บทความวิจัย

การสร้างชั้นภูมิโดยใช้เทคนิคการจัดกลุ่มด้วยอัลกอริทึมเคมีนสำหรับการสุ่มตัวอย่างแบบแบ่งชั้นภูมิ

วิชญ์วิสิฐ เกษรสิทธิ์* ปรีชญา หะสะเล็ม และ จิราวัลย์ จิตรถเวช
สาขาสถิติ คณะสถิติประยุกต์ สถาบันบัณฑิตพัฒนบริหารศาสตร์

* ผู้นิพนธ์ประสานงาน โทรศัพท์ 09 7201 3776 อีเมล: witwisit.kes@gmail.com
DOI: 10.14416/j.kmutnb.2018.12.004
รับเมื่อ 12 กันยายน 2561 แก้ไขเมื่อ 13 พฤศจิกายน 2561 ตอบรับเมื่อ 23 พฤศจิกายน 2561 เผยแพร่ออนไลน์ 6 ธันวาคม 2561

บทคัดย่อ

งานวิจัยนี้ได้นำเสนอเทคนิคการสร้างชั้นภูมิด้วยเทคนิคการจัดกลุ่มโดยอัลกอริทึมเคมีนในการประมาณค่าเฉลี่ยประชากรสำหรับการสุ่มตัวอย่างแบบชั้นภูมิ หน่วยตัวอย่างในแต่ละชั้นภูมิสุ่มด้วยวิธีการสุ่มแบบง่ายไม่คืนที่และกระจายตามสัดส่วนกับจำนวนประชากรในชั้นภูมินั้นๆ การศึกษาใช้จำนวนชั้นภูมิเท่ากับ 4, 5 และ 6 ชั้นภูมิ กำหนดค่าสัมประสิทธิ์สหสัมพันธ์สูงสุดระหว่างตัวแปรช่วยกับตัวแปรที่สนใจ เท่ากับ 0.50, 0.70 และ 0.90 และขนาดตัวอย่างเท่ากับ 50, 100, 150, 200 และ 300 หน่วย ประสิทธิภาพของตัวประมาณค่าโดยการสร้างชั้นภูมิตามที่เสนอเปรียบเทียบตัวประมาณค่าโดยการสร้างชั้นภูมิด้วยวิธีของดาเลเนียสและฮอดจ์ ผลการศึกษาพบว่าตัวประมาณค่าเฉลี่ยจากการสุ่มตัวอย่างแบบแบ่งชั้นภูมิที่สร้างชั้นภูมิด้วยอัลกอริทึมเคมีน มีประสิทธิภาพสูงกว่าการสร้างชั้นภูมิโดยวิธีของดาเลเนียสและฮอดจ์ในการจำลองทุกกรณี

คำสำคัญ: การประมาณค่าเฉลี่ยประชากร, การสุ่มตัวอย่างแบบแบ่งชั้นภูมิ, เทคนิคการจัดกลุ่มข้อมูลด้วยอัลกอริทึมเคมีน

การอ้างอิงบทความ: วิชญ์วิสิฐ เกษรสิทธิ์ ปรีชญา หะสะเล็ม และ จิราวัลย์ จิตรถเวช, "การสร้างชั้นภูมิโดยใช้เทคนิคการจัดกลุ่มด้วยอัลกอริทึมเคมีนสำหรับการสุ่มตัวอย่างแบบแบ่งชั้นภูมิ," *วารสารวิชาการพระจอมเกล้าพระนครเหนือ*, ปีที่ 29, ฉบับที่ 2, หน้า 321–331, เม.ย.–มิ.ย. 2562.

วารสารวิชาการพระจอมเกล้าพระนครเหนือ ปีที่ 29 ฉบับที่ 2 เม.ย.–มิ.ย. 2562
The Journal of KMUTNB., Vol. 29, No. 2, Apr.–Jun. 2019

Research Article

Stratum Boundaries Construction Using *K*-means Clustering Algorithm in Stratified Random Sampling

Witwisit Kesornsit*, Prechaya Hasalem and Jirawan Jitthavech
School of Applied Statistics, National Institute of Development Administration, Bangkok, Thailand

* Corresponding Author, Tel. 09 7201 3776, E–mail: witwisit.kes@gmail.com DOI: 10.14416/j.kmutnb.2018.12.004
Received 12 September 2018; Revised 13 November 2018; Accepted 23 November 2018; Published online: 6 December 2018

Abstract

This research presented the stratum boundaries construction techniques by *K*-means clustering algorithm for estimating the population mean in stratified random sampling. Samples are selected by simple random sampling method without replacement and allocated in accordance with the population size in each stratum. In this study, the population was separated into 4, 5 and 6 strata with the maximum correlation coefficients between the auxiliary variables and the variables of interest equal to 0.50, 0.70 and 0.90 and sample sizes of 50, 100, 150, 200 and 300. The efficiency of the estimator by the proposed strata construction relative to the estimate by Dalenius & Hodges strata construction method is used in the estimator evaluation. The estimator when constructing the strata by *K*-means clustering algorithm was more efficient in all simulation cases.

Keywords: Population Mean Estimation, Stratified Random Sampling, *K*-Means Clustering

Please cite this article as: W. Kesornsit, P. Hasalem, and J. Jitthavech, "Stratum boundaries construction using *K*-means clustering algorithm in stratified random sampling," *The Journal of KMUTNB*, vol. 29, no. 2, pp. 321–331, Apr.–Jun. 2019 (in Thai).

1. บทนำ

การสุ่มตัวอย่างแบบแบ่งชั้นภูมิเป็นวิธีการที่ใช้กันอย่างแพร่หลายมากที่สุดไม่ว่าจะเป็นงานวิจัยการสำรวจขนาดใหญ่หรือขนาดเล็ก โดยเฉพาะอย่างยิ่งงานวิจัยที่ครอบคลุมหลายพื้นที่ ครอบคลุมบุคคลหลายกลุ่มหรือหลายอาชีพ มักใช้วิธีการสุ่มตัวอย่างด้วยวิธีการนี้ [1], [2] การสุ่มตัวอย่างแบบแบ่งชั้นภูมิเป็นวิธีการสุ่มตัวอย่างจากประชากรหนึ่ง โดยทำการแบ่งประชากรออกเป็นส่วนๆ ที่ไม่ทับซ้อนกัน ซึ่งแต่ละส่วนเรียกว่าชั้นภูมิ (Stratum) แล้วทำการสุ่มตัวอย่างจากชั้นภูมินั้นๆ โดยใช้แบบเดียวกันหรือแตกต่างกันก็ได้ เช่น การสุ่มตัวอย่างแบบแบ่งชั้นภูมิอย่างง่าย (Stratified Random Sampling) การสุ่มตัวอย่างแบบแบ่งชั้นภูมิอย่างมีระบบ (Stratified Systematic Sampling) เป็นต้น

การสุ่มตัวอย่างแบบแบ่งชั้นภูมิเป็นวิธีการสุ่มตัวอย่างที่ให้ค่าประมาณของพารามิเตอร์ที่มีความถูกต้องหรือความแม่นยำสูงกว่าวิธีการสุ่มตัวอย่างโดยวิธีอื่นๆ ที่ใช้ขนาดตัวอย่างจำนวนเท่ากัน หรือเสียค่าใช้จ่ายจำนวนเท่ากัน เป็นวิธีการที่สามารถทำการวิเคราะห์ข้อมูลเฉพาะบางส่วนของประชากรคือวิเคราะห์ข้อมูลแยกชั้นภูมิกันได้ ทำให้ได้สารสนเทศในแต่ละชั้นภูมิตามความต้องการของผู้วิจัย นอกจากนั้นยังสามารถให้น้ำหนักความสำคัญแก่หน่วยตัวอย่างบางหน่วยให้สูงกว่าหน่วยอื่นๆ รวมทั้งสามารถบริหารการสำรวจได้สะดวกมากขึ้น โดยสามารถแบ่งการเก็บรวบรวมข้อมูลไปตามชั้นภูมิต่างๆ ก่อให้เกิดการประหยัดเวลาและค่าใช้จ่ายในการสำรวจได้ [1], [3]

เนื่องด้วยการใช้วิธีการสุ่มตัวอย่างแบบแบ่งชั้นภูมิมีสิ่งที่ควรพิจารณาหลายประการ เช่น การเลือกใช้วิธีการสุ่มตัวอย่างจากแต่ละชั้นภูมิ การเลือกตัวแปรที่ใช้เป็นเกณฑ์ในการแบ่งประชากรเป็นชั้นภูมิหรือการสร้างชั้นภูมิ การกระจายตัวอย่างไปตามชั้นภูมิ การกำหนดจำนวนชั้นภูมิ และการหาขอบเขตของชั้นภูมิ เพื่อจะทำให้ความแปรปรวนของตัวประมาณพารามิเตอร์ที่สนใจมีค่าต่ำที่สุด โดยเฉพาะอย่างยิ่งวิธีการแบ่งประชากรออกเป็นชั้นภูมิที่แตกต่างกันย่อมมีผลต่อการประมาณค่าพารามิเตอร์ที่สนใจ ดังนั้นผู้วิจัยจึงต้องเลือกวิธีการแบ่งประชากรออกเป็นชั้นภูมิให้เหมาะสมที่สุด [1], [4]

ปัจจุบันวิธีการแบ่งประชากรออกเป็นชั้นภูมิที่ใช้ส่วนใหญ่เป็นไปตามลักษณะทางภูมิศาสตร์ โดยเฉพาะอย่างยิ่งงานวิจัยที่ต้องการเสนอผลการวิจัยเป็นรายเขตตามภูมิศาสตร์ หากไม่มีข้อจำกัดดังกล่าว จำนวนชั้นภูมิที่ควรใช้ และหลักการในการแบ่งชั้นภูมิต้องให้หน่วยตัวอย่างในชั้นภูมิมีความคล้ายคลึงกันและหน่วยตัวอย่างในแต่ละชั้นภูมิมีความแตกต่างกัน เพื่อให้ขนาดตัวอย่างที่ต้องสุ่มมาเพื่อทำการศึกษาลดลง และความแปรปรวนของตัวประมาณค่าเท่าเดิมเมื่อเทียบกับวิธีการสุ่มแบบอย่างง่าย และเป็นที่ทราบกันดีว่าเมื่อเพิ่มจำนวนชั้นภูมิความแปรปรวนของตัวประมาณจะลดลงแต่ค่าใช้จ่ายในการสำรวจจะเพิ่มขึ้น ดังนั้น ตัวแปรที่ดีที่สุดที่ใช้ในการแบ่งชั้นภูมิคือตัวแปรที่ต้องการศึกษา หากไม่มีข้อมูลในอดีตของตัวแปรที่สนใจมาช่วยในการแบ่งชั้นภูมิ ตัวแปรที่เหมาะสมในการแบ่งชั้นภูมิคือตัวแปรที่มีความสัมพันธ์หรือเกี่ยวข้องกับตัวแปรที่สนใจมากที่สุดมาช่วยในการแบ่งชั้นภูมิ ตัวแปรนี้เรียกว่าตัวแปรช่วย [5]

ในการศึกษาครั้งนี้ ใช้วิธีการสร้างชั้นภูมิโดยใช้อัลกอริทึมเคมีน เปรียบเทียบกับวิธีรากที่สองของความถี่สะสมสำหรับตัวแปรเชิงปริมาณที่มีความสัมพันธ์กับตัวแปรที่สนใจ y สูง [6] ของดาเลเนียสและฮอดจ์โดยใช้เกณฑ์ของประสิทธิภาพของตัวประมาณค่าสัมพัทธ์ของตัวประมาณค่าเฉลี่ยของประชากรในการเปรียบเทียบ จากวิธีการสร้างชั้นภูมิโดยวิธีของดาเลเนียสและฮอดจ์จำเป็นต้องเลือกตัวแปรช่วย x ที่เหมาะสมที่สุด 1 ตัว ในการแบ่งประชากรออกเป็นชั้นภูมิ ในที่นี้ใช้ตัวแปรที่มีความสัมพันธ์กับตัวแปรที่สนใจ y มากที่สุดในการแบ่งชั้นภูมิ วิธีการสร้างชั้นภูมิโดยใช้อัลกอริทึมเคมีน ซึ่งเป็นเทคนิคการจัดกลุ่มที่สามารถใช้ตัวแปร x ได้ s ตัว $(x_1, x_2,...,x_s)$ และเปรียบเทียบการประมาณค่าเฉลี่ยประชากรของการสุ่มตัวอย่างแบบแบ่งชั้นภูมิทั้ง 2 วิธี

1.1 การประมาณค่าเฉลี่ยประชากร

การสุ่มตัวอย่างแบบแบ่งชั้นภูมิ (Stratified Sampling) ที่มีการสุ่มตัวอย่างแบบง่ายในทุกชั้นภูมิเรียกว่าการสุ่มตัวอย่างแบบแบ่งชั้นภูมิอย่างง่าย สำหรับการวิจัยครั้งนี้จะศึกษาภายใต้การสุ่มตัวอย่างแบบแบ่งชั้นภูมิอย่างง่ายแบบไม่คืนที่

กรณีการสุ่มตัวอย่างด้วยวิธีการแบ่งประชากรขนาด N ออกเป็น L ชั้นภูมิ โดยที่ชั้นภูมิที่ i มีหน่วยตัวอย่างอยู่ N_i หน่วยซึ่ง $\sum_{i=1}^{L} N_i = N$ แล้วสุ่มตัวอย่างขนาด n_i จากชั้นภูมิที่ i ด้วยวิธีการสุ่มตัวอย่างแบบสุ่มอย่างง่ายไม่คืนที่โดยอิสระกัน ซึ่ง $\sum_{i=1}^{L} n_i = n$ ดังนั้นสามารถประมาณค่าเฉลี่ย μ_y ของประชากรโดยไม่เอนเอียงโดยการใช้น้ำหนักของชั้นภูมิต่างๆ ได้ดังสมการที่ (1) [7]

$$\overline{y} = \sum_{i=1}^{L} W_i \overline{y}_i \tag{1}$$

ซึ่งเป็นตัวประมาณไม่เอนเอียงของสมการที่ (2)

$$\mu_y = \frac{T_y}{N} = \frac{T(y)}{N} = \sum_{i=1}^{L}\sum_{j=1}^{N_i} Y_{ij} \tag{2}$$

เมื่อ $W_i = \frac{N_i}{N}$ เป็นน้ำหนักของชั้นภูมิที่ i และ $\overline{y}_i = \frac{1}{n_i}\sum_{j=1}^{n_i} y_{ij}$ โดยที่ y_{ij} คือค่าสังเกตของตัวแปร Y ที่สนใจศึกษาจากหน่วยตัวอย่างที่ j ซึ่งสุ่มจากชั้นภูมิที่ i ด้วยวิธีการสุ่มอย่างง่ายไม่คืนที่ [1], [3], [5] และความแปรปรวนของ $\overline{y}$ แสดงได้ดังสมการที่ (3)

$$V(\overline{y}) = S^2(\overline{y}) = \sum_{i=1}^{L} W_i^2 (1 - f_i)\frac{S_i^2}{n_i} \tag{3}$$

เมื่อ $f_i = \frac{n_i}{N_i}$, $S_i^2 = \frac{1}{N_i - 1}\sum_{j=1}^{N_i}(Y_{ij} - \mu_i)^2$ และ $\mu_i = \frac{1}{N_i}\sum_{j=1}^{N_i} Y_{ij}$

โดยตัวประมาณไม่เอนเอียงของ $V(\overline{y})$ ดังสมการที่ (4)

$$v(\overline{y}) = s^2(\overline{y}) = \sum_{i=1}^{L} W_i^2 (1 - f_i)\frac{s_i^2}{n_i} \tag{4}$$

เมื่อ $s_i^2 = \frac{1}{n_i - 1}\sum_{j=1}^{n_i}(y_{ij} - \overline{y}_i)^2$ คือความแปรปรวนของตัวอย่างที่สุ่มจากชั้นภูมิที่ i [1], [5], [8]

1.2 การกระจายตัวอย่างไปตามชั้นภูมิ

การกระจายตัวอย่างไปตามชั้นภูมิเป็นการกำหนดขนาดของตัวอย่างจากแต่ละชั้นภูมิ เมื่อทราบหรือกำหนดขนาดตัวอย่าง n ที่จะใช้แล้ว เมื่อใช้การสุ่มตัวอย่างแบบสุ่มอย่างง่ายในแต่ละชั้นภูมิโดยอิสระกันแล้วจะกระจาย n ไปตามชั้นภูมิต่างๆ เป็น $n_1, n_2, \ldots, n_L$ โดยที่ $\sum_{i=1}^{L} n_i = n$ ซึ่งการกระจายตัวอย่างขนาด n ไปตามชั้นภูมิมักพิจารณาจากความคลาดเคลื่อนของตัวประมาณหรือค่าใช้จ่ายในการสำรวจ นอกจากหลักเกณฑ์ดังกล่าวแล้วยังมีวิธีการกระจายตัวอย่างไปตามชั้นภูมิที่นิยมใช้กันอีกหลายวิธี ได้แก่ แบบแบ่งเท่ากัน แบบไม่มีกฎเกณฑ์ แบบให้เหมาะสมที่สุดแบบเนย์มัน และแบบได้สัดส่วนกับจำนวนที่มี ซึ่งในการวิจัยครั้งนี้ผู้วิจัยเลือกเทคนิคการกระจายตัวอย่างแบบได้สัดส่วนกับจำนวนหน่วยที่มี [1], [5]

การกระจายตัวอย่างแบบได้สัดส่วนกับจำนวนที่มีเป็นกรณีที่ให้สัดส่วนการสุ่มในชั้นภูมิต่างๆ เท่ากันหมดและเท่ากับสัดส่วนการสุ่มตัวอย่างจากประชากร กล่าวคือ $f_i = \frac{n_i}{N_i} = \frac{n}{N} = f$ ซึ่งจะได้ดังสมการที่ (5)

$$n_i = \frac{N_i}{N} \cdot n = W_i \cdot n \tag{5}$$

โดยที่ $i = 1, 2, \ldots, L$ ในกรณีดังกล่าวนี้จะได้ดังสมการที่ (6)

$$V(\overline{y}) = \frac{1 - f}{n} \cdot \sum_{i=1}^{L} W_i S_i^2 \tag{6}$$

1.3 การสร้างชั้นภูมิ

การสุ่มตัวอย่างแบบแบ่งชั้นภูมิจะมีประสิทธิภาพในการสุ่มตัวอย่างก็ต่อเมื่อจำนวนชั้นภูมิและขอบเขตของชั้นภูมิมีความเหมาะสม โดยทั่วไปผู้ใช้งานจะนิยมแบ่งชั้นภูมิตามลักษณะทางภูมิศาสตร์หรือตำแหน่งที่อยู่ของหน่วยตัวอย่าง สำหรับกรณีที่ชุดข้อมูลมีแต่ตัวแปรเชิงปริมาณจะมีวิธีการแบ่งประชากรออกเป็นชั้นภูมิโดยใช้ค่าของตัวแปรช่วย x ที่เรียกว่าตัวแปรแบ่งชั้นภูมิ ซึ่งมีความสัมพันธ์สูงกับตัวแปร y ที่สนใจซึ่งเป็นที่คาดหวังได้ว่าตัวแปร x มีความสัมพันธ์กับ y สูงจะให้ค่าประมาณของ y แม่นยำขึ้น [5], [9]

จากการศึกษาของดาเลเนียสได้แสดงว่าถ้า x กับ y มีความสัมพันธ์เชิงเส้นกันแล้ว การแบ่งประชากรออกเป็นชั้นภูมิตามค่าของ x จะทำให้การประมาณค่าพารามิเตอร์

เกี่ยวกับ y มีความแม่นยำสูงเมื่อกระจายตัวอย่างไปตามชั้นภูมิแบบได้สัดส่วนกับจำนวนหน่วยในชั้นภูมิ ซึ่งจาก $V(\bar{y}) = \sum_{i=1}^{L} W_i^2 (1-f) \frac{S_i^2}{n_i}$ จะเห็นได้ว่า $V(\bar{y})$ มีค่าน้อยเมื่อ S_i^2 มีค่าน้อย ดังนั้นจึงควรสร้างชั้นภูมิโดยให้หน่วยตัวอย่างในชั้นภูมิเดียวกันมีค่าของตัวแปรที่ศึกษาใกล้เคียงกัน [1], [9]

การหาขอบเขตของชั้นภูมิโดยกำหนดให้ตัวแปร x เป็นตัวแปรแบ่งชั้นภูมิ และเมื่อต้องการหาค่าของ $x_1, x_2, ..., x_{L-1}$ ซึ่ง $x_0 < x_1 < ... < x_{L-1} < x_L$ โดยที่ x_0 คือค่าต่ำสุดของ x และ y คือค่าสูงสุดของ x ในประชากร ดังนั้นเพื่อให้ความแปรปรวนของตัวประมาณพารามิเตอร์มีค่าต่ำที่สุดจากการสุ่มตัวอย่างจากแต่ละชั้นภูมิจึงมีนักวิจัยศึกษาวิธีการสร้างชั้นภูมิไว้จำนวนมาก ได้แก่ แบบได้สัดส่วนกับจำนวนที่มีแบบเนย์มัน แบบเท่ากัน แบบดาเลเนียสและฮอดจ์หรือเรียกว่ารากที่สองของความถี่สะสม แบบดาเลเนียสและเกอร์นีย์ แบบมหาลาโนบิส แบบอาโอยามา และแบบเอคมัน เป็นต้น สำหรับการวิจัยครั้งนี้จะทำการสร้างชั้นภูมิโดยวิธีดาเลเนียสและฮอดจ์หรือรากที่สองของความถี่สะสมของ Dalenius & Hodges [6] ซึ่งเสนอให้แบ่งค่าสะสมของรากที่สองของความถี่ของ x ออกเป็น L ส่วน ในส่วนละเท่าๆ กัน ภายใต้การกระจายแบบเนย์มันเรียกว่ารากที่สองของความถี่สะสม [1], [6] แสดงดังสมการที่ (7)

$$x_i = x_{i-1} + \frac{\sqrt{\sum f}}{L} \tag{7}$$

โดยที่ $i = 1, 2, ..., L$ และ $\frac{\sqrt{\sum f}}{L}$ คือรากที่สองของความถี่สะสมทั้งหมด

1.4 เทคนิคการจัดกลุ่ม (Clustering Techniques)

การจัดกลุ่มเป็นวิธีการในการแบ่งกลุ่มของข้อมูลโดยไม่ทราบมาก่อนว่าข้อมูลมีทั้งหมดกี่กลุ่มและข้อมูลแต่ละกลุ่มมีลักษณะเป็นอย่างไร แต่จะแบ่งกลุ่มโดยการพิจารณาจากลักษณะของข้อมูลที่นำมาวิเคราะห์ [10] อัลกอริทึมสำหรับการแบ่งกลุ่มข้อมูลมีหลายประเภท การวิจัยในครั้งนี้จะทำการจัดกลุ่มข้อมูลด้วยเทคนิคเคมีนซึ่งเป็นอัลกอริทึมในการจัดกลุ่มที่ต้องกำหนดจำนวนกลุ่มข้อมูลตามผู้วิจัย เพื่อจะสามารถกำหนดจำนวนกลุ่มให้เท่ากับจำนวนชั้นภูมิที่ผู้วิจัยต้องการเปรียบเทียบผลการวิจัย

การจัดกลุ่มข้อมูลด้วยอัลกอริทึมเคมีนเป็นวิธีการแบ่งกลุ่มข้อมูลตามค่า k ที่กำหนด ซึ่งกระบวนการทำงานจะทำการเลือกค่า k เริ่มต้นสำหรับเป็นค่ากลางในการจัดกลุ่มโดยการสุ่มค่าเริ่มต้นจำนวน k ค่าเรียกว่าจุดศูนย์กลางการจัดกลุ่ม (Centroid) หลังจากนั้นนำข้อมูลทั้งหมดจัดเข้ากลุ่มที่มีจุดศูนย์กลางที่อยู่ใกล้วัตถุนั้นมากที่สุดโดยคำนวณจากการวัดระยะห่างระหว่างจุดที่น้อยที่สุด และปรับค่า k ใหม่ให้เป็นค่ากลางของกลุ่มข้อมูล และหยุดเมื่อค่า k ไม่เปลี่ยนแปลง [10], [11] ดังนั้นการใช้การจัดกลุ่มแบบเคมีนจึงสามารถนำมาใช้งานเมื่อทราบจำนวนกลุ่มที่ต้องการจำแนกที่แน่นอน ทั้งนี้การวัดระยะห่างระหว่างข้อมูลสามารถใช้วิธีการคำนวณระยะห่างได้หลากหลาย เช่น ระยะทางแบบยุคลิด (Euclidean Distance) เป็นต้น [12], [13]

สำหรับการวัดระยะห่างของข้อมูลโดยวิธี ระยะทางแบบยุคลิดเป็นการวัดระยะห่างระหว่างจุด 2 จุด เช่น จุด p และจุด q ในระบบพิกัดคาร์ทีเซียน (Cartesian Coordinates) ถ้ามี $p = (p_1, p_2, ..., p_n)$, $q = (q_1, q_2, ..., q_n)$ เป็นจุดสองจุดในปริภูมิเวกเตอร์แบบยุคลิด n ปริภูมิ (Euclidean n-space) ซึ่งระยะทางระหว่าง p ไปจนถึง q หรือ q ไปจนถึง p จะสามารถคำนวณได้ดังสมการที่ (8)

$$\begin{aligned} d(p,q) &= \sqrt{(p_1-q_1)^2+(p_2-q_2)^2+...+(p_n-q_n)^2} \\ &= \sqrt{\sum_{i=1}^{n}(p_i-q_i)^2} \end{aligned} \tag{8}$$

โดยที่ $d(p,q)$ คือระยะทางระหว่าง p ไปจนถึง q หรือ q ไปจนถึง p และ n คือจำนวนมิติของข้อมูล [13], [14]

2. วิธีการดำเนินการวิจัย

การวิจัยครั้งนี้ใช้วิธีการจำลองข้อมูลเพื่อศึกษาประสิทธิภาพของวิธีการสร้างชั้นภูมิด้วยอัลกอริทึมเคมีนกับการสร้างชั้นภูมิโดยใช้วิธีดาเลเนียสและฮอดจ์หรือรากที่สองของความถี่สะสมของตัวแปรเชิงปริมาณ และเปรียบเทียบโดยใช้ค่าประมาณค่าเฉลี่ยประชากร ในการวิเคราะห์ข้อมูล

ใช้โปรแกรม SAS Version 9.4 และ SAS Enterprise Miner version 14.2 มีขั้นตอนการวิจัย ดังนี้

1) จำลองข้อมูลด้วยเทคนิคมอนติคาร์โลประกอบด้วยตัวแปรทั้งหมด 9 ตัวแปร ประกอบด้วย ตัวแปรที่สนใจ y ซึ่งเป็นตัวแปรที่ต้องการหาค่าประมาณ ตัวแปรช่วยซึ่งเป็นตัวแปรเชิงกลุ่มจำนวน 1 ตัวแปร (x_8) จำนวน 5 กลุ่ม และตัวแปรเชิงปริมาณจำนวน 7 ตัวแปร $(x_1 - x_7)$ มีค่าเฉลี่ยเท่ากับ 50 ความแปรปรวนเท่ากับ 20 ค่าสัมประสิทธิ์สหสัมพันธ์ระหว่างตัวแปรช่วย x_i กับตัวแปร y มีค่าสูงสุดเท่ากับ 0.50 0.70 และ 0.90 จำนวนหนึ่งตัวแปรและตัวแปรอื่นๆ มีค่าแตกต่างไปจากตัวแปรที่มีค่าสัมประสิทธิ์สหสัมพันธ์สูงสุด โดยทำการจำลองข้อมูล 1,000,000 หน่วย

2) กำหนดจำนวนชั้นภูมิ ซึ่งจำนวนชั้นภูมิส่วนใหญ่จะกำหนดตามลักษณะภูมิศาสตร์หรือตำแหน่งที่อยู่ของหน่วยตัวอย่าง ซึ่งเป็นที่ทราบกันดีว่าจำนวนชั้นภูมิที่มากขึ้นส่งผลให้ความแปรปรวนของตัวประมาณมีค่าลดลง และสำหรับการกำหนดจำนวนชั้นภูมิด้วยวิธีอื่นๆ นั้น คอแครนเสนอข้อสังเกตว่าจำนวนชั้นภูมิ L มากกว่า 6 ชั้นภูมิจะไม่มีประโยชน์ในการลดลงของความแปรปรวนของตัวประมาณ [15] ดังนั้นสำหรับการวิจัยครั้งนี้กำหนดจำนวนชั้นภูมิเป็น 4, 5 และ 6 ชั้นภูมิ

3) กำหนดตัวแปรช่วยในการสร้างชั้นภูมิโดยวิธีดาเลเนียสและฮอดจ์หรือรากที่สองของความถี่สะสมของตัวแปรเชิงปริมาณจาก $x_1 - x_7$ โดยพิจารณาจากตัวแปรที่มีความสัมพันธ์กับตัวแปร y มากที่สุดเป็นตัวแปรช่วยในการแบ่งชั้นภูมิและสร้างชุดข้อมูลที่มีการแบ่งชั้นภูมิ 3 ขนาดคือ แบ่งเป็น 4, 5 และ 6 ชั้นภูมิ และสร้างชั้นภูมิด้วยเทคนิคการจัดกลุ่มข้อมูลของตัวแปร $x_1 - x_8$ ด้วยอัลกอริทึมเคมีนซึ่งทำการสร้างชุดข้อมูลที่มีการแบ่งชั้นภูมิเป็น 3 ขนาดเช่นกัน

4) กำหนดขนาดตัวอย่าง n ขนาด 50, 100, 150, 200 และ 300 หน่วยตัวอย่าง การกำหนดตัวอย่างในแต่ละชั้นภูมิของทั้งสองวิธีเป็นสัดส่วนกับขนาดของหน่วยตัวอย่างในชั้นภูมิและสุ่มตัวอย่างจากแต่ละชั้นภูมิด้วยวิธีการสุ่มตัวอย่างแบบง่ายไม่คืนที่ตามขนาดตัวอย่าง n ที่คำนวณได้ และกระทำซ้ำ 1,000 รอบ ในแต่ละสถานการณ์ และประมาณค่าเฉลี่ยประชากรโดยการหาผลรวมของค่าเฉลี่ยแต่ละชั้นภูมิคูณกับค่าน้ำหนักของแต่ละชั้นภูมิ คำนวณค่าความคลาดเคลื่อนกำลังสองเฉลี่ย (*MSE*) ค่าความคลาดเคลื่อนสัมพัทธ์ของตัวประมาณค่า (*Relative Error*)

5) ทดสอบแนวคิดโดยใช้ข้อมูลจริงซึ่งเป็นปริมาณฝุ่นละอองในอากาศในเมืองต่างๆ ของประเทศจีนที่จัดเก็บข้อมูลโดย U.S. Diplomatic เป็นเวลา 3 ปี [16] ชุดข้อมูลจัดเก็บอยู่ที่ UCI Machine Learning Repository ทำการแบ่งชั้นภูมิโดยวิธีรากที่สองของความถี่สะสมสำหรับตัวแปรเชิงปริมาณ และเทคนิคการจัดกลุ่มข้อมูลด้วยอัลกอริทึมเคมีนทำการสุ่มตัวอย่างจากชั้นภูมิด้วยวิธีการสุ่มตัวอย่างแบบง่ายไม่คืนที่ตามขนาดตัวอย่างที่คำนวณได้และกระทำซ้ำ 100 รอบในแต่ละสถานการณ์ และคำนวณตัวประมาณค่าเฉลี่ยประชากร

3. ผลการวิจัย

3.1 ผลการวิจัยจากข้อมูลจำลอง

จากข้อมูลจำลองทั้ง 3 ชุด ข้อมูลที่มีค่าสัมประสิทธิ์สหสัมพันธ์ระหว่างตัวแปรช่วย x_i กับตัวแปร y มีค่าสูงสุดเท่ากับ 0.50, 0.70 และ 0.90 จำนวนหนึ่งตัวแปรและตัวแปรอื่นๆ มีค่าแตกต่างไปจากตัวแปรที่มีค่าสัมประสิทธิ์สหสัมพันธ์สูงสุดซึ่งผู้วิจัยกำหนดให้ตัวแปร x_7 มีค่าสัมประสิทธิ์สหสัมพันธ์สูงสุดและมีขนาดประชากรเท่ากับ 1,000,000 หน่วย ผู้วิจัยได้ดำเนินการสร้างชั้นภูมิของข้อมูลแต่ละชุดโดยใช้วิธีดาเลเนียสและฮอดจ์หรือรากที่สองของความถี่สะสมโดยใช้ตัวแปรที่มีค่าสัมประสิทธิ์สหสัมพันธ์ของตัวแปรช่วยกับตัวแปร y มากที่สุดเป็นตัวแปรช่วยในการสร้างชั้นภูมิแสดงค่าสัมประสิทธิ์สหสัมพันธ์ดังตารางที่ 1 และการสร้างชั้นภูมิของข้อมูลแต่ละชุดโดยใช้อัลกอริทึมเคมีน ได้ผลการสร้างชั้นภูมิซึ่งสามารถแสดงร้อยละของจำนวนหน่วยตัวอย่างของประชากรในแต่ละชั้นภูมิแสดงผลดังตารางที่ 2

การสุ่มตัวอย่างแบบแบ่งชั้นภูมิโดยสุ่มตัวอย่างจากแต่ละชั้นภูมิด้วยวิธีการสุ่มตัวอย่างแบบง่ายไม่คืนที่ และกำหนดการสร้างชั้นภูมิด้วยอัลกอริทึมเคมีน และการสร้างชั้นภูมิโดยใช้วิธีรากที่สองของความถี่สะสมสำหรับตัวแปร

เชิงปริมาณโดยกำหนดตัวแปรที่มีค่าสัมประสิทธิ์สหสัมพันธ์ของตัวแปรช่วย x_i กับตัวแปร y สูงสุดเป็นตัวแปรในการสร้างชั้นภูมิดังตารางที่ 1 และคำนวณค่าเฉลี่ยประชากรเท่ากับ 50.00 คำนวณค่าประมาณค่าเฉลี่ยประชากรจากการสุ่มตัวอย่างและเปรียบเทียบประสิทธิภาพของตัวประมาณค่าโดยใช้เกณฑ์ค่าความคลาดเคลื่อนกำลังสองเฉลี่ยโดยที่ตัวประมาณ $\hat{\theta}_1$ จะมีประสิทธิภาพมากกว่า $\hat{\theta}_2$ ก็ต่อเมื่อ $MSE(\hat{\theta}_1) < MSE(\hat{\theta}_2)$ และเกณฑ์ค่าความคลาดเคลื่อนสัมพัทธ์ซึ่งผลการวิจัยแสดงผลดังตารางที่ 3

ตารางที่ 1 สัมประสิทธิ์สหสัมพันธ์ของตัวแปรช่วย $x_1 - x_7$ กับตัวแปร y ของข้อมูลจำลอง

	x_1	x_2	x_3	x_4	x_5	x_6	x_7
y	0.40	0.39	0.20	0.19	0.30	0.29	0.50
y	0.49	0.40	0.49	0.39	0.49	0.60	0.70
y	0.49	0.70	0.59	0.79	0.50	0.69	0.90

จากตารางที่ 3 พบว่าเมื่อพิจารณาค่าความคลาดเคลื่อนกำลังสองเฉลี่ย พบว่าตัวประมาณค่าเฉลี่ยประชากรโดยการสร้างชั้นภูมิด้วยอัลกอริทึมเคมีนมีค่าความคาดเคลื่อนกำลังสองเฉลี่ยน้อยกว่าค่าความคาดคลาดเคลื่อนกำลังสองเฉลี่ยของตัวประมาณจากการสร้างชั้นภูมิโดยใช้วิธีดาเลเนียสและฮอดจ์หรือรากที่สองของความถี่สะสมสำหรับตัวแปรเชิงปริมาณในทุกกรณี ทั้งขนาดตัวอย่าง จำนวนชั้นภูมิ และค่าสัมประสิทธิ์สหสัมพันธ์ และมีค่าความคลาดเคลื่อนกำลังสองเฉลี่ยลดลงเรื่อยๆ เมื่อขนาดตัวอย่าง และจำนวนชั้นภูมิเพิ่มมากขึ้น

สำหรับกรณีการสร้างชั้นภูมิด้วยอัลกอริทึมเคมีนจะพบว่าเมื่อขนาดตัวอย่างในการสร้างชั้นภูมิมีค่าเพิ่มขึ้นโดยค่าสัมประสิทธิ์สหสัมพันธ์เท่ากัน และจำนวนชั้นภูมิเท่ากันจะส่งผลให้ค่าคาดเคลื่อนกำลังสองเฉลี่ยลดลงตามลำดับขนาดตัวอย่าง เมื่อค่าสัมประสิทธิ์สหสัมพันธ์ระหว่างตัวแปรอิสระกับตัวแปรตามมีค่ามากขึ้นโดยขนาดตัวอย่าง

ตารางที่ 2 ร้อยละของจำนวนหน่วยตัวอย่างของประชากรในแต่ละชั้นภูมิของข้อมูลจำลอง

วิธี	L	r	ร้อยละของจำนวนหน่วยตัวอย่างของประชากรในแต่ละชั้นภูมิ					
			N_1	N_2	N_3	N_4	N_5	N_6
Dalenius	4	0.50	17.18	30.99	34.83	17.00	-	-
		0.70	16.38	32.43	33.31	17.89	-	-
		0.90	16.78	32.60	33.13	17.50	-	-
	5	0.50	11.60	23.01	29.28	24.70	11.42	-
		0.70	10.94	24.29	27.32	25.33	12.12	-
		0.90	12.23	23.53	27.31	25.12	11.81	-
	6	0.50	8.14	17.77	22.26	24.57	18.52	8.74
		0.70	7.65	18.80	22.36	22.76	19.91	8.52
		0.90	8.63	18.34	22.41	22.67	18.89	9.07
K-means	4	0.50	27.15	25.45	23.69	23.71	-	-
		0.70	27.08	24.50	26.88	21.54	-	-
		0.90	35.18	14.31	18.16	32.36	-	-
	5	0.50	19.75	19.17	19.57	19.48	22.04	-
		0.70	22.16	19.63	22.17	15.12	20.93	-
		0.90	30.69	20.12	28.82	6.97	13.40	-
	6	0.50	16.07	16.05	17.08	17.06	17.85	15.89
		0.70	16.19	19.04	11.77	17.20	15.48	20.33
		0.90	19.83	4.72	14.85	27.93	25.28	7.39

ตารางที่ 3 ค่าประมาณค่าเฉลี่ยประชากรจากการสุ่มตัวอย่างและการเปรียบเทียบประสิทธิภาพของตัวประมาณค่าจากข้อมูลจำลอง

r	วิธี	n	จำนวน 4 ชั้นภูมิ			จำนวน 5 ชั้นภูมิ			จำนวน 6 ชั้นภูมิ		
			$\bar{y}_{st}$	MSE	$RE(\bar{y}_{st})$	$\bar{y}_{st}$	MSE	$RE(\bar{y}_{st})$	$\bar{y}_{st}$	MSE	$RE(\bar{y}_{st})$
0.50	Dalenius	50	50.31	21.39	-	49.99	19.84	-	49.99	21.16	-
		100	50.28	10.43	-	50.00	10.79	-	50.00	11.07	-
		150	50.00	7.44	-	50.00	6.98	-	49.99	6.87	-
		200	50.00	5.45	-	50.21	5.21	-	50.00	5.64	-
		300	50.00	3.77	-	50.14	3.81	-	50.00	3.91	-
	K-means	50	50.00	14.47	147.82	50.00	14.44	137.40	50.02	14.45	146.44
		100	50.00	7.65	136.34	50.00	7.15	150.91	49.99	7.08	156.36
		150	50.00	5.07	146.75	50.00	5.07	137.67	50.00	5.07	135.50
		200	50.00	3.93	138.68	50.00	4.14	125.85	50.00	4.31	130.86
		300	50.00	2.83	133.22	50.00	2.83	134.63	50.00	2.81	139.15
0.70	Dalenius	50	50.01	16.99	-	50.01	15.10	-	50.00	17.04	-
		100	50.00	8.15	-	50.00	8.90	-	50.00	7.96	-
		150	50.00	5.31	-	50.00	5.83	-	50.00	5.59	-
		200	49.99	4.14	-	50.00	4.37	-	50.00	4.15	-
		300	50.00	3.22	-	50.00	3.09	-	50.00	2.95	-
	K-means	50	50.01	7.94	213.98	50.00	7.51	201.07	49.99	6.26	272.20
		100	50.00	4.31	189.10	50.00	3.92	227.04	50.00	3.77	211.14
		150	50.01	2.76	192.39	50.00	2.74	212.77	50.00	2.43	230.04
		200	50.00	2.31	179.22	50.00	2.17	201.38	50.00	1.92	216.15
		300	50.00	1.71	188.30	50.00	1.57	196.82	50.00	1.56	189.10
0.90	Dalenius	50	49.97	7.20	-	49.99	6.98	-	50.04	6.91	-
		100	50.00	3.41	-	50.00	3.15	-	49.99	2.98	-
		150	50.00	2.32	-	50.00	2.21	-	50.00	2.00	-
		200	50.00	1.94	-	50.01	1.77	-	50.00	1.67	-
		300	50.00	1.49	-	50.00	1.44	-	50.00	1.30	-
	K-means	50	50.01	3.20	225.00	50.01	3.13	223.00	50.00	3.44	200.80
		100	50.00	1.93	176.68	50.00	1.50	210.00	50.00	1.49	200.00
		150	50.01	1.42	163.38	50.00	1.18	187.29	50.00	1.03	194.17
		200	50.01	1.28	151.56	50.00	1.04	170.19	50.00	0.98	170.41
		300	50.00	0.99	150.51	50.00	0.86	167.44	50.00	0.77	168.83

เท่ากัน และจำนวนชั้นภูมิเท่ากันจะส่งผลให้ค่าคลาดเคลื่อนกำลังสองเฉลี่ยลดลงตามลำดับค่าสัมประสิทธิ์สหสัมพันธ์ และเมื่อจำนวนชั้นภูมิเพิ่มมากขึ้นโดยขนาดตัวอย่างเท่ากันและค่าสัมประสิทธิ์สหสัมพันธ์เท่ากัน จะส่งผลให้ค่าความคลาดเคลื่อนกำลังสองเฉลี่ยลดลงด้วยเช่นเดียวกัน

นอกจากนี้เมื่อพิจารณาค่าความคลาดเคลื่อนสัมพัทธ์ $RE(\bar{y}_{st})$ พบว่าการสร้างชั้นภูมิด้วยอัลกอริทึมเคมีนจะมีประสิทธิภาพสูงสุดในทุกกรณี โดยให้ค่าร้อยละของสัดส่วนของค่าความคลาดเคลื่อนของค่าประมาณระหว่างการสร้างชั้นภูมิด้วยวิธีดาเลเนียสและฮอดจ์หรือรากที่สองของความถี่สะสมกับการสร้างชั้นภูมิด้วยอัลกอริทึมเคมีนที่มากกว่าร้อยละ 100 ในทุกกรณี และสามารถสังเกตได้ว่าเมื่อค่าสัมประสิทธิ์สหสัมพันธ์เพิ่มขึ้นโดยขนาดตัวอย่างเท่ากัน และจำนวนชั้นภูมิเท่ากันจะทำให้ค่าความคลาดเคลื่อนสัมพัทธ์

$RE(\bar{y}_{st})$ มีค่ามากขึ้นและจะมีค่าลดลงเมื่อค่าสัมประสิทธิ์สหสัมพันธ์สูงมากๆ โดยเฉพาะเมื่อค่าสัมประสิทธิ์สหสัมพันธ์เท่ากับ 0.90

3.2 ผลการวิจัยจากข้อมูลปริมาณฝุ่นละอองในอากาศซึ่งเป็นข้อมูลจริง

ผู้วิจัยใช้ข้อมูลฝุ่นละอองในอากาศซึ่งเป็นข้อมูลจริงในการทดสอบผลการวิจัยและข้อมูลปริมาณฝุ่นละอองในอากาศมีขนาดประชากรเท่ากับ 11,769 หน่วย ตัวแปรที่ใช้ในการวิจัยจำนวน 11 ตัวแปร ประกอบด้วย x_1 เป็นอุณหภูมิจุดน้ำค้าง x_2 เป็นอุณหภูมิทั่วไป x_3 เป็นความชื้น x_4 เป็นความกดอากาศ x_5 เป็นความเร็วลมสะสม x_6 เป็นการเร่งรัดรายชั่วโมง x_7 เป็นปริมาณฝนสะสม x_8 เป็นทิศทางลมจำนวน 5 ระดับ x_9 เป็นฤดูกาลจำนวน 4 ระดับ x_{10} เป็นเมืองต่างๆ ของประเทศจีนจำนวน 5 ระดับ และตัวแปรสุดท้าย y เป็นค่าฝุ่นละอองในอากาศกำหนดเป็นตัวแปรที่ต้องการประมาณค่า ตัวแปรช่วยในการสร้างชั้นภูมิโดยวิธีดาเลเนียสและฮอดจ์หรือรากที่สองของความถี่สะสมสำหรับตัวแปรเชิงปริมาณ $x_1 - x_7$ และค่าสัมประสิทธิ์สหสัมพันธ์ของตัวแปรช่วย $x_1 - x_7$ กับปริมาณฝุ่นละอองในอากาศ y เท่ากับ 0.12, 0.10, 0.09, 0.16, 0.14, 0.08 และ 0.18 ตามลำดับ การสร้างชั้นภูมิด้วยเทคนิคการจัดกลุ่มข้อมูลของตัวแปร $x_1 - x_{10}$ ด้วยอัลกอริทึมเคมีน แสดงจำนวนประชากรในแต่ละชั้นภูมิได้ดังตารางที่ 4

จากข้อมูลปริมาณฝุ่นละอองในอากาศซึ่งเป็นข้อมูลจริงพบว่าข้อมูลจะมีค่าสัมประสิทธิ์สหสัมพันธ์ระหว่างตัวแปรช่วย $x_1 - x_7$ กับตัวแปร y น้อยประมาณเท่ากับ 0.10 สามารถคำนวณค่าเฉลี่ยประชากรเท่ากับ 34.64 เมื่อคำนวณค่าประมาณค่าเฉลี่ยประชากรจากการสุ่มตัวอย่างและเปรียบเทียบประสิทธิภาพของตัวประมาณค่าดังตารางที่ 5 พบว่าตัวประมาณค่าเฉลี่ยประชากรโดยการสร้างชั้นภูมิด้วย

ตารางที่ 4 ร้อยละของจำนวนหน่วยตัวอย่างของประชากรในแต่ละชั้นภูมิของข้อมูลปริมาณฝุ่นละอองในอากาศซึ่งเป็นข้อมูลจริง

วิธี	L	ร้อยละของจำนวนหน่วยตัวอย่างของประชากรในแต่ละชั้นภูมิ					
		N_1	N_2	N_3	N_4	N_5	N_6
Dalenius	4	64.81	21.33	10.10	3.76	-	-
	5	55.39	23.21	12.48	6.38	2.53	-
	6	48.57	23.76	13.81	7.27	4.79	1.81
K-means	4	56.48	35.78	2.43	5.31	-	-
	5	34.03	2.43	4.68	54.02	4.84	-
	6	4.71	46.35	1.64	14.05	32.42	0.83

ตารางที่ 5 ค่าประมาณค่าเฉลี่ยประชากรจากการสุ่มตัวอย่างและการเปรียบเทียบประสิทธิภาพของตัวประมาณค่าของข้อมูลปริมาณฝุ่นละอองในอากาศซึ่งเป็นข้อมูลจริง

วิธี	n	จำนวน 4 ชั้นภูมิ			จำนวน 5 ชั้นภูมิ			จำนวน 6 ชั้นภูมิ		
		$\bar{y}_{st}$	MSE	$RE(\bar{y}_{st})$	$\bar{y}_{st}$	MSE	$RE(\bar{y}_{st})$	$\bar{y}_{st}$	MSE	$RE(\bar{y}_{st})$
Dalenius	50	34.38	12.68	-	34.39	11.19	-	34.68	11.12	-
	100	34.28	7.73	-	34.43	5.99	-	34.30	5.60	-
	150	34.44	4.14	-	34.45	4.10	-	34.90	3.95	-
	200	34.54	3.12	-	34.48	2.78	-	34.39	2.61	-
	300	34.73	2.32	-	34.51	1.96	-	34.89	1.45	-
K-means	50	34.11	8.67	146.25	34.43	8.22	136.13	35.10	7.53	147.68
	100	34.02	4.07	189.93	34.42	4.03	148.64	34.63	3.86	145.08
	150	34.63	3.08	134.42	34.63	3.08	133.12	34.99	3.03	130.36
	200	34.84	3.03	102.97	34.57	2.34	118.80	34.41	2.14	121.96
	300	34.91	1.81	128.18	34.48	1.76	111.36	34.59	1.27	114.17

อัลกอริทึมเคมีนมีค่าความคาดเคลื่อนกำลังสองเฉลี่ยน้อยกว่าค่าความคาดคลาดเคลื่อนกำลังสองเฉลี่ยของตัวประมาณจากการสร้างชั้นภูมิโดยใช้วิธีดาเลเนียสและฮอดจ์หรือรากที่สองของความถี่สะสมสำหรับตัวแปรเชิงปริมาณในทุกกรณี ทั้งขนาดตัวอย่าง และจำนวนชั้นภูมิ

นอกจากนี้เมื่อพิจารณาค่าความคลาดเคลื่อนสัมพัทธ์ $RE(\bar{y}_{st})$ พบว่าการสร้างชั้นภูมิด้วยอัลกอริทึมเคมีนจะมีประสิทธิภาพสูงสุดในทุกกรณี โดยให้ค่าร้อยละของสัดส่วนของค่าความคาดเคลื่อนของค่าประมาณระหว่างการสร้างชั้นภูมิด้วยวิธีดาเลเนียสและฮอดจ์หรือรากที่สองของความถี่สะสมกับการสร้างชั้นภูมิด้วยอัลกอริทึมเคมีนที่มากกว่าร้อยละ 100 ในทุกกรณี

4. อภิปรายผลและสรุป

ตัวประมาณค่าเฉลี่ยจากการสุ่มตัวอย่างแบบแบ่งชั้นภูมิโดยสุ่มตัวอย่างจากแต่ละชั้นภูมิด้วยวิธีการสุ่มตัวอย่างแบบง่ายไม่คืนที่และกำหนดการสร้างชั้นภูมิด้วยอัลกอริทึมเคมีนที่ผู้วิจัยนำเสนอจะมีประสิทธิภาพมากกว่าการสร้างชั้นภูมิโดยใช้วิธีดาเลเนียสและฮอดจ์หรือรากที่สองของความถี่สะสมสำหรับตัวแปรเชิงปริมาณในทุกกรณี ทั้งขนาดตัวอย่าง จำนวนชั้นภูมิ และค่าสัมประสิทธิ์สหสัมพันธ์ โดยเมื่อขนาดตัวอย่างเพิ่มขึ้น จำนวนชั้นภูมิเพิ่มขึ้น และค่าสัมประสิทธิ์สหสัมพันธ์ระหว่างตัวแปรอิสระกับตัวแปรตามมีค่ามากขึ้นจะส่งผลให้ค่าคาดเคลื่อนกำลังสองเฉลี่ยลดลง สำหรับการสร้างชั้นภูมิโดยใช้วิธีดาเลเนียสและฮอดจ์หรือรากที่สองของความถี่สะสมจะมีประสิทธิภาพดีที่สุดเมื่อค่าสัมประสิทธิ์สหสัมพันธ์ระหว่างตัวแปรอิสระกับตัวแปรตามมีค่าสูงมากๆ แต่การสร้างชั้นภูมิด้วยอัลกอริทึมเคมีนที่ผู้วิจัยนำเสนอจะมีประสิทธิภาพจะมีประสิทธิภาพดีสำหรับทุกค่าสัมประสิทธิ์สหสัมพันธ์ และมีประสิทธิภาพดีกว่าวิธีดาเลเนียสและฮอดจ์หรือรากที่สองของความถี่สะสมในทุกกรณี

กรณีการวิจัยเชิงสำรวจจะนิยมทำการสร้างชั้นภูมิตามลักษณะภูมิศาสตร์ เพื่อความสะดวกในการลงพื้นที่เก็บข้อมูลวิจัย การคำนวณค่าใช้จ่าย การนำเสนอผลการวิจัยตามลักษณะภูมิศาสตร์ เป็นต้น แต่สำหรับงานวิจัยบางเรื่องที่ไม่มีตัวแปรเชิงภูมิศาสตร์หรือตำแหน่งที่อยู่ของหน่วยตัวอย่างเป็นตัวแปรในการวิจัย ผู้วิจัยจำเป็นต้องใช้ค่าของตัวแปรช่วย x ที่เรียกว่าตัวแปรแบ่งชั้นภูมิตัวใดตัวหนึ่ง ซึ่งมีความสัมพันธ์สูงกับตัวแปร y ที่สนใจแต่เป็นที่คาดหวังได้ว่าตัวแปร x มีความสัมพันธ์กับ y สูงจะให้ค่าประมาณของ y แม่นยำขึ้น ซึ่งวิธีการดังกล่าวจะเป็นการสร้างชั้นภูมิที่ขึ้นกับตัวแปรช่วยเพียงตัวเดียว เทคนิคการสร้างชั้นภูมิโดยใช้เทคนิคเคมีน จึงเป็นอีกวิธีหนึ่งที่สามารถสร้างชั้นภูมิโดยใช้ตัวแปร x หลายตัวในการจัดกลุ่มของหน่วยตัวอย่างในแต่ละชั้นภูมิ ซึ่งจะทำให้การแบ่งชั้นภูมิด้วยวิธีนี้ให้ประสิทธิภาพของตัวประมาณค่าดีกว่าการสร้างชั้นภูมิโดยวิธีของดาเลเนียสและฮอดจ์

เอกสารอ้างอิง

[1] P. Suwatthee, *Sample Surveys: Sampling Designs and Analysis*. Bangkok: National Institute of Development Administration, 2009.

[2] K. Silpakob and W. Chaimongkol, "Estimation of population mean with missing data in stratified random sampling," *Burapha Science Journal*, vol. 22, no. 2, pp. 202–217, 2017.

[3] M. E. Thompson, *Theory of Sample Surveys.* London: Chapman & Hall, 1997.

[4] M. H. Hansen, W. N. Hurwitz, and W. G. Madow, *Sample Survey Methods and Theory*. Canada: John Wiley & Sons, lnc., 1960.

[5] P. Suwatthee, *Theory of Sampling Designs*. Bangkok: National Institute of Development Administration, 2011.

[6] T. Dalenius and J. L. Hodges, "Minimum variance stratification," *Journal of the American Statistical Association*, no. 285, pp. 88, 1959.

[7] Y. Olufadi, "Dual to ratio-cum-product estimator in simple and stratified random sampling," *Pakistan Journal of Statistics and Operation Research*, vol. 9, no. 3, pp. 305–319, 2013.

[8] A. K. Gupta and D. G. Kabe, *Theory of Sample Surveys*. Singapore: World Scientific Publishing, 2011.

[9] T. Dalenius, "A First Couse in Survey Sampling," in P. krishnaiash and C. R. Rao (Eds.), *Handbook of Statistics Volume 6: Sampling*, North-Holland: Elsevier Science B.V., 1988, pp. 15–46.

[10] S. Wichaidit, "DNA microarray data analysis model using clustering algorithm for disease diagnosis," M.S. thesis, Department of Computer Science, Faculty of Science, Price of Songkla University, 2008.

[11] W. Chongnguluam, "Parallelize rough k-medoids clustering on multicore processor," M.S. thesis, School of Computer Engineering, Faculty of Engineering, Suranaree University of Technology, 2012.

[12] W. Pimpaporn and P. Meesad, "A comparative efficiency of clustering using dynamic feature selection optimization of subspace clustering algorithms," *Information Technology Journal*, vol. 10, no. 2, pp. 43–51, 2014.

[13] J. Kaufman and P. J. Rousseeuw, *Finding Groups in Data: An Introduction to Cluster Analysis*. New York: John Wiley, 1990.

[14] G. J. Myatt and W. P. Johnson, *Making Sense of Data II: A Practical Guide to Data Visualization, Advanced Data Mining Methods, and Applications*. Canada: John Wiley & Sons, Inc. 2009.

[15] W. G. Cochran, *Sampling Techniques*, 3rd ed. New York : Wiley, c1977., 1977.

[16] X. Liang, S. Li, S. Zhang, H. Huang, and S. X. Chen, "$PM_{2.5}$ data reliability, consistency, and air quality assessment in five Chinese cities," *Journal of Geophysical Research: Atmospheres*, vol. 121, no. 17, 2016.