}$ โดยลำดับสุดท้ายทุกตัวมีค่า MSE เท่ากัน

ขนาดตัวอย่าง เท่ากับ 500 เมื่อร้อยละข้อมูลสูญหาย เท่ากับ 30 สามารถเรียงลำดับ ค่า MSE ต่ำสุดได้ดังนี้ $(\bar{y}_{F3})_2$ กับ $\bar{y}_{E3}$ โดย MSE ของ $(\bar{y}_{F3})_2$ เท่ากับ $\bar{y}_{E3}$ ถัดมา คือ $(\bar{y}_{F2})_2$ และ $\bar{y}_{E2}$ ตามลำดับ

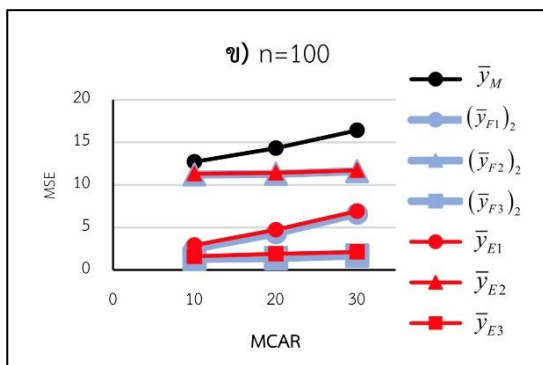
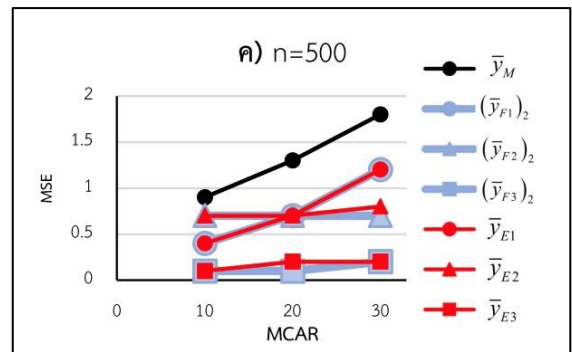
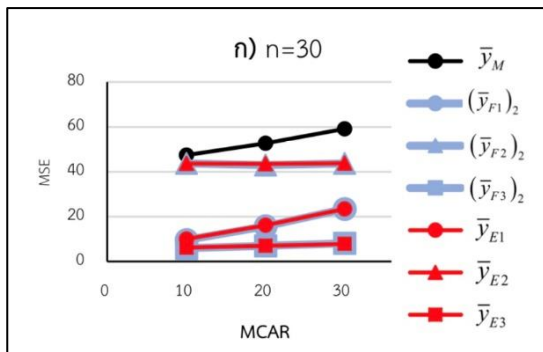
นอกจากนี้เมื่อพิจารณาอิทธิพลของร้อยละการสูญหายที่มีต่อค่า MSE ของวิธีต่าง ๆ จากรูปที่ 1 พบว่าขนาดตัวอย่าง เท่ากับ 30 ร้อยละการสูญหายที่แตกต่างกันมีผลต่อวิธี $\bar{y}_M$ กับวิธีที่ใช้ตัวแปรเสริม ได้แก่ $(\bar{y}_{F1})_2$ และ $\bar{y}_{E1}$ โดยร้อยละข้อมูลสูญหายเพิ่มขึ้น ค่า MSE ของวิธีดังกล่าวเพิ่มขึ้นอย่างชัดเจน

ขนาดตัวอย่าง เท่ากับ 100 ร้อยละการสูญหายที่แตกต่างกันมีผลต่อวิธี $\bar{y}_M$ กับวิธีที่ใช้ตัวแปรเสริม ได้แก่ $(\bar{y}_{F1})_2$ และ $\bar{y}_{E1}$ โดยร้อยละข้อมูลสูญหายเพิ่มขึ้น ค่า MSE ของวิธีดังกล่าวเพิ่มขึ้นอย่างชัดเจน

ขนาดตัวอย่าง เท่ากับ 500 ร้อยละการสูญหายที่แตกต่างกันมีผลต่อวิธี $\bar{y}_M$ กับวิธีที่ใช้ตัวแปรเสริม ได้แก่ $(\bar{y}_{F1})_2$ และ $\bar{y}_{E1}$ โดยร้อยละข้อมูลสูญหายเพิ่มขึ้น ค่า MSE ของวิธีดังกล่าวเพิ่มขึ้นอย่างชัดเจน

ตารางที่ 1 แสดงค่า MSE ของตัวประมาณค่าใด ภายใต้ข้อมูลสูญหายแบบ MCAR สำหรับขนาดตัวอย่างเท่ากับ 30 100 และ 500

n	M	$\bar{y}_M$	$(\bar{y}_{F1})_2$	$(\bar{y}_{F2})_2$	$(\bar{y}_{F3})_2$	$(\bar{y}_{F1})_3$	$(\bar{y}_{F2})_3$	$(\bar{y}_{F3})_3$	$\bar{y}_{E1}$	$\bar{y}_{E2}$	$\bar{y}_{E3}$
30	10	47.4	9.6	43.7	5.9	57.6	48.4	58.7	10.0	43.7	6.2
	20	52.7	15.7	43.0	7.0	62.7	55.2	65.2	16.2	43.5	7.0
	30	59.1	23.4	43.6	8.1	68.9	63.2	73.1	23.4	43.8	7.8
100	10	12.7	2.5	11.2	1.3	24.0	14.2	25.6	2.9	11.3	1.6
	20	14.3	4.3	11.3	1.4	25.5	17.6	28.8	4.7	11.4	1.9
	30	16.4	6.6	11.6	1.7	27.5	21.9	33.0	6.9	11.7	2.1
500	10	0.9	0.4	0.7	0.1	43.8	20.4	63.3	0.4	0.7	0.1
	20	1.3	0.7	0.7	0.1	45.0	44.8	89.7	0.7	0.7	0.2
	30	1.8	1.2	0.7	0.2	45.4	75.0	119.4	1.2	0.8	0.2



รูปที่ 1 เปรียบเทียบ MSE ของข้อมูลสูญหายแบบ MCAR กำหนดร้อยละข้อมูลสูญหายได้แก่ 10 20 30

3.1.2 MAR

จากผลการศึกษาดังตารางที่ 2 พบว่า $(\bar{y}_{F1})_2$ $(\bar{y}_{F2})_2$ $(\bar{y}_{F3})_2$ $\bar{y}_{E1}$ $\bar{y}_{E2}$ และ $\bar{y}_{E3}$ มีค่า MSE ต่ำกว่า $\bar{y}_M$ เมื่อพิจารณาค่า MSE โดยจำแนกตามขนาดตัวอย่างและร้อยละการสูญหาย พบว่า

ขนาดตัวอย่าง เท่ากับ 30 เมื่อร้อยละข้อมูล
สูญหาย เท่ากับ 10 สามารถเรียงลำดับ ค่า MSE ต่ำสุด
ได้ดังนี้ $\bar{y}_{E3}$ $(\bar{y}_{F3})_2$ และ $(\bar{y}_{F1})_2$ ตามลำดับ

ขนาดตัวอย่าง เท่ากับ 30 เมื่อร้อยละข้อมูล
สูญหาย เท่ากับ 20 สามารถเรียงลำดับ ค่า MSE ต่ำสุด
ได้ดังนี้ $\bar{y}_{E3}$ $(\bar{y}_{F3})_2$ และ $(\bar{y}_{F2})_2$ ตามลำดับ

ขนาดตัวอย่าง เท่ากับ 30 เมื่อร้อยละข้อมูล
สูญหาย เท่ากับ 30 สามารถเรียงลำดับ ค่า MSE ต่ำสุด
ได้ดังนี้ $\bar{y}_{E3}$ $\bar{y}_{E2}$ และ $(\bar{y}_{F2})_2$ ตามลำดับ

ขนาดตัวอย่าง เท่ากับ 100 เมื่อร้อยละข้อมูล
สูญหาย เท่ากับ 10 สามารถเรียงลำดับ ค่า MSE ต่ำสุด
ได้ดังนี้ $(\bar{y}_{F3})_2$ $\bar{y}_{E3}$ และ $(\bar{y}_{F2})_2$ ตามลำดับ

ขนาดตัวอย่าง เท่ากับ 100 เมื่อร้อยละข้อมูล
สูญหาย เท่ากับ 20 สามารถเรียงลำดับ ค่า MSE ต่ำสุด
ได้ดังนี้ $\bar{y}_{E3}$ $(\bar{y}_{F3})_2$ และ $(\bar{y}_{F2})_2$ ตามลำดับ

ขนาดตัวอย่าง เท่ากับ 100 เมื่อร้อยละข้อมูล
สูญหาย เท่ากับ 30 สามารถเรียงลำดับ ค่า MSE ต่ำสุด
ได้ดังนี้ $\bar{y}_{E3}$ $\bar{y}_{E2}$ และ $(\bar{y}_{F2})_2$ ตามลำดับ

ขนาดตัวอย่าง เท่ากับ 500 เมื่อร้อยละข้อมูล

สูญหาย เท่ากับ 10 สามารถเรียงลำดับ ค่า MSE ต่ำสุด
ได้ดังนี้ $(\bar{y}_{F3})_2$ $(\bar{y}_{F2})_2$ และ $\bar{y}_{E3}$ ตามลำดับ

ขนาดตัวอย่าง เท่ากับ 500 เมื่อร้อยละข้อมูล
สูญหาย เท่ากับ 20 สามารถเรียงลำดับ ค่า MSE ต่ำสุด
ได้ดังนี้ $(\bar{y}_{F3})_2$ $(\bar{y}_{F2})_2$ และ $\bar{y}_{E3}$ ตามลำดับ

ขนาดตัวอย่าง เท่ากับ 500 เมื่อร้อยละข้อมูล
สูญหาย เท่ากับ 30 สามารถเรียงลำดับ ค่า MSE ต่ำสุด
ได้ดังนี้ $\bar{y}_{E3}$ $\bar{y}_{E2}$ และ $(\bar{y}_{F2})_2$ ตามลำดับ

เมื่อพิจารณาอิทธิพลของร้อยละการสูญหายที่มี
ต่อค่า MSE ของวิธีต่าง ๆ จากรูปที่ 2 พบว่า

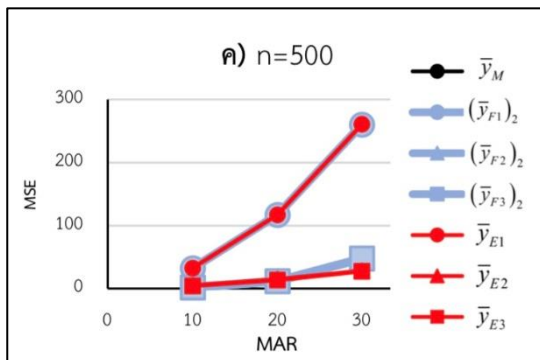
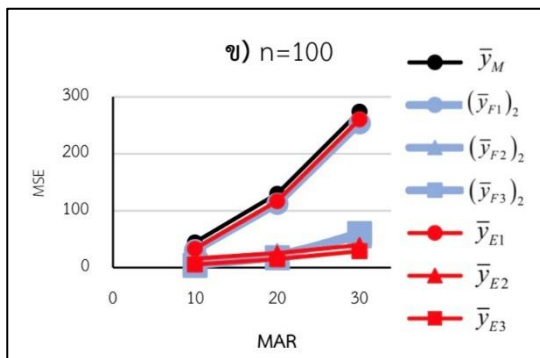
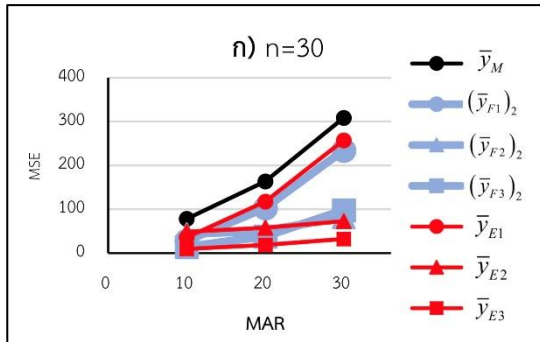
ขนาดตัวอย่าง เท่ากับ 30 ร้อยละการสูญหายที่
แตกต่างกันมีผลต่อค่า MSE ของทุกวิธี โดยร้อยละข้อมูล
สูญหายเพิ่มขึ้น ค่า MSE เพิ่มขึ้น

ขนาดตัวอย่าง เท่ากับ 100 ร้อยละการสูญหาย
ที่แตกต่างกันมีผลต่อค่า MSE ของทุกวิธี โดยร้อยละ
ข้อมูลสูญหายเพิ่มขึ้น ค่า MSE เพิ่มขึ้น

ขนาดตัวอย่าง เท่ากับ 500 ร้อยละการสูญหาย
ที่แตกต่างกันมีผลต่อค่า MSE ของทุกวิธี โดยร้อยละ
ข้อมูลสูญหายเพิ่มขึ้น ค่า MSE เพิ่มขึ้น

ตารางที่ 2 แสดงค่า MSE ของตัวประมาณค่าใด ภายใต้ข้อมูลสูญหายแบบ MAR สำหรับขนาดตัวอย่างเท่ากับ 30 100 และ 500

<i>n</i>	<i>M</i>	$\bar{y}_M$	$(\bar{y}_{F1})_2$	$(\bar{y}_{F2})_2$	$(\bar{y}_{F3})_2$	$(\bar{y}_{F1})_3$	$(\bar{y}_{F2})_3$	$(\bar{y}_{F3})_3$	$\bar{y}_{E1}$	$\bar{y}_{E2}$	$\bar{y}_{E3}$
30	10	78.2	28.3	43.6	14.3	90.5	86.0	99.6	35.8	48.7	9.5
	20	163.6	103.6	51.0	38.3	177.6	194.4	211.2	117.8	58.0	18.6
	30	308.7	234.8	81.8	98.0	325.1	383.3	404.7	257.5	72.9	32.5
100	10	44.6	30.6	12.2	4.6	58.6	78.3	94.3	33.2	15.8	5.5
	20	129.6	112.7	21.7	18.2	145.5	263.5	284.0	117.1	25.3	14.9
	30	274.1	253.9	55.9	60.5	292.6	603.5	630.4	260.4	39.5	28.9
500	10	33.1	32.0	2.0	1.6	85.7	2238.9	2328.9	32.3	4.8	4.2
	20	117.8	116.8	12.0	11.7	176.2	9366.5	9469.3	117.0	14.3	13.7
	30	261.0	260.2	47.5	47.6	324.8	24159.9	24281.6	260.4	28.1	27.5



รูปที่ 2 เปรียบเทียบ MSE ของข้อมูลสูญหายแบบ MAR กำหนดร้อยละข้อมูลสูญหาย ได้แก่ 10 20 30

3.1.3 MNAR

จากผลการศึกษาดังตารางที่ 3 พบว่า $(\bar{y}_{F1})_2$ $(\bar{y}_{F2})_2$ $(\bar{y}_{F3})_2$ $\bar{y}_{E1}$ $\bar{y}_{E2}$ และ $\bar{y}_{E3}$ มีค่า MSE ต่ำกว่า $\bar{y}_M$ เมื่อพิจารณาค่า MSE โดยจำแนกตามขนาดตัวอย่างและร้อยละการสูญหาย พบว่า

ขนาดตัวอย่าง เท่ากับ 30 เมื่อร้อยละข้อมูลสูญหาย เท่ากับ 10 สามารถเรียงลำดับ ค่า MSE ต่ำสุด ได้นี้ $\bar{y}_{E3}$ $(\bar{y}_{F3})_2$ และ $(\bar{y}_{F1})_2$ ตามลำดับ

ขนาดตัวอย่าง เท่ากับ 30 เมื่อร้อยละข้อมูลสูญหาย เท่ากับ 20 สามารถเรียงลำดับ ค่า MSE ต่ำสุด ได้นี้ $\bar{y}_{E3}$ $(\bar{y}_{F3})_2$ และ $(\bar{y}_{F2})_2$ ตามลำดับ

ขนาดตัวอย่าง เท่ากับ 30 เมื่อร้อยละข้อมูลสูญหาย เท่ากับ 30 สามารถเรียงลำดับ ค่า MSE ต่ำสุด ได้นี้ $\bar{y}_{E3}$ $(\bar{y}_{F2})_2$ และ $\bar{y}_{E2}$ ตามลำดับ

ขนาดตัวอย่าง เท่ากับ 100 เมื่อร้อยละข้อมูลสูญหาย เท่ากับ 10 สามารถเรียงลำดับ ค่า MSE ต่ำสุด ได้นี้ $(\bar{y}_{F3})_2$ $\bar{y}_{E3}$ และ $(\bar{y}_{F2})_2$ ตามลำดับ

ขนาดตัวอย่าง เท่ากับ 100 เมื่อร้อยละข้อมูลสูญหาย เท่ากับ 20 สามารถเรียงลำดับ ค่า MSE ต่ำสุด ได้นี้ $(\bar{y}_{F3})_2$ $\bar{y}_{E3}$ และ $(\bar{y}_{F2})_2$ ตามลำดับ

ขนาดตัวอย่าง เท่ากับ 100 เมื่อร้อยละข้อมูลสูญหาย เท่ากับ 30 สามารถเรียงลำดับ ค่า MSE ต่ำสุด ได้นี้ $\bar{y}_{E3}$ $(\bar{y}_{F2})_2$ และ $\bar{y}_{E2}$ ตามลำดับ

ขนาดตัวอย่าง เท่ากับ 500 เมื่อร้อยละข้อมูลสูญหาย เท่ากับ 10 สามารถเรียงลำดับ ค่า MSE ต่ำสุด ได้นี้ $(\bar{y}_{F3})_2$ $(\bar{y}_{F2})_2$ และ $\bar{y}_{E3}$ ตามลำดับ

ขนาดตัวอย่าง เท่ากับ 500 เมื่อร้อยละข้อมูลสูญหาย เท่ากับ 20 สามารถเรียงลำดับ ค่า MSE ต่ำสุด ได้นี้ $(\bar{y}_{F3})_2$ $(\bar{y}_{F2})_2$ และ $\bar{y}_{E3}$ ตามลำดับ

ขนาดตัวอย่าง เท่ากับ 500 เมื่อร้อยละข้อมูลสูญหาย เท่ากับ 30 สามารถเรียงลำดับ ค่า MSE ต่ำสุด ได้นี้ $\bar{y}_{E3}$ $\bar{y}_{E2}$ และ $(\bar{y}_{F2})_2$ กับ $(\bar{y}_{F3})_2$ โดย

MSE ของ $(\bar{y}_{F2})_2$ เท่ากับ $(\bar{y}_{F3})_2$ ตามลำดับ

นอกจากนี้เมื่อพิจารณาอิทธิพลของร้อยละการสูญหายที่มีต่อค่า MSE ของวิธีต่าง ๆ จากรูปที่ 3 พบว่า

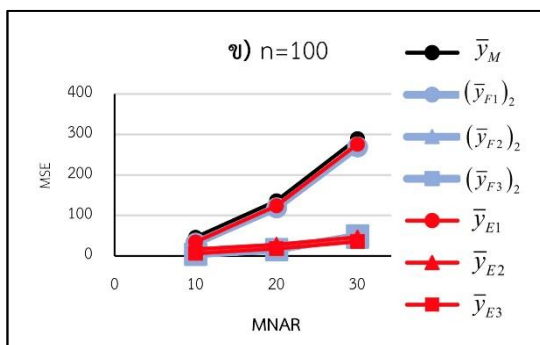
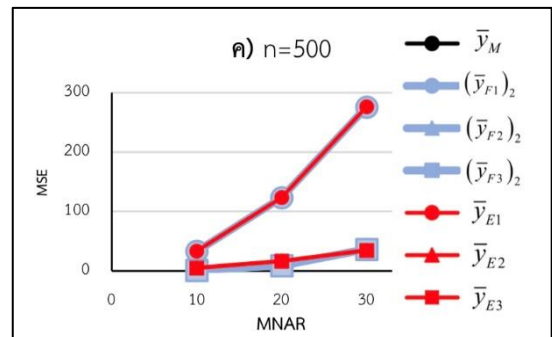
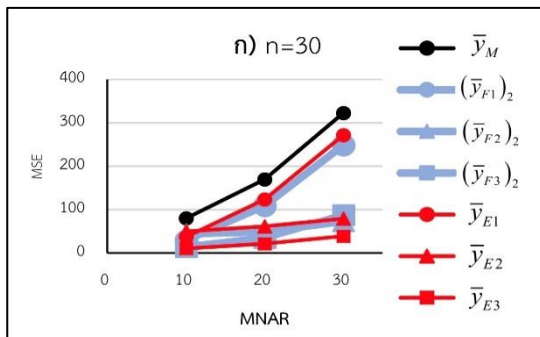
ขนาดตัวอย่าง เท่ากับ 30 ร้อยละการสูญหายที่แตกต่างกันมีผลต่อค่า MSE ของทุกวิธี โดยร้อยละข้อมูลสูญหายเพิ่มขึ้น ค่า MSE เพิ่มขึ้น

ขนาดตัวอย่าง เท่ากับ 100 ร้อยละการสูญหายที่แตกต่างกันมีผลต่อค่า MSE ของทุกวิธี โดยร้อยละข้อมูลสูญหายเพิ่มขึ้น ค่า MSE เพิ่มขึ้น

ขนาดตัวอย่าง เท่ากับ 500 ร้อยละการสูญเสียข้อมูลสูญหายเพิ่มขึ้น ค่า MSE เพิ่มขึ้น
หายที่แตกต่างกันมีผลต่อค่า MSE ของทุกวิธี

ตารางที่ 3 แสดงค่า MSE ของตัวประมาณค่าใด ภายใต้ข้อมูลสูญหายแบบ MNAR สำหรับขนาดตัวอย่างเท่ากับ 30 100 และ 500

n	M	$\bar{y}_M$	$(\bar{y}_{F1})_2$	$(\bar{y}_{F2})_2$	$(\bar{y}_{F3})_2$	$(\bar{y}_{F1})_3$	$(\bar{y}_{F2})_3$	$(\bar{y}_{F3})_3$	$\bar{y}_{E1}$	$\bar{y}_{E2}$	$\bar{y}_{E3}$
30	10	79.5	29.1	43.6	13.6	91.8	87.3	100.9	36.8	49.5	10.1
	20	169.5	108.9	48.8	35.0	183.6	200.6	217.4	123.5	60.9	21.2
	30	322.8	248.7	72.6	87.1	339.2	398.1	419.4	271.7	79.1	38.7
100	10	45.7	31.5	12.0	4.1	59.7	79.3	95.3	34.2	16.5	6.1
	20	135.7	118.7	18.9	15.2	151.7	270.5	291.1	123.3	28.1	17.6
	30	289.5	268.9	45.9	49.6	308.2	620.3	647.3	275.7	46.3	35.5
500	10	34.1	33.0	1.6	1.2	86.8	2197.8	2287.7	33.3	5.4	4.7
	20	124.2	123.2	9.1	8.9	182.8	9244.1	9346.6	123.5	17.1	16.4
	30	276.8	275.9	36.6	36.6	341.5	23605.5	23728.3	276.1	35.0	34.4



รูปที่ 3 เปรียบเทียบ MSE ของข้อมูลสูญหายแบบ MNAR กำหนดร้อยละข้อมูลสูญหาย ได้แก่ 10 20 30

3.2 อภิปรายผล

ข้อมูลที่ใช้ในการศึกษาครั้งนี้ได้กำหนดให้ตัวแปรที่สนใจศึกษาและตัวแปรเสริม (Y, X) คือ $PM_{2.5}$ และ PM_{10} ตามลำดับ จากข้อมูลพบว่า ρ เท่ากับ 0.97 แสดงให้เห็นว่าทั้ง 2 ตัวแปรมีความสัมพันธ์กันระดับสูง

ในทิศทางเดียวกัน [13] เมื่อพิจารณาภาพรวมจากผลการศึกษาทั้งหมดที่ได้ตั้งตารางที่ 1-3 และรูปที่ 1-3 ด้วยค่า MSE แสดงให้เห็นว่า $(\bar{y}_{F1})_2$ $(\bar{y}_{F2})_2$ $(\bar{y}_{F3})_2$ $\bar{y}_{E1}$ $\bar{y}_{E2}$ และ $\bar{y}_{E3}$ เป็นวิธีที่ใช้ข้อมูลของตัวแปรเสริมในการคำนวณโดยมีค่า MSE ต่ำว่า $\bar{y}_M$ ซึ่งเป็นวิธีที่ใช้เฉพาะข้อมูลที่สนใจศึกษา ดังนั้นวิธีที่ใช้ตัวแปรเสริมหรือข้อมูลที่มีความสัมพันธ์กับตัวแปรที่สนใจศึกษาสามารถเพิ่มประสิทธิภาพการประมาณค่าเฉลี่ยได้ซึ่งมีความสอดคล้องกับงานวิจัยของ V. Dubey and H. K. Sharma [19] และ A. Audu et al. [20]

อีกทั้งเมื่อพิจารณาวิธีที่มีการใช้ข้อมูลจากตัวแปรเสริมในการคำนวณ ได้แก่ วิธีแฟคเตอร์ พบว่า $(\bar{y}_{F3})_2$ และ $(\bar{y}_{F1})_2$ มีประสิทธิภาพดีกว่า $(\bar{y}_{F2})_2$ และวิธีเลขชี้กำลัง พบว่า $\bar{y}_{E3}$ และ $\bar{y}_{E1}$ มีประสิทธิภาพดีกว่า $\bar{y}_{E2}$ จากการพิจารณาการใช้ข้อมูลจากตัวแปรเสริม พบว่า $(\bar{y}_{F3})_2$ $(\bar{y}_{F1})_2$ $\bar{y}_{E3}$ $\bar{y}_{E1}$ มีการใช้ $\bar{X}$ ในการคำนวณ แต่ $(\bar{y}_{F2})_2$ $\bar{y}_{E2}$ ไม่มีการใช้ $\bar{X}$ ในการคำนวณ จึงแสดงให้เห็นว่า ภายใต้วิธีเดียวกันนั้นการใช้ข้อมูล $\bar{X}$ มีประสิทธิภาพดีกว่าการไม่ใช้ข้อมูล $\bar{X}$ ซึ่งสอดคล้องกับงานวิจัยของ A. K. Singh et al. [12]

รูปแบบการสุ่มหายที่แตกต่างกันภายใต้สถานการณ์เดียวกันมีผลต่อประสิทธิภาพของการประมาณค่าเฉลี่ยที่ได้จากวิธีการประมาณค่าสุ่มหาย ซึ่งค่า MSE ในรูปแบบการสุ่มหายแบบ MAR และ MNAR มีค่ามากกว่า MCAR ดังนั้นประสิทธิภาพของการประมาณค่าสุ่มหายภายใต้การสุ่มหายแบบ MCAR จึงดีกว่าการสุ่มหายแบบ MAR และ MNAR โดยมีความสอดคล้องกับงานวิจัยของ K. Srihaset et al. [21]

นอกจากนี้ยังมีปัจจัยเกี่ยวกับขนาดตัวอย่างและระดับการสุ่มหายของข้อมูลมีผลต่อประสิทธิภาพของตัวประมาณที่ได้จากวิธีต่าง ๆ เช่นกัน เมื่อขนาดตัวอย่างเพิ่มขึ้นประสิทธิภาพของวิธีการเพิ่มสูงขึ้นซึ่งมีทิศทางเดียวกัน แต่ถ้าวระดับการสุ่มหายเพิ่มขึ้นประสิทธิภาพ

ของวิธีการลดลงซึ่งมีทิศทางตรงกันข้ามโดยสอดคล้องกับงานวิจัยของ M. Rueda and S. Gonzalez [22] และ G. N. Singh and S. Suman [23]

4. สรุป

งานวิจัยนี้สนใจศึกษาประสิทธิภาพการประมาณค่าเฉลี่ยจากวิธีการประมาณค่าสุ่มหายที่อาศัยข้อมูลตัวแปรเสริม ภายใต้รูปแบบการสุ่มหายที่แตกต่างกัน ได้แก่ แบบ MCAR MAR และ MNAR สำหรับการศึกษาตัวประมาณค่าเฉลี่ยในงานวิจัย ได้แก่ $\bar{y}_M$ $(\bar{y}_{F1})_2$ $(\bar{y}_{F2})_2$ $(\bar{y}_{F3})_2$ $(\bar{y}_{F1})_3$ $(\bar{y}_{F2})_3$ $(\bar{y}_{F3})_3$ $\bar{y}_{E1}$ $\bar{y}_{E2}$ และ $\bar{y}_{E3}$ โดย $\bar{y}_M$ ได้มาจากวิธีค่าเฉลี่ยซึ่งเป็นวิธีเบื้องต้นที่ใช้เฉพาะตัวแปรที่สนใจศึกษา ส่วน $(\bar{y}_{F1})_2$ $(\bar{y}_{F2})_2$ $(\bar{y}_{F3})_2$ $(\bar{y}_{F1})_3$ $(\bar{y}_{F2})_3$ $(\bar{y}_{F3})_3$ ได้มาจากวิธีแฟคเตอร์ และ $\bar{y}_{E1}$ $\bar{y}_{E2}$ และ $\bar{y}_{E3}$ ได้มาจากวิธีเลขชี้กำลัง สำหรับวิธีวิธีแฟคเตอร์และวิธีเลขชี้กำลังเป็นการใช้ข้อมูลตัวแปรเสริมร่วมกับตัวแปรที่สนใจศึกษา ในงานวิจัยนี้ใช้ข้อมูลจริงเกี่ยวกับมลพิษทางอากาศ คือ ฝุ่นละอองขนาดเล็กเพื่อศึกษาประสิทธิภาพการประมาณค่าเฉลี่ยจากวิธีการประมาณค่าสุ่มหายที่ได้นำเสนอ ภายใต้ขนาดตัวอย่าง 30 100 และ 500 ในแต่ละขนาดตัวอย่างกำหนดให้มีร้อยละข้อมูลสุ่มหาย คือ 10 20 และ 30

จากการศึกษาพบว่า ภายใต้รูปแบบการสุ่มหาย ได้แก่ แบบ MCAR MAR และ MNAR การประมาณค่าเฉลี่ยจากวิธีแฟคเตอร์ คือ $(\bar{y}_{F3})_2$ และวิธีเลขชี้กำลัง คือ $\bar{y}_{E3}$ มีประสิทธิภาพดีที่สุด อีกทั้งระดับข้อมูลสุ่มหายและขนาดตัวอย่างมีอิทธิพลต่อประสิทธิภาพของตัวประมาณค่าอีกด้วย โดยขนาดตัวอย่างเพิ่มขึ้นประสิทธิภาพของวิธีการเพิ่มสูงขึ้น แต่ถ้าวระดับการสุ่มหายเพิ่มขึ้นประสิทธิภาพของวิธีการเพิ่มลดลง

ข้อเสนอแนะในการทำงานวิจัยครั้งต่อไป โดยงานวิจัยนี้ได้ใช้ข้อมูลของฝุ่นละอองขนาดเล็ก จังหวัด

อุปสรรคเหล่านี้ในการศึกษา ซึ่งกำหนดตัวแปรที่สนใจศึกษาและตัวแปรเสริมคือ $PM_{2.5}$ และ PM_{10} จากการศึกษาพบว่า ρ เท่ากับ 0.97 ซึ่งเป็นการศึกษาความสัมพันธ์ระหว่าง 2 ตัวแปรเพียงระดับเดียว และไม่ได้ประมาณค่าเฉลี่ยภายใต้ตัวแบบการแจกแจงความน่าจะเป็นใด ๆ ดังนั้นในงานวิจัยครั้งต่อไปอาจศึกษาระดับความสัมพันธ์ระหว่างตัวแปรที่สนใจศึกษากับตัวแปรเสริมให้มีความหลากหลายมากขึ้น เพื่อศึกษาอิทธิพลของระดับความสัมพันธ์ระหว่าง 2 ตัวแปรที่มีต่อประสิทธิภาพของการประมาณค่าเฉลี่ยจากวิธีที่ได้นำเสนอในงานวิจัย นอกจากนี้ยังสามารถศึกษาเพิ่มเติมภายใต้ข้อมูลที่สนใจโดยระบุตัวแบบการแจกแจงความน่าจะเป็น ซึ่งอาจจะพิจารณาจาก Bivariate Random Variables เพื่อประโยชน์ต่องานวิจัยในอนาคต

5. เอกสารอ้างอิง

- [1] R. Little and D. Rubin, *Statistical analysis with missing data*. 3rd ed. New York: Wiley, 2019.
- [2] M. N. Norazian et al., "Estimation of missing values in air pollution data using single imputation techniques," *ScienceAsia*, vol. 34, pp. 341-345, 2008.
- [3] N. A. Zainuri, A. A. Jemain and N. Muda, "A Comparison of Various Imputation Methods for Missing Values in Air Quality Data," *Sains Malaysiana*, vol. 44, no. 3, pp. 449-456, 2015.
- [4] K. Chodjuntug and N. Lawson, "Imputation for estimating the population mean in the presence of nonresponse, with application to fine particle density in Bangkok," *Mathematical Population Studies*, vol. 29, no. 4, pp. 204-225, 2022a.
- [5] K. Chodjuntug and N. Lawson, "A chain regression exponential type imputation method for mean estimation in the presence of missing data," *Songklanakarin Journal of Science & Technology*, vol. 44, no. 4, pp. 1109-1118, 2022b.
- [6] H. Lee, E. Rancourt and C. E. Sarndall, "Experiments with variance estimation from survey data with imputed values." *Journal of Official Statistics*, vol. 10, no. 3, pp. 231-243, 1994.
- [7] W. G. Cochran, "The estimation of yield of cereal experiments by sampling for the ratio of gain to total produce." *Journal of Agricultural Science*, vol. 30, pp. 262-275, 1940.
- [8] S. Singh and S. Horn, "Compromised imputation in survey sampling." *Metrika*, vol. 51, pp. 267-276, 2000.
- [9] D. Shukla and N. Thakur, "Estimation of mean with imputation of missing data using factor type estimator." *Statistics in Transition*. vol. 9, pp. 33-48, 2008.
- [10] V. K. Singh and D. Shukla, "One parameter family of factor type ratio estimator," *Metron*, vol. 45, pp. 273-283, 1987.
- [11] D. Shukla, N. Singh and T. S. Pathak, "Some new aspects on imputation in sampling," *African Journal of Mathematics and Computer Science Research*, vol. 6, no. 1, pp. 5-15, 2013.

- [12] A. K. Singh, P. Singh, and V. Singh, "Estimation of mean with imputation of missing data using exponential-type estimators," *Journal of Statistics Applications and Probability Letters*, vol. 9, no. 2, pp. 71-80, 2022.
- [13] K. Chodjuntug. *Biostatistics*, 1st ed. Ubon Ratchathani: Ubon Ratchathani University Printing House, 2024.
- [14] S. Salao, N. Potatong, P. Kammoongkun and P. Choomanee, "Chemical composition analysis for source identification of PM_{2.5} in Muang District of Ubon Ratchathani Province," *Burapha science Journal*, vol. 26, no. 1, pp. 438-453, 2021.
- [15] S.K. Yadav, G.K. Vishwakarma and D.K. Sharm, "A computational strategy for estimation of mean using optimal imputation in presence of missing observation." *Scientific Reports*, vol.14, no. 6433, pp1-13, 2024.
- [16] T. Dachochaiporn, and K. Chodjuntug, "Estimating the Mean of PM_{2.5} with Missing Data in the Area Around Electricity Generating Authority of Thailand Using the Improved Compromised Imputation Method." *Current Applied Science and Technology*, vol.23, no. 5, pp 1-14, 2023.
- [17] S. Bahl and R. K. Tuteja, "Ratio and product type exponential estimator," *Journal of the Statistics Applications and Probability*, vol. 12, no. 1, pp. 159-163, 1991.
- [18] A. K. Singh, P. Singh, and V. K. Singh, "Exponential-Type Compromised Imputation in Survey Sampling," *Journal of the Statistics Applications and Probability*, vol. 3, no. 2, pp. 211-217, 2014.
- [19] V. Dubey and H. K. Sharma, "On estimating population variance using auxiliary information," *Statistics in Transition*, vol. 9, no. 1, pp. 7-18, 2008.
- [20] A. Audu et al, "On the efficiency of imputation estimators using auxiliary attribute," *Continental Journal Applied Sciences*, vol. 15, no. 1, pp. 1-13, 2020.
- [21] K. Srihaset, S. Kanjanawasee and D. Srisukho, "A comparison of the quality of missing data treatment methods for examinees' ability parameters estimation," *Journal of Research Methodology*, vol. 26, no. 2, pp. 169-187, 2013.
- [22] M. Rueda and S. Gonzalez, "A new ratio-type imputation with random disturbance," *Applied Mathematics Letters*, vol. 21, pp. 978-982, 2008.
- [23] G. N. Singh and S. Suman, "Estimation of population mean using imputation methods for missing data under two-phase sampling design," *Journal of Statistical Theory and Practice*, vol. 13, no. 19, pp. 1-24, 2019.