บทความวิชาการ

แบบจำลองสำหรับค่าสุดขีด: แบบจำลองสถิติอันดับ *r* อันดับที่ใหญ่ที่สุด

พรรณรัตน์ ก้วยเจริญพานิชก์
คณะสหวิทยาการ วิทยาเขตหนองคาย มหาวิทยาลัยขอนแก่น
ประภาวรรณ เสนาเพ็ง และ ปิยภัทร บุษบาบดินทร์*
ภาควิชาคณิตศาสตร์ คณะวิทยาศาสตร์ มหาวิทยาลัยมหาสารคาม

* ผู้นิพนธ์ประสานงาน โทรศัพท์ 08 9542 6396 อีเมล: piyapatr.b@msu.ac.th DOI: 10.14416/j.kmutnb.2023.07.010
รับเมื่อ 13 มิถุนายน 2564 แก้ไขเมื่อ 9 สิงหาคม 2564 ตอบรับเมื่อ 13 สิงหาคม 2564 เผยแพร่ออนไลน์ 13 กรกฎาคม 2566

บทคัดย่อ

การวิเคราะห์ค่าสุดขีด (Extreme Value Analysis) เป็นการวิเคราะห์ข้อมูลที่มีค่าสูงสุดหรือต่ำสุด โดยเป็นข้อมูลที่อยู่ในส่วนปลายหางของข้อมูลซึ่งมีจำนวนข้อมูลน้อยมาก ทำให้ผลของการวิเคราะห์มีความคลาดเคลื่อนส่งผลต่อการประมาณค่าพารามิเตอร์ของแบบจำลองที่ผิดพลาดและการประมาณระดับการเกิดซ้ำไม่แม่นยำเท่าที่ควร ดังนั้นการลดความคลาดเคลื่อนที่เกิดจากการมีจำนวนข้อมูลน้อยจึงมีความจำเป็นอย่างยิ่ง ซึ่งวิธีการที่ใช้กันอย่างแพร่หลาย คือ การเพิ่มขนาดหรือเพิ่มจำนวนข้อมูล เป็นการเพิ่มข้อมูลเข้าไปในการวิเคราะห์ค่าสุดขีด ซึ่งจะสามารถเพิ่มได้เฉพาะกรณีที่ผู้วิเคราะห์เลือกใช้วิธีบล็อกเวลา (Block Time Method) ในการวิเคราะห์เท่านั้น โดยการแจกแจงที่ใช้ในกรณีนี้เรียกว่า การแจกแจงค่าสุดขีดนัยทั่วไปสำหรับสถิติอันดับ r อันดับที่ใหญ่ที่สุด (Generalized Extreme Value Distribution for the r Largest Order Statistics; GEVr) ถือว่าเป็นวิธีการที่จะทำให้แบบจำลองที่ได้จากการวิเคราะห์นี้มีความเหมาะสมกับข้อมูลมากขึ้น และการประมาณค่าพารามิเตอร์และการประมาณระดับการเกิดซ้ำแม่นยำมากยิ่งขึ้น

คำสำคัญ: ค่าสุดขีด วิธีบล็อกเวลา การแจกแจงค่าสุดขีดนัยทั่วไป สถิติอันดับ การแจกแจงค่าสุดขีดนัยทั่วไปสำหรับสถิติอันดับ r อันดับที่ใหญ่ที่สุด

การอ้างอิงบทความ: พรรณรัตน์ ก้วยเจริญพานิชก์, ประภาวรรณ เสนาเพ็ง และ ปิยภัทร บุษบาบดินทร์, “แบบจำลองสำหรับค่าสุดขีด: แบบจำลองสถิติอันดับ r อันดับที่ใหญ่ที่สุด,” *วารสารวิชาการพระจอมเกล้าพระนครเหนือ*, ปีที่ 33, ฉบับที่ 3, หน้า 1–10, เลขที่บทความ 233-285155, ก.ค.–ก.ย. 2566.

Academic Article

Extreme Value Model: The *r* Largest Order Statistic Model

Pannarat Guayjarernpanishk
Faculty of Interdisciplinary Studies, Nong Khai Campus, Khon Kaen University, Nong Khai, Thailand
Prapawan Senapeng and Piyapatr Busababodhin*
Department of Mathematics, Faculty of Science, Mahasarakham University, Maha Sarakham, Thailand

* Corresponding Author, Tel. 08 9542 6396, E–mail: piyapatr.b@msu.ac.th DOI: 10.14416/j.kmutnb.2023.07.010
Received 13 June 2021; Revised 9 August 2021; Accepted 13 August 2021; Published online: 13 July 2023

Abstract

Extreme Value Analysis is the analysis of data with the largest and smallest values in the data set. They are at the tail end of the data distribution. Only a small fraction available can give results that are very misleading, resulting in inaccuracy in parameter estimation of the model as well as inaccurate return level estimates. Reduction of the aberration due to such a small amount of data is essential. A widely used method of reducing the error is to increase the sample size or increase the amount of data into extreme value analysis. Data can only be added if the block time method is used in the analysis. The distribution in this case is called the Generalized Extreme Value Distribution for the r largest order statistics (GEVr). The model derived from the analysis appears to be more suitable for the data. Moreover, parameter estimation and return level estimation would yield more reliable results.

Keywords: Extreme Value, Block Time Method, Generalized Extreme Value Distribution, Order Statistic, Generalized Extreme Value Distribution for the r Largest Order Statistics

Please cite this article as: P. Guayjarernpanishk, P. Senapeng, and P. Busababodhin, "Extreme value model: The *r* largest order statistic model," *The Journal of KMUTNB*, vol. 33, no. 3, pp. 1–10, ID. 233-285155, Jul.–Sep. 2023 (in Thai).

1. บทนำ

การหาแบบจำลองค่าสุดขีด (Extreme Value Model) สำหรับข้อมูลที่มีลักษณะสุดขีดหรือมีค่าสุดขีด ไม่ว่าจะเป็นค่าสูงสุดขีดหรือค่าต่ำสุดขีด ในปัจจุบันมีการประยุกต์ใช้กันอย่างแพร่หลายในหลากหลายสาขา เช่น อุทกวิทยา อุตุนิยมวิทยา วิศวกรรมโครงสร้าง เศรษฐศาสตร์ เป็นต้น สำหรับวิธีที่ใช้ในการสร้างแบบจำลองค่าสุดขีดมีอยู่ 2 วิธี คือ วิธีบล็อกเวลา (Block Time Method) และวิธีเกณฑ์ (Threshold Method) โดยวิธีแรกจะใช้การแจกแจงค่าสุดขีดนัยทั่วไป (Generalized Extreme Value Distribution; GEV) ส่วนวิธีที่สองใช้การแจกแจงพาเรโตนัยทั่วไป (Generalized Pareto Distribution; GPD) [1] ซึ่ง Bernulli เป็นผู้ค้นพบการแจกแจงค่าสุดขีดเป็นคนแรกใน พ.ศ. 2252 โดยใน พ.ศ. 2457 Fuller ได้นำมาประยุกต์ใช้ จึงถือว่าเป็นจุดเริ่มต้นของการนำข้อมูลที่มีลักษณะสุดขีดมาวิเคราะห์ ศึกษาเพิ่มเติมได้จาก [2]–[8]

ขั้นตอนในการหาแบบจำลองค่าสุดขีด ขั้นตอนแรกเป็นขั้นตอนที่สำคัญ เพื่อให้ได้แบบจำลองที่เหมาะสมและแม่นยำที่สุด คือ ขั้นตอนในการนำข้อมูลมาวิเคราะห์ ซึ่งวิธีการนำข้อมูลมาวิเคราะห์ทั้ง 2 วิธี ที่กล่าวมาแล้วข้างต้น มีวิธีการได้มาซึ่งข้อมูลที่แตกต่างกัน โดยวิธีแรกข้อมูลที่ถูกนำมาวิเคราะห์จะเป็นข้อมูลที่มีค่าสุดขีด (ค่าสูงสุดหรือค่าต่ำสุด) ในแต่ละคาบเวลา โดยในแต่ละคาบเวลาจะมีช่วงความห่างของระยะเวลาเท่าๆ กัน เช่น ข้อมูลรายสัปดาห์ ข้อมูลรายเดือน หรือข้อมูลรายปี เป็นต้น ส่วนวิธีที่สองข้อมูลที่ถูกนำมาวิเคราะห์จะเป็นข้อมูลที่มีค่ามากกว่าค่าเกณฑ์ (Threshold) ในกรณีที่เป็นค่าสูงสุด หรือเป็นข้อมูลที่มีค่าต่ำกว่าค่าเกณฑ์ในกรณีเป็นค่าต่ำสุด ซึ่งข้อมูลโดยส่วนใหญ่ที่นำมาวิเคราะห์ในลักษณะนี้จะเป็นข้อมูลรายวัน เมื่อได้วิธีการที่เหมาะสมในการวิเคราะห์แล้ว ขั้นตอนต่อมาคือการทดสอบกระบวนการของข้อมูล ซึ่งเป็นการทดสอบว่าข้อมูลที่นำมาวิเคราะห์อยู่ภายใต้กระบวนการแบบใด ระหว่างกระบวนการคงที่ (Stationary Process) หรือกระบวนการไม่คงที่ (Non-Stationary Process) [9] เนื่องจากกระบวนการทั้งสองมีวิธีการวิเคราะห์ และการเลือกตัวแบบที่เหมาะสมแตกต่างกัน ถ้าหากข้อมูลมีตัวแปรอื่นๆ มาเกี่ยวข้องด้วย เช่น ตัวแปรด้านเวลาหรือฤดูกาล เป็นต้น การวิเคราะห์ข้อมูลก็ควรอยู่ภายใต้กระบวนการไม่คงที่ เช่น ข้อมูลทางอุตุนิยมวิทยา ซึ่งฤดูกาลส่งผลต่อข้อมูลที่นำมาวิเคราะห์ หรือข้อมูลค่าสินไหมทดแทนของการประกันภัยพืชผล ซึ่งค่าสินไหมทดแทนจะมีค่าสูงในฤดูแล้ง เป็นต้น หากไม่มีการพิจารณาลักษณะของข้อมูลอาจจะทำให้การประมาณค่าพารามิเตอร์ของแบบจำลองผิดพลาด ส่งผลต่อการนำไปใช้ โดยเฉพาะข้อมูลที่ต้องการความแม่นยำของแบบจำลองเป็นอย่างมาก สามารถศึกษาเพิ่มเติมได้จาก [10]–[14]

วัตถุประสงค์หลักในการสร้างแบบจำลอง คือ ค่าที่ได้จากแบบจำลองที่มีความแม่นยำสูง ซึ่งการที่ค่าที่ได้จากแบบจำลองจะมีความแม่นยำสูงจำเป็นต้องใช้ข้อมูลจำนวนที่เพียงพอในการวิเคราะห์ ในกรณีที่ใช้วิธีเกณฑ์ในการสร้างแบบจำลอง ถ้าข้อมูลที่ผ่านเกณฑ์มีจำนวนไม่เพียงพอสามารถแก้ปัญหาได้โดยกำหนดค่าเกณฑ์ใหม่ เพื่อให้จำนวนข้อมูลเพิ่มขึ้น เช่น ถ้าข้อมูลเป็นค่าสูงสุด อาจจะปรับค่าเกณฑ์ให้น้อยลง แต่ถ้าข้อมูลเป็นค่าต่ำสุด จะทำการปรับค่าเกณฑ์ให้มากขึ้น จากนั้นทำการวิเคราะห์ข้อมูลตามขั้นตอนโดยใช้ GPD ตามปกติ หากใช้วิธีบล็อกเวลาในการสร้างแบบจำลอง วิธีการเพิ่มจำนวนข้อมูลจะไม่เหมือนวิธีเกณฑ์ แต่สามารถเพิ่มจำนวนข้อมูลโดยการเลือกข้อมูลที่มีค่าสุดขีดในลำดับถัดๆ เข้ามาในการวิเคราะห์ เช่น เลือกข้อมูลที่มีค่าสูงสุด 5 อันดับแรกในแต่ละคาบเวลามาวิเคราะห์ ดังนั้นจำนวนข้อมูลจะเพิ่มขึ้น 5 เท่าของจำนวนข้อมูลเดิม เป็นต้น ซึ่งการเพิ่มข้อมูลเข้าในวิธีการนี้จะทำให้การวิเคราะห์ข้อมูลโดยใช้ GEV แบบเดิมไม่สามารถวิเคราะห์ได้ เนื่องจาก GEV ใช้วิเคราะห์สำหรับข้อมูลค่าสุดขีดเพียงค่าเดียวต่อคาบเวลาเท่านั้น

ดังนั้นในบทความนี้จะกล่าวถึงขั้นตอนการวิเคราะห์เมื่อใช้ข้อมูลสุดขีดของ r อันดับแรกในแต่ละคาบเวลากรณีค่าสุดขีดที่พิจารณาเป็นค่าสูงสุด หรือการหาแบบจำลองสถิติอันดับ r อันดับที่ใหญ่ที่สุด โดยจะกล่าวถึงการแจกแจงค่าสุดขีดนัยทั่วไป การแจกแจงค่าสุดขีดนัยทั่วไปสำหรับสถิติอันดับ r อันดับที่ใหญ่ที่สุด การประมาณค่าพารามิเตอร์ การเลือกแบบจำลอง และตัวอย่างการวิเคราะห์ ในหัวข้อถัดตามลำดับ

2. การแจกแจงค่าสุดขีดนัยทั่วไป

ใน พ.ศ. 2498 Jenkinson [15] ได้พัฒนาการแจกแจงค่าสุดขีดนัยทั่วไปขึ้นเป็นครั้งแรก โดยพบว่าเป็นการแจกแจงที่เหมาะสำหรับการวิเคราะห์ค่าสุดขีดในคาบเวลาที่มีความห่างของระยะเวลาเท่าๆ กัน ซึ่งวิธีการเลือกข้อมูล คือ เลือกข้อมูลที่มีค่าสุดขีด (ค่าสูงสุดหรือต่ำสุดเพียง 1 ค่า) ในแต่ละคาบเวลามาวิเคราะห์ ในกรณีค่าสุดขีดเป็นค่าสูงสุดจะมีแนวคิดดังนี้

กำหนดให้ $M_n = \max(X_1, X_2,..., X_n)$ เมื่อ $X_1, X_2,..., X_n$ เป็นตัวแปรที่อิสระกันและมีฟังก์ชันการแจกแจง $F(x)$ แบบเดียวกัน [4]

ทฤษฎีบทที่ 1 ถ้ามีลำดับค่าคงที่ $\{a_n > 0\}$ และ $\{b_n\}$ ดังสมการที่ (1)

$$\Pr\{(M_n - b_n)/a_n \le x\} \to G(x) \text{ เมื่อ } n \to 0 \quad (1)$$

สำหรับฟังก์ชันการแจกแจงไม่แปรสภาพ (Non-degenerate Distribution) G แล้วจะได้ว่า G เป็นสมาชิกของวงศ์ GVE ดังสมการที่ (2)

$$G(x) = \exp\left\{-\left[1+\xi\left(\frac{x-\mu}{\sigma}\right)\right]^{-\frac{1}{\xi}}\right\} \quad (2)$$

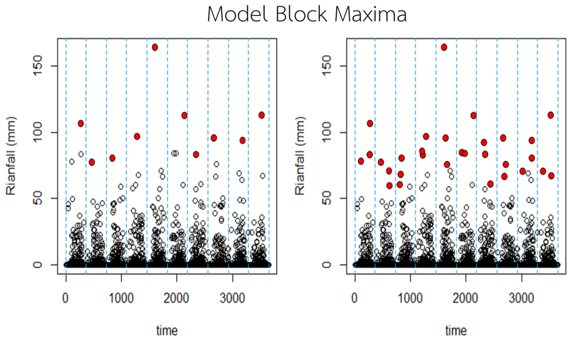
นิยามบน $\left\{x : 1+\xi\left(\frac{x-\mu}{\sigma}\right) > 0\right\}$ เมื่อ $-\infty < \mu < \infty$,

$-\infty < \xi < \infty$ และ $\sigma > 0$

ถ้า $\xi \to 0$ หรือ $\xi = 0$ จะได้

$$G(x) = \exp\left\{-\exp\left[-\left(\frac{x-\mu}{\sigma}\right)\right]\right\} \quad (3)$$

เมื่อ μ พารามิเตอร์บ่งตำแหน่ง (Location Parameter) σ พารามิเตอร์บ่งขนาด (Scale Parameter) และ ξ พารามิเตอร์บ่งรูปร่าง (Shape Parameter) จากสมการที่ (2) ถ้า $\xi < 0$ จะเรียกการแจกแจงค่าสุดขีดนัยทั่วไปว่า "การแจกแจงไวบูล (Weibull Distribution)" และถ้า $\xi > 0$



(ก) (ข)

รูปที่ 1 การเลือกค่าสูงสุดกรณีที่ r = 1 (ก) และ r = 3 (ข)

จะเรียกว่า "การแจกแจงฟรีเชท (Fréchet Distribution)" และในกรณีที่ $\xi \to 0$ จะเรียก สมการที่ (3) ว่า "การแจกแจงกัมเบล (Gumbel Distribution)

3. การแจกแจงค่าสุดขีดนัยทั่วไปสำหรับสถิติอันดับ r อันดับที่ใหญ่ที่สุด

สำหรับการแก้ปัญหากรณีที่จำนวนข้อมูลไม่เพียงพอในการวิเคราะห์หาแบบจำลอง สามารถแก้ปัญหาโดยการเพิ่มจำนวนข้อมูล ด้วยการนำข้อมูลที่มีค่าสุดขีดในอันดับรองๆ ลงมา r อันดับแรก (The r Largest Order Statistic) เช่น เลือก 3 อันดับสูงสุด (r = 3) หรือ 10 อันดับสูงสุด (r = 10) เป็นต้น มาวิเคราะห์ร่วมด้วย วิธีการนี้ใช้ได้เฉพาะกรณีที่หาแบบจำลองโดยวิธีบล็อกเวลาเท่านั้น โดยถ้า r = 1 จะได้ว่าการหาแบบจำลองค่าสุดขีดจะใช้ GEV ในการวิเคราะห์ การเลือกข้อมูลค่าสูงสุดมาวิเคราะห์เมื่อ r = 1 และ r = 3 แสดงดังรูปที่ 1

แนวคิดและวิธีการการเพิ่มจำนวนข้อมูลเป็นดังนี้

ให้ $X_1, X_2,..., X_n$ เป็นลำดับของตัวแปรที่เป็นอิสระกันและมีการแจกแจงเดียวกัน และกำหนดให้ $M_n^{(k)} = k$–th largest $\{X_1, X_2,..., X_n\}$ เป็นสถิติอันดับที่ k ที่ใหญ่ที่สุด [4]

ทฤษฎีบทที่ 2 ถ้ามีลำดับค่าคงที่ $\{a_n > 0\}$ และ $\{b_n\}$ ซึ่ง $\Pr\{(M_n - b_n)/a_n \le x\} \to G(x)$ เมื่อ $n \to 0$

สำหรับบางฟังก์ชันการแจกแจงไม่แปรสภาพ G ซึ่ง G คือ ฟังก์ชัน GEV ที่นิยามในสมการที่ (2) แล้วจะได้ว่า สำหรับ

ค่า k ใดๆ

$$\Pr\left\{\left(M_n^{(k)} - b_n\right)/a_n \le x\right\} \to G_k(x)$$

นิยามบน $\left\{x : 1+\xi\left(\frac{x-\mu}{\sigma}\right) > 0\right\}$ เมื่อ

$$G_k(x) = \exp\{-\tau(x)\}\sum_{s=0}^{k-1}\frac{\tau(x)^s}{s!} \tag{4}$$

และ $\tau(x) = \left[1+\xi\left(\frac{x-\mu}{\sigma}\right)\right]^{-\frac{1}{\xi}}$

จากทฤษฎีบทที่ 2 ถ้า $M_n^{(k)}$ หรือสถิติอันดับที่ k ที่ใหญ่ที่สุดในคาบเวลามีลักษณะทำนองเดียวกันกับค่าสูงสุดแล้วจะได้ว่า การแจกแจงในสมการที่ (4) จะมีพารามิเตอร์ต่างๆ สอดคล้องกับ GEV ซึ่งทฤษฎีบทที่ 2 เป็นกรณีเฉพาะของข้อมูลของสถิติอันดับที่ k ที่ใหญ่ที่สุดเพียงข้อมูลตัวเดียวในแต่ละคาบเวลา แต่ข้อมูลที่ต้องการนำมาวิเคราะห์ในครั้งนี้คือข้อมูลที่มีค่าสูงสุด r อันดับแรก

กำหนดให้ $\mathbf{M}_n^{(r)} = (M_n^{(1)}, M_n^{(2)},..., M_n^{(r)})$ เป็นเวกเตอร์ของสถิติอันดับ r อันดับที่ใหญ่ที่สุดในแต่ละคาบเวลา ดังนั้นสมการที่ (4) จึงเป็นเพียงวงศ์ของการแจกแจงของแต่ละสมาชิกของ $\mathbf{M}_n^{(r)}$ ไม่ใช่การแจกแจงร่วม (The Joint Distribution) ของ $\mathbf{M}_n^{(r)}$ นอกจากนั้นแล้วสมาชิกแต่ละตัวของ $\mathbf{M}_n^{(r)}$ ยังไม่เป็นอิสระกัน ($M_n^{(2)}$ ไม่มีโอกาสที่จะมีค่ามากกว่า $M_n^{(1)}$) ดังนั้นต้องทำการปรับรูป (Re-scaling) เพื่อให้ข้อมูลเป็นไปคุณสมบัติของการเป็นอิสระกัน ดังนี้

$\tilde{\mathbf{M}}_n^{(r)} = \left(\frac{M_n^{(1)}-b_n}{a_n}, ..., \frac{M_n^{(r)}-b_n}{a_n}\right)$ และจะได้การแจกแจงร่วมของ $\tilde{\mathbf{M}}_n^{(r)}$ ได้ดังทฤษฎีบทที่ 3 [4]

ทฤษฎีบทที่ 3 ถ้ามีลำดับค่าคงที่ $\{a_n > 0\}$ และ $\{b_n\}$ ซึ่ง

$$\Pr\left\{\left(M_n - b_n\right)/a_n \le x\right\} \to G(x) \quad \text{เมื่อ} \quad n \to 0$$

สำหรับบางฟังก์ชันการแจกแจงไม่แปรสภาพ G แล้วจะได้ว่า สำหรับค่า r ใดๆ ฟังก์ชันความหนาแน่นความน่าจะเป็นร่วม (The Joint Probability Density Function) ของ $\tilde{\mathbf{M}}_n^{(r)}$ เมื่อ $n \to 0$ คือ

$$f(x^{(1)},...,x^{(r)}) = \left[\begin{array}{l}\exp\left\{-\left[1+\xi\left(\frac{x^{(r)}-\mu}{\sigma}\right)\right]^{-\frac{1}{\xi}}\right\} \\ \times\prod_{k=1}^{r}\sigma^{-1}\left[1+\xi\left(\frac{x^{(r)}-\mu}{\sigma}\right)\right]^{-\frac{1}{\xi}-1}\end{array}\right] \tag{5}$$

เมื่อ $-\infty < \mu < \infty$, $-\infty < \xi < \infty$, $\sigma > 0$, $x^{(r)} \le x^{(r-1)} \le ... \le x(^1)$

และ $1+\xi\left(\frac{x^{(k)}-\mu}{\sigma}\right) > 0;\ k = 0, 1, 2, ..., r$

กรณีที่ $\xi = 0$ หรือ $\xi \to 0$ จะได้ฟังก์ชันความหนาแน่นความน่าจะเป็นร่วม $\tilde{\mathbf{M}}_n^{(r)}$ ดังสมการที่ (6)

$$f(x^{(1)},...,x^{(r)}) = \left[\begin{array}{l}\exp\left\{-\exp\left[-\left(\frac{x^{(r)}-\mu}{\sigma}\right)\right]\right\} \\ \times\prod_{k=1}^{r}\sigma^{-1}\exp\left[-\left(\frac{x^{(r)}-\mu}{\sigma}\right)\right]\end{array}\right] \tag{6}$$

จะเรียก $f(x^{(1)},...,x^{(r)})$ ในสมการที่ (5) และสมการที่ (6) ว่าการแจกแจงค่าสุดขีดนัยทั่วไปสำหรับสถิติอันดับ r อันดับที่ใหญ่ที่สุด (The Generalized Extreme Value Distribution for the r Largest Order Statistics; GEVr)

กรณีที่ r = 1 สมการที่ (5) และสมการที่ (6) จะลดรูปเป็นฟังก์ชันการแจกแจงความน่าจะเป็นของวงศ์ GEV กรณี $\xi \ne 0$ และ $\xi = 0$ ตามลำดับ

4. การประมาณค่าพารามิเตอร์

ในบทความนี้จะนำเสนอการประมาณค่าพารามิเตอร์ที่ได้จากสมการที่ (5) และสมการที่ (6) การด้วยวิธีประมาณค่าภาวะน่าจะเป็นสูงสุด (Maximum Likelihood Estimation; MLE) ซึ่งเป็นวิธีประมาณค่าที่มีความแกร่งมากและเหมาะสำหรับข้อมูลที่มีจำนวนมาก

ขั้นตอนการหาตัวประมาณพารามิเตอร์ ด้วยวิธี MLE มีดังต่อไปนี้

1) สร้างฟังก์ชันภาวะน่าจะเป็น (Likelihood Function) ของฟังก์ชันความหนาแน่นความน่าจะเป็นร่วมของ $\tilde{\mathbf{M}}_n^{(r)}$ จาก

$$L(\mu,\sigma,\xi)=\prod_{i=1}^{m} f(x_i^{(1)},x_i^{(2)},\ldots,x_i^{(r)})$$

เมื่อ $f(x_i^{(1)},x_i^{(2)},\ldots,x_i^{(r)})$ เป็นฟังก์ชันความหนาแน่นความน่าจะเป็นร่วมของ $\tilde{\mathbf{M}}_n^{(r)}$ ที่มี μ, σ และ ξ เป็นพารามิเตอร์ จะได้

1.1) กรณี $\xi \neq 0$ จากสมการที่ (5) จะได้ดังสมการที่ (7)

$$L(\mu,\sigma,\xi)=\prod_{i=1}^{m}\left(\begin{array}{c}\exp\left\{-\left[1+\xi\left(\dfrac{x_i^{(r)}-\mu}{\sigma}\right)\right]^{-\frac{1}{\xi}}\right\}\\ \times\prod_{k=1}^{r}\sigma^{-1}\left[1+\xi\left(\dfrac{x_i^{(r)}-\mu}{\sigma}\right)\right]^{-\frac{1}{\xi}-1}\end{array}\right) \tag{7}$$

นิยามบน $1+\xi\left(\dfrac{x^{(k)}-\mu}{\sigma}\right)>0;\ k=0,1,2,\ldots,r$ และ $i=1,2,\ldots,m$

1.2) กรณี $\xi = 0$ จากสมการที่ (6) จะได้ดังสมการที่ (8)

$$L(\mu,\sigma,\xi)=\prod_{i=1}^{m}\left(\begin{array}{c}\exp\left\{-\exp\left[-\left(\dfrac{x^{(r)}-\mu}{\sigma}\right)\right]\right\}\\ \times\prod_{k=1}^{r}\sigma^{-1}\exp\left[-\left(\dfrac{x^{(r)}-\mu}{\sigma}\right)\right]\end{array}\right) \tag{8}$$

2) สร้างฟังก์ชันล็อคภาวะน่าจะเป็น (Log-likelihood Function) ที่ได้จากขั้นตอนที่ 1 ดังนี้

$$l(\mu,\sigma,\xi)=\ln L(\mu,\sigma,\xi)$$

3) ประมาณค่าพารามิเตอร์ μ, σ และ ξ ทำได้โดยการหาอนุพันธ์ย่อย (Partial Derivative) จากขั้นตอนที่ 2 โดยกำหนดให้มีค่าเท่ากับ 0 แล้วแก้สมการหาค่าประมาณพารามิเตอร์ $(\hat{\mu},\hat{\sigma},\hat{\xi})$ ดังต่อไปนี้

กรณี $\xi \neq 0$

- หาค่าประมาณพารามิเตอร์ μ $(\hat{\mu})$ ได้จาก

$$\frac{\partial}{\partial\mu}l(\mu,\sigma,\xi)=\frac{\partial}{\partial\mu}\ln\prod_{i=1}^{m}\left(\begin{array}{c}\exp\left\{-\left[1+\xi\left(\dfrac{x_i^{(r)}-\mu}{\sigma}\right)\right]^{-\frac{1}{\xi}}\right\}\\ \times\prod_{k=1}^{r}\sigma^{-1}\left[1+\xi\left(\dfrac{x_i^{(r)}-\mu}{\sigma}\right)\right]^{-\frac{1}{\xi}-1}\end{array}\right)=0$$

- หาค่าประมาณพารามิเตอร์ σ $(\hat{\sigma})$ ได้จาก

$$\frac{\partial}{\partial\sigma}l(\mu,\sigma,\xi)=\frac{\partial}{\partial\sigma}\ln\prod_{i=1}^{m}\left(\begin{array}{c}\exp\left\{-\left[1+\xi\left(\dfrac{x_i^{(r)}-\mu}{\sigma}\right)\right]^{-\frac{1}{\xi}}\right\}\\ \times\prod_{k=1}^{r}\sigma^{-1}\left[1+\xi\left(\dfrac{x_i^{(r)}-\mu}{\sigma}\right)\right]^{-\frac{1}{\xi}-1}\end{array}\right)=0$$

- หาค่าประมาณพารามิเตอร์ ξ $(\hat{\xi})$ ได้จาก

$$\frac{\partial}{\partial\xi}l(\mu,\sigma,\xi)=\frac{\partial}{\partial\xi}\ln\prod_{i=1}^{m}\left(\begin{array}{c}\exp\left\{-\left[1+\xi\left(\dfrac{x_i^{(r)}-\mu}{\sigma}\right)\right]^{-\frac{1}{\xi}}\right\}\\ \times\prod_{k=1}^{r}\sigma^{-1}\left[1+\xi\left(\dfrac{x_i^{(r)}-\mu}{\sigma}\right)\right]^{-\frac{1}{\xi}-1}\end{array}\right)=0$$

กรณี $\xi = 0$

- หาค่าประมาณพารามิเตอร์ μ $(\hat{\mu})$ ได้จาก

$$\frac{\partial}{\partial\mu}l(\mu,\sigma,\xi)=\frac{\partial}{\partial\mu}\ln\prod_{i=1}^{m}\left(\begin{array}{c}\exp\left\{-\exp\left[-\left(\dfrac{x^{(r)}-\mu}{\sigma}\right)\right]\right\}\\ \times\prod_{k=1}^{r}\sigma^{-1}\exp\left[-\left(\dfrac{x^{(r)}-\mu}{\sigma}\right)\right]\end{array}\right)=0$$

- หาค่าประมาณพารามิเตอร์ σ ($\hat{\sigma}$) ได้จาก

$$\frac{\partial}{\partial\sigma} l(\mu,\sigma,\xi) = \frac{\partial}{\partial\sigma}\ln\prod_{i=1}^{m}\left(\begin{array}{l}\exp\left\{-\exp\left[-\left(\frac{x^{(r)}-\mu}{\sigma}\right)\right]\right\} \\ \times\prod_{k=1}^{r}\sigma^{-1}\exp\left[-\left(\frac{x^{(r)}-\mu}{\sigma}\right)\right]\end{array}\right) = 0$$

เมื่อได้ค่าประมาณมิเตอร์ $(\hat{\mu},\hat{\sigma},\hat{\xi})$ แล้วการวิเคราะห์ขั้นตอนต่อไปคือ การหาแบบจำลองที่เหมาะสมกับข้อมูลที่สุด

5. การเลือกแบบจำลอง

วิธีการที่ใช้ในการเปรียบเทียบแบบจำลองว่า แบบจำลองสถิติอันดับ r อันดับที่ใหญ่ที่สุดอันดับใดเหมาะสมกับข้อมูลที่สุด เมื่อกำหนด r_j เป็นสถิติอันดับ อันดับที่ต้องการเปรียบเทียบ และ r_i เป็นสถิติอันดับ อันดับตั้งต้น จะใช้สถิติดีเวียนซ์ (Deviance Statistic; D) ในการทดสอบ โดยมีสมมุติฐานในการทดสอบดังนี้

H_0: แบบจำลองของสถิติอันดับ อันดับที่ต้องการเปรียบเทียบที่ใหญ่ที่สุด ($\lambda(r_j)$) มีความเหมาะสม

H_1: แบบจำลองของสถิติอันดับ อันดับที่ต้องการเปรียบเทียบที่ใหญ่ที่สุด ($\lambda(r_j)$) ไม่มีความเหมาะสม

ซึ่งตัวสถิติ D นิยามดังสมการที่ (9) [16]

$$D(i,j) = 2\{\ln\lambda(r_i) - \ln\lambda(r_j)\};\chi_1^2 \tag{9}$$

สำหรับ $i \neq j$ เมื่อ $\ln\lambda(r_i)$ และ $\ln\lambda(r_j)$ เป็นค่าล็อกภาวะน่าจะเป็นสูงสุด (Maximized Log-likelihood) ของแบบจำลอง $\lambda(r_i)$ และ $\lambda(r_j)$ ตามลำดับ

โดยจะปฏิเสธสมมติฐานหลัก (H_0) ที่ระดับนัยสำคัญ (Level of Significance: α) ถ้า $D > c_\alpha$ เมื่อ c_α คือ ควอนไทล์ (Quantile) ที่ $(1-\alpha)$ ของ χ_1^2 ดังนั้นถ้าปฏิเสธแบบจำลองของสถิติอันดับ r_j อันดับที่ใหญ่ที่สุดมีความเหมาะสม แล้วจะได้ว่าแบบจำลองของสถิติอันดับ r_i อันดับที่ใหญ่ที่สุดมีความเหมาะสม ซึ่งการทดสอบด้วยตัวสถิติ D จะทำได้ครั้งละ 1 คู่ ดังนั้นทำการทดสอบทีละคู่ไปจนครบ แล้วนำผลที่ได้มาวิเคราะห์หาแบบจำลองที่เหมาะสมที่สุด

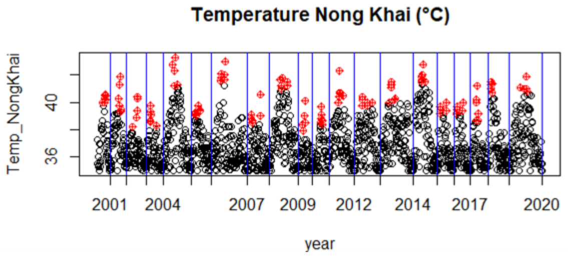
เมื่อได้แบบจำลองที่เหมาะสมที่สุดแล้วจะทำการทดสอบภาวะสารูปสนิทดี (The Goodness of Fit) ของแบบจำลองด้วยสถิติทดสอบโคโมโกรอฟสเมอรนอฟ (Kolmogorov-Smirnov Test; K-S Test) เช่นเดียวกับการใช้ GEV ในการวิเคราะห์ข้อมูล จากนั้นนำแบบจำลองที่ได้ไปประมาณระดับการเกิดซ้ำ (Return Level Estimation) ณ เวลา T ปี ($\hat{R}_T$) ทำนองเดียวกันกับการหาระดับการเกิดซ้ำของ GEV โดยสามารถหาค่าประมาณได้จากสมการที่ (10)

$$\hat{R}_T = \begin{cases}\hat{\mu} - \dfrac{\hat{\sigma}}{\hat{\xi}}\left\{1 - \left[-\log\left(1-\dfrac{1}{T}\right)\right]^{-\hat{\xi}}\right\}; & \hat{\xi} \neq 0 \\ \hat{\mu} - \hat{\sigma}\log\left\{-\log\left(1-\dfrac{1}{T}\right)\right\} & ;\ \hat{\xi} = 0\end{cases} \tag{10}$$

6. ตัวอย่างการวิเคราะห์ข้อมูล

6.1. ข้อมูล

ข้อมูลที่ใช้เป็นตัวอย่างในการวิเคราะห์เป็นข้อมูลอุณหภูมิสูงสุดรายปีของจังหวัดหนองคาย ตั้งแต่ พ.ศ. 2544–2563 เมื่อใช้ข้อมูลสูงสุด r ลำดับแรกในแต่ละปี (r = 1, 2, 3, 4, 5 และ 6) ซึ่งนำข้อมูลไปพล็อตกราฟโดยวิธีบล็อกเวลาเมื่อ r = 6 ดังรูปที่ 2



รูปที่ 2 อุณหภูมิสูงสุดรายปีของจังหวัดหนองคายตั้งแต่ พ.ศ. 2554–2563 โดยวิธีบล็อกเวลารายปี เมื่อ r = 6 [17]

ตารางที่ 1 ค่าประมาณพารามิเตอร์ และค่าล็อกภาวะน่าจะเป็นสูงสุด ($\ln \lambda(r)$) ของแต่ละแบบจำลอง

แบบจำลองที่	r	$\ln \lambda(r)$	ค่าประมาณพารามิเตอร์			การแจกแจงที่เหมาะสม
			$\hat{\mu}$ (CI 95%)	$\hat{\sigma}$ (CI 95%)	$\hat{\xi}$ (CI 95%)	
1	1	–29.71	40.564 (40.132, 40.983)	0.863 (0.546, 1.181)	0.082 (–0.241, 0.405)	กัมเบล
2	2	–42.08	41.038 (40.616, 41.458)	1.048 (0.643, 1.084)	–0.300 (–0.476, –0.124)	ไวบูล
3	3	–43.07	41.265 (40.895, 41.635)	0.962 (0.805, 1.118)	–0.333 (–0.478, –0.189)	ไวบูล
4	4	–40.43	41.337 (40.994, 41.681)	0.918 (0.788, 1.048)	–0.330 (–0.460, –0201)	ไวบูล
5	5	–35.58	41.460 (41.137, 41.783)	0.880 (0.774, 0.986)	–0.376 (–0.487, –0.266)	ไวบูล
6	6	–28.45	41.552 (41.246, 41.859)	0.849 (0.788, 1.048)	–0.405 (–0.460, –0.201)	ไวบูล

6.2. ค่าประมาณพารามิเตอร์และรูปแบบของพารามิเตอร์ที่เหมาะสม

นำข้อมูลสูงสุดรายปีเมื่อใช้ r = 1, 2, 3, 4, 5 และ 6 ตามลำดับ ที่กล่าวมาข้างต้นมาวิเคราะห์ จากผลวิเคราะห์จะได้ ค่าประมาณพารามิเตอร์ที่เหมาะสม และค่าล็อกภาวะน่าจะเป็นสูงสุด ($\ln \lambda(r)$) ของแต่ละรูปแบบแสดงในตารางที่ 1

จากตารางที่ 1 เมื่อนำค่า $\ln \lambda(r)$ มาเรียงจากมากไปหาน้อยจะได้ แบบจำลองที่ 6 (–28.45) แบบจำลองที่ 1 (–29.71) แบบจำลองที่ 5 (–35.58) แบบจำลองที่ 4 (–40.43) แบบจำลองที่ 2 (–42.08) และแบบจำลองที่ 3 (–43.07) ตามลำดับ

ทำการเปรียบเทียบแบบจำลองรายคู่ (1, 2), (2, 3), (3, 4), (4, 5) และ (5, 6) ตามลำดับ ด้วยสถิติดีเวียนซ์ (D) ที่ระดับนัยสำคัญ (α) 0.01 จะได้ χ^1_2 = 6.64 ซึ่งผลการวิเคราะห์ค่าสถิติดีเวียนซ์และค่าพี (*p*-value) ของการเปรียบเทียบรายคู่แสดงในตารางที่ 2

ตารางที่ 2 ค่าสถิติดีเวียนซ์ (D) และค่าพี (*p*-value) ของการเปรียบเทียบรายคู่

	$D(1, 2)$	$D(2, 3)$	$D(3, 4)$	$D(4, 5)$	$D(5, 6)$
D *p*-value	–24.73 > 0.05	–1.99 > 0.05	5.28 > 0.05	9.71 < 0.05	14.26 < 0.05

จากตารางที่ 2 พบว่า ในการเปรียบเทียบแบบจำลองที่ 1 กับแบบจำลองที่ 2 จะได้ $D(1, 2) = -24.73 < \chi_1^2$ (6.64) ดังนั้นยอมรับแบบจำลองที่ 2 นั่นคือ แบบจำลองที่ 2 มีความเหมาะสม การเปรียบเทียบแบบจำลองที่ 2 กับแบบจำลองที่ 3 จะได้ $D(2, 3) = -1.99 < \chi_1^2$ ดังนั้นยอมรับแบบจำลองที่ 3 นั่นคือ รูปแบบที่ 3 มีความเหมาะสม การเปรียบเทียบแบบจำลองที่ 3 กับแบบจำลองที่ 4 จะได้ $D(3, 4) = 5.28 < \chi_1^2$ ดังนั้นยอมรับแบบจำลองที่ 4 นั่นคือ รูปแบบที่ 4 มีความเหมาะสม การเปรียบเทียบแบบจำลองที่ 4 กับแบบจำลองที่ 5 จะได้ D(4, 5) = 9.71 > χ_1^2 ดังนั้น ปฏิเสธแบบจำลองที่ 5 นั่นคือ แบบจำลองที่ 4 มีความเหมาะสม และการเปรียบเทียบแบบจำลองที่ 5 กับแบบจำลองที่ 6 จะได้ D(5, 6) = 14.26 > χ_1^2 ดังนั้น ปฏิเสธแบบจำลองที่ 6 นั่นคือ รูปแบบที่ 5 มีความเหมาะสม ซึ่งแบบจำลองที่ 4 เหมาะสมมากกว่าแบบจำลองที่ 5 จึงสามารถสรุปได้ว่าแบบจำลองของสถิติอันดับ 4 อันดับที่ใหญ่ที่สุด เป็นแบบจำลองเหมาะสมที่สุด ดังนั้นสามารถเขียนแบบจำลองที่เหมาะสำหรับข้อมูลอุณหภูมิสูงสุดรายปีของจังหวัดหนองคายได้ดังนี้ $X \sim \mathrm{GEVr}\,(\mu, \sigma, \xi)$

เมื่อ μ = 41.34, σ = 0.92 และ ξ = –0.33 ซึ่งมีการแจกแจงไวบูล

นำแบบจำลองที่ได้มาประมาณระดับการเกิดซ้ำ 2, 5,

10 และ 20 ปี ของอุณหภูมิสูงสุดรายปีของจังหวัดหนองคาย ซึ่งผลการประมาณที่ได้แสดงในตารางที่ 3

ตารางที่ 3 ค่าประมาณของอุณหภูมิสูงสุด (องศาเซลเซียส) ที่เกิดซ้ำในรอบปี (T)

T	2	5	10	20
ค่าประมาณ	41.65 (41.33, 41.98)	42.42. (40.84, 44.01)	42.79 (37.13, 48.46)	43.07 (27.59, 58.55)

จากตารางที่ 3 พบว่า ค่าประมาณระดับการเกิดซ้ำของอุณหภูมิสูงสุดรายปีของจังหวัดหนองคายในรอบปีการเกิดซ้ำ 2 ปี จะมีอุณหภูมิสูงสุดรายปีประมาณ 41.65 องศาเซลเซียส ด้วยความน่าจะเป็น 0.5 สำหรับรอบปีการเกิดซ้ำ 5 ปี จะมีอุณหภูมิสูงสุดรายปีประมาณ 42.42 องศาเซลเซียส ด้วยความน่าจะเป็น 0.2 สำหรับรอบปีการเกิดซ้ำ 10 ปี จะมีอุณหภูมิสูงสุดรายปีประมาณ 42.79 องศาเซลเซียส ด้วยความน่าจะเป็น 0.1 และรอบปีการเกิดซ้ำ 20 ปี จะอุณหภูมิสูงสุดรายปีประมาณ 43.07 องศาเซลเซียส ด้วยความน่าจะเป็น 0.05

7. สรุป

บทความนี้ผู้เขียนจึงได้กล่าวถึงวิธีการวิเคราะห์ข้อมูลที่มีจำนวนข้อมูลไม่เพียงพอ โดยใช้แนวคิดของการแจกแจงสำหรับสถิติอันดับ r อันดับที่ใหญ่ที่สุด (the r Largest Order Statistic Model) สำหรับวิธีวิธีบล็อกเวลา (Block Time Method) ผ่านการแจกแจงค่าสุดขีดนัยทั่วไป (GEV) และเลือกแบบจำลองที่ดีที่สุดโดยใช้สถิติดีเวียนซ์ (Deviance Statistic) ใช้ในการทดสอบ และรวมไปถึงการประมาณระดับการเกิดซ้ำ (Return Level Estimation) ณ เวลาต่างๆ ของแบบจำลองที่ได้ พร้อมทั้งยกตัวอย่างการวิเคราะห์ข้อมูลโดยใช้ข้อมูลอุณหภูมิสูงสุดรายปีของจังหวัดหนองคายตั้งแต่ พ.ศ. 2554–2563 ซึ่งผลการวิเคราะห์พบว่า แบบจำลองที่เหมาะสมสำหรับข้อมูลอุณหภูมิสูงสุดรายปีของจังหวัดหนองคายคือแบบจำลองของสถิติอันดับ 4 อันดับที่ใหญ่ที่สุด มีความเหมาะสมแบบจำลองภายใต้กระบวนการคงที่ ดังนี้

$$X \sim \mathrm{GEVr}\,(\mu, \sigma, \xi)$$

เมื่อ $\mu = 41.34$, $\sigma = 0.92$ และ $\xi = -0.33$ ซึ่งมีการแจกแจงไวบูล และมีระดับการเกิดซ้ำเพิ่มขึ้นเมื่อรอบปีการเกิดซ้ำเพิ่มขึ้น

เอกสารอ้างอิง

[1] P. Busababodhin and A. Keawmun, “Extreme values statistics,” *The Journal of KMUTNB*, vol. 25, no. 2, pp. 55–65, 2015 (in Thai)

[2] S. Kotz and S. Nadaraja, *Extreme Value Distributions: Theory and Applications*, Singapore: Imperial College Press, 2000.

[3] J. Beirlant, Y. Goegebeur, J. Segers, J. Teugels, D. D. Waal, and C. Ferro, *Statistics of Extremes: Theory and Applications*, New York: John Wiley & Sons, 2004.

[4] S. Coles, *An Introduction to Statistical Modeling of Extreme Values*, London: Springer-Varlag, 2001.

[5] S. Coles and S. Nadaraja, *An Introduction to Statistical Modeling of Extreme Values*, Great Britain: Springer-Varlag London Limited, 2001.

[6] P. Embrecht, C. Kluppelberg, and T. Mikosch, *Modeling extremal events for insurance and finance*, Berlin: Springer Verlag, 1997.

[7] E. J. Gumbel, *Statistics of Extremes*, New York: Columbia University Press, 1958.

[8] J. Galambos, *The asymptotic theory of extreme order statistics*, New York: Wiley, 1978.

[9] P. Guayjarernpanishk, T Phupiewpha, and P. Busababodhin, “Extreme value analysis: Non–stationary process,” *The Journal of KMUTNB*, vol. 32, no. 2, 2021 (in Thai).

[10] B. Finkenstadt and H. Rootzen, *Extreme values in finance, telecommunications, and the*

environment, London: Chapman and Hall/CRC Press, 2004.

[11] P. Amphanthong and P. Busababodhin, “Modeling and prediction of exchange rate and billion gold price of Thailand,” *International Journal of Statistics and Economics*, vol. 16, no. 3, pp. 81–92, 2015.

[12] P. Busababodhin, “Modeling on maximum rainfall and temperature based on extreme value copula analysis,” in *Proceeding of 10th Conference on Extreme Value Analysis (EVA2017)*, 2017, pp. 14.

[13] R. D. Reiss and M. Thomas, *Statistical Analysis of Extremes value with Applications to Insurance, Finance, Hydrology and Other Fields*, Germany: Springer, 2007.

[14] T. An and M. D. Pandey, “A comparison of methods of extreme wind speed estimation,” *Technical note Journal of Wind Engineering and Industrial Aerodynamics*, vol. 93, no. 7, pp. 535–545, 2005.

[15] A. F. Jenkinson, “The frequency distribution of the annual maximum (or minimum) values of meteorological elements,” *Quarterly Journal of the Royal Meteorological Society*, vol. 81, no. 348, pp. 158–171, 1955.

[16] M. M. Nemukula and C. Sigauke, “Modelling average maximum daily temperature using *r* largest order statistics: An application to South African data,” *Jàmbá: Journal of Disaster Risk Studies*, vol. 10, no. 1, pp. 1–11, 2018.

[17] Meteorological Department of Thailand. (2021, May). Thailand and Around Area Earthquake Report. Bangkok, Thailand [Online] (in Thai). Available: https://www.tmd.go.th