บทความวิชาการ

การวิเคราะห์ค่าสุดขีด: ภายใต้กระบวนการไม่คงที่

พรรณรัตน์ ก้วยเจริญพานิชก์
คณะวิทยาศาสตร์ประยุกต์และวิศวกรรมศาสตร์ มหาวิทยาลัยขอนแก่น วิทยาเขตหนองคาย
ทศพล ภูผิวฟ้า และ ปิยภัทร บุษบาบดินทร์*
ภาควิชาคณิตศาสตร์ คณะวิทยาศาสตร์ มหาวิทยาลัยมหาสารคาม

* ผู้นิพนธ์ประสานงาน โทรศัพท์ 08 9542 6396 อีเมล: piyapatr.b@msu.ac.th
DOI: 10.14416/j.kmutnb.2021.05.041
รับเมื่อ 17 มิถุนายน 2563 แก้ไขเมื่อ 4 สิงหาคม 2563 ตอบรับเมื่อ 10 สิงหาคม 2563 เผยแพร่ออนไลน์ 27 พฤษภาคม 2564

บทคัดย่อ

ค่าสุดขีด (Extreme Value) หมายถึง เซตของข้อมูลที่เป็นค่าสูงสุดหรือค่าต่ำสุด ที่อยู่ในเหตุการณ์สุดขีด (Extreme Event) ที่เกิดขึ้นในธรรมชาติ ดังนั้นการหาโอกาสที่จะเกิดเหตุการณ์สุดขีดในอดีตว่าจะเกิดขึ้นได้อีกในอนาคตหรือไม่นั้น คือ การที่นักวิเคราะห์พยายามสร้างแบบจำลองที่ดีที่สุดสำหรับค่าสุดขีดที่ศึกษา ซึ่งนักวิเคราะห์ส่วนใหญ่มักจะตัดข้อมูลดังกล่าวทิ้งไปไม่นำมาพิจารณาในการสร้างแบบจำลอง เนื่องจากการวิเคราะห์ข้อมูลเหล่านี้มีความซับซ้อนและยุ่งยาก แต่ในความเป็นจริงถ้านักวิเคราะห์ต้องการทราบความน่าจะเป็นหรือโอกาสของเหตุการณ์ที่มีค่าสูงสุดหรือต่ำสุดซึ่งอยู่ในส่วนปลายหางที่มีจำนวนข้อมูลน้อยมาก ซึ่งการวิเคราะห์ข้อมูลที่มีคุณสมบัติเป็นค่าสุดขีด เงื่อนไขหนึ่งที่จำเป็นต้องตรวจสอบก่อนจะนำข้อมูลไปวิเคราะห์ต่อเพื่อหาค่าพารามิเตอร์ของแบบจำลองคือ ข้อมูลที่นำมาวิเคราะห์อยู่ภายใต้กระบวนการแบบใด ระหว่างกระบวนการคงที่ (Stationary Process) หรือกระบวนการไม่คงที่ (Non-stationary Process) เนื่องจากกระบวนการทั้งสองมีขั้นตอนการวิเคราะห์ และวิธีการเลือกตัวแบบที่เหมาะสมที่แตกต่างกัน ดังนั้นถ้าหากไม่มีขั้นตอนการพิจารณาลักษณะของข้อมูลอาจจะทำให้ผลประมาณค่าพารามิเตอร์ของแบบจำลองผิดพลาด และส่งผลถึงการนำไปใช้ต่อที่ไม่เกิดประโยชน์ และอาจจะส่งผลร้ายแรง โดยเฉพาะการวิเคราะห์ข้อมูลในด้านที่ต้องใช้ความแม่นยำของแบบจำลองเป็นอย่างยิ่ง

คำสำคัญ: ค่าสุดขีด การแจกแจงค่าสุดขีดนัยทั่วไป กระบวนการคงที่ กระบวนการไม่คงที่

การอ้างอิงบทความ: พรรณรัตน์ ก้วยเจริญพานิชก์, ทศพล ภูผิวฟ้า และ ปิยภัทร บุษบาบดินทร์, “การวิเคราะห์ค่าสุดขีด: ภายใต้กระบวนการไม่คงที่,” *วารสารวิชาการพระจอมเกล้าพระนครเหนือ*, ปีที่ 32, ฉบับที่ 2, หน้า 506–515, เม.ย.–มิ.ย. 2565.

Academic Article

Extreme Value Analysis: Non–stationary Process

Pannarat Guayjarernpanishk
Faculty of Applied Science and Engineering, Khon Kaen University, Nong Khai Campus, Nong Khai, Thailand
Tossapol Phupiewpha 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.2021.05.041
Received 17 June 2020; Revised 4 August 2020; Accepted 10 August 2020; Published online: 27 May 2021

Abstract

Extreme value means a set of data that is the highest or lowest value in an extreme event that naturally occurs. Therefore, it is intended to find the opportunity to experience the extreme events in the past that will happen in the future. This includes the analysts to create the best model for the extreme values study. Most analysts tend to exclude the data and do not consider it in creating the model because the data is complicated and complex. However, in reality, they want to know the probability or opportunity of the event with the highest or lowest value, which is at the tail end with very little amount of data. In data analysis of the extreme features, it is necessary to check for the model parameters and consider the type of data to be analyzed whether the process is stationary or unstable process (non-stationary process). Since both processes have different analysis procedures and methods for selecting the suitable model, then, if there is no data considering process, the results might be incorrect causing an error in estimated parameter values of the model. Consequently, this might lead to useless utilization and serious outcomes particularly the data analysis process that requires high precision of the model.

Keywords: Extreme Value, Generalized Extreme Value Distribution, Stationary Process, Non-stationary Process

Please cite this article as: P. Guayjarernpanishk, T. Phupiewpha, and P. Busababodhin, "Extreme value analysis: Non–stationary process," *The Journal of KMUTNB*, vol. 32, no. 2, pp. 506–515, Apr.–Jun. 2022 (in Thai).

1. บทนำ

ปัจจุบันในประเทศไทยได้มีการนำเอาข้อมูลที่มีลักษณะสุดขีดหรือมีค่าสุดขีด (Extreme Value) มาวิเคราะห์หาแบบจำลองกันอย่างแพร่หลาย ไม่ว่าจะเป็นทางด้านวิศวกรรมศาสตร์ เศรษฐศาสตร์ อุทกวิทยา การเงิน และการประกันภัย เป็นต้น ซึ่งจุดเริ่มต้นของการค้นพบการแจกแจงค่าสุดขีดคือ ค.ศ. 1709 โดย Bernulli และถูกนำมาประยุกต์ใช้ใน ค.ศ. 1914 โดย Fuller หลังจากนั้นได้มีการพัฒนามาเรื่อยๆ จากนักคณิตศาสตร์ เช่น ใน ค.ศ. 2000 โดย Kotz, และ Nadaraja [1] โดยสามารถศึกษาเพิ่มเติมได้จาก [2]–[6] และ [14] ซึ่งการวิเคราะห์ข้อมูลที่มีลักษณะค่าสุดขีดนี้มีประโยชน์อย่างมากในการเตรียมรับมือกับเหตุการณ์ที่ก่อให้เกิดความเสียหายขนาดใหญ่ เพื่อเป็นการลดความสูญเสียที่จะเกิดขึ้นหากเกิดเหตุการณ์เช่นนั้นขึ้นอีกในอนาคต ดังนั้นการวิเคราะห์ข้อมูลสุดขีดที่ถูกต้องจึงเป็นสิ่งสำคัญ เพื่อให้ได้ผลลัพธ์จากการวิเคราะห์มีความแม่นยำมากที่สุด และสามารถนำไปใช้เป็นข้อมูลเบื้องต้นในการจัดการในขั้นตอนต่อไป

สำหรับการวิเคราะห์ข้อมูลที่มีค่าสุดขีด เงื่อนไขหนึ่งที่จำเป็นต้องตรวจสอบก่อนจะนำข้อมูลไปวิเคราะห์ต่อเพื่อหาค่าพารามิเตอร์ของแบบจำลองคือ ข้อมูลที่นำมาวิเคราะห์อยู่ภายใต้กระบวนการแบบใด ระหว่างกระบวนการคงที่ (Stationary Process) หรือกระบวนการไม่คงที่ (Non-stationary Process) ซึ่งกระบวนการทั้งสองมีขั้นตอนการวิเคราะห์ และวิธีการเลือกตัวแบบที่เหมาะสมที่แตกต่างกัน ซึ่งโดยทั่วไปข้อมูลทางการประกันภัย อุทกวิทยา เศรษฐศาสตร์เกิดขึ้นภายใต้กระบวนการไม่คงที่ โดยมีตัวแปรอื่นๆ เช่น ตัวแปรด้านเวลา หรือฤดูกาลเข้ามาเกี่ยวข้อง ซึ่งหากไม่มีขั้นตอนการพิจารณาลักษณะของข้อมูล อาจจะทำให้ผลประมาณค่าพารามิเตอร์ของแบบจำลองผิดพลาด และส่งผลถึงการนำไปใช้ต่อที่ไม่เกิดประโยชน์ และอาจจะส่งผลร้ายแรงโดยเฉพาะการวิเคราะห์ข้อมูลในด้านที่ต้องใช้ความแม่นยำของแบบจำลองเป็นอย่างยิ่ง ตัวอย่างเช่น บริษัทประกันภัยต้องเตรียมเงินสำรองเพื่อจ่ายค่าสินไหมเท่าไรถึงจะเพียงพอหากเกิดความเสียหายร้ายแรงกับภัยที่รับทำประกันไว้ หรือการสร้างกำแพงกั้นน้ำเพื่อป้องกันน้ำท่วม หรือการศึกษาเพิ่มเติมจาก [7]–[11] จำเป็นต้องได้ข้อมูลที่แม่นยำในการวางแผนสร้างกำแพงว่าต้องสร้างสูงกี่เมตรจึงจะเพียงพอต่อการป้องกันน้ำท่วมได้ เป็นต้น ดังนั้นขั้นตอนนี้จึงเป็นสิ่งที่ต้องพิจารณาเป็นขั้นตอนแรกก่อนทำการวิเคราะห์ในขั้นตอนต่อไป ซึ่งในบทความนี้จะกล่าวถึงค่าสุดขีดภายใต้กระบวนการไม่คงที่ของการแจกแจงค่าสุดขีดนัยทั่วไป

โดยหัวข้อถัดจะกล่าวถึง การแจกแจงค่าสุดขีดนัยทั่วไป ค่าสุดขีดภายใต้กระบวนการไม่คงที่ การประมาณค่าพารามิเตอร์ การเลือกแบบจำลอง และตัวอย่างการวิเคราะห์ตามลำดับ

2. การแจกแจงค่าสุดขีดนัยทั่วไป (Generalized Extreme Value Distribution; GEV)

การแจกแจงค่าสุดขีดนัยทั่วไป (GEV) ถูกพัฒนาขึ้นครั้งแรกโดย Jenkinson ใน ค.ศ. 1955 [12] เป็นวิธีการวิเคราะห์ข้อมูลที่มีค่าสุดขีดวิธีหนึ่ง ซึ่งเหมาะสำหรับการวิเคราะห์ค่าสุดขีดในช่วงเวลาที่มีคาบเท่าๆ กัน เช่น รายสัปดาห์ รายเดือน รายสามเดือน หรือรายปี ซึ่งวิธีการเลือกข้อมูลที่นำมาใช้ในการวิเคราะห์คือ เลือกข้อมูลที่มีค่าสูงสุดในแต่ละช่วงคาบเวลา โดยมีแนวคิดดังนี้

สมมติให้เป็น X_i เมื่อ $i = 1, 2, ..., n$ เป็นตัวแปรที่เป็นอิสระกันและมีฟังก์ชันการแจกแจง $F(x)$ แบบเดียวกัน และกำหนดให้ $X_{(n)} = \max(X_1, X_2, \ldots, X_n)$ โดยในการวิเคราะห์โดยใช้ GEV จะมีพารามิเตอร์ที่เกี่ยวข้องอยู่ 3 ตัว คือ μ พารามิเตอร์ระบุตำแหน่ง (Location Parameter) σ พารามิเตอร์ระบุขนาด (Scale Parameter) และ ξ พารามิเตอร์ระบุรูปร่าง (Shape Parameter)

ใน ค.ศ. 1978 Galambos [13] ได้สร้างฟังก์ชันการแจกแจงสะสม (Cumulative Distribution Function; CDF) ของ GEV สำหรับ $-\infty < x < \infty$ ดังนี้

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

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

ถ้า $\xi = 0$ หรือ $\xi \to 0$ จะได้

$$G(x;\mu,\sigma,\xi) = \exp\left\{-\exp\left[-\left(\frac{x-\mu}{\sigma}\right)\right]\right\} \qquad (2)$$

จากสมการที่ (1) ถ้า $\xi < 0$ จะเรียกการแจกแจงค่าสุดขีดนัยทั่วไปว่า "การแจกแจงไวบูล (Weibull Distribution)" และถ้า $\xi > 0$ จะเรียกว่า "การแจกแจงฟรีเชท (Fréchet Distribution)" และจากสมการที่ (2) เมื่อ $\xi \to 0$ จะเรียกว่า "การแจกแจงกัมเบล (Gumbel Distribution)" โดย GEV มีฟังก์ชันการแจกแจงความน่าจะเป็น (Probability Distribution Function; pdf) ดังนี้

$$g(x;\mu,\sigma,\xi) = \begin{cases} \frac{1}{\sigma}\left[1+\xi\left(\frac{x-\mu}{\sigma}\right)\right]^{-1/\xi-1} \exp\left\{-\left[1+\xi\left(\frac{x-\mu}{\sigma}\right)\right]^{-1/\xi}\right\} & ;\xi \neq 0 \\ \frac{1}{\sigma}\exp\left(\frac{x-\mu}{\sigma}\right)^{-1} \exp\left\{-\exp\left[-\left(\frac{x-\mu}{\sigma}\right)\right]\right\} & ;\xi = 0 \end{cases} \qquad (3)$$

3. ค่าสุดขีดภายใต้กระบวนการไม่คงที่ (Non–stationary Process in Extreme Value)

ส่วนใหญ่ข้อมูลทางด้านการประกันภัย อุตุนิยมวิทยา หรือเศรษฐศาสตร์ ที่นำมาวิเคราะห์ค่าสุดขีดได้รับอิทธิภพจากตัวแปรอื่นๆ เช่น ข้อมูลมีการเปลี่ยนแปลงเมื่อเวลาเปลี่ยนไป หรือข้อมูลมีแนวโน้ม ซึ่งทำให้ข้อมูลที่นำมาวิเคราะห์เป็นข้อมูลที่อยู่ภายใต้กระบวนการไม่คงที่ การวิเคราะห์ข้อมูลลักษณะนี้จะแตกต่างจากการวิเคราะห์ข้อมูลที่อยู่ภายใต้กระบวนการคงที่ โดยที่ลักษณะข้อมูลที่อยู่ภายใต้กระบวนการไม่คงที่แสดงดังรูปที่ 1

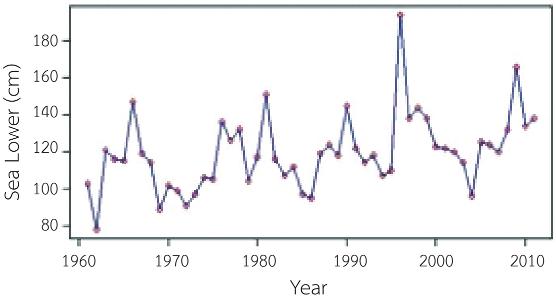
รูปแบบของแบบจำลองการแจกแจง GEV ภายใต้กระบวนไม่คงที่ ที่ใช้อธิบายการแจกแจงของ X_t ณ เวลา t เมื่อ t = 1, 2, .., m เป็นดังนี้

$X_t \sim \mathrm{GEV}\left(\mu(t), \sigma(t), \xi(t)\right)$

ตัวอย่างพารามิเตอร์ภายใต้กระบวนการไม่คงที่

$\mu(t) = \beta_0$

$\mu(t) = \beta_0 + \beta_1 t$



รูปที่ 1 ระดับน้ำทะเลสูงสุด (เซนติเมตร) เมืองเวนิส ตั้งแต่ ค.ศ. 1960–2011

$\mu(t) = \beta_0 + \beta_1 t + \beta_2 t^2$

$\mu(t) = \beta_0 + \beta_1 \exp(-\beta_1 t)$

$\sigma(t) = \beta_0$

$\sigma(t) = \exp(\beta_0 + \beta_1 t)$

$\sigma(t) = \exp(\beta_0 + \beta_1 t + \beta_2 t^2)$

$\xi(t) = \beta_0$

$\xi(t) = \beta_0 + \beta_1 t$

$\xi(t) = \exp(\beta_0 + \beta_1 t)$

ตัวอย่างรูปแบบของพารามิเตอร์ที่สามารถเป็นไปได้

รูปแบบ 1 : μ, σ และ ξ เป็นค่าคงที่

รูปแบบ 2 : $\mu(t) = \beta_0 + \beta_1 t$
โดยที่ σ และ ξ เป็นค่าคงที่

รูปแบบ 3 : $\mu(t) = \beta_0 + \beta_1 t + \beta_2 t^2$
โดยที่ σ และ ξ เป็นค่าคงที่

รูปแบบ 4 : $\mu(t) = \beta_0 + \beta_1 \exp(-\beta_2 t)$
โดยที่ σ และ ξ เป็นค่าคงที่

รูปแบบ 5 : $\sigma(t) = \exp(\beta_0 + \beta_1 t)$
โดยที่ μ, ξ เป็นค่าคงที่

รูปแบบ 6 : $\mu(t) = \beta_0 + \beta_1 t$
และ $\sigma(t) = \exp(\beta_0 + \beta_1 t)$
โดยที่ ξ เป็นค่าคงที่

รูปแบบ 7 : $\mu(t) = \beta_0 + \beta_1 t + \beta_2 t^2$
และ $\sigma(t) = \exp(\beta_0 + \beta_1 t)$
โดยที่ ξ เป็นค่าคงที่

รูปแบบ 8 : $\mu(t) = \beta_0 + \beta_1 \exp(-\beta_2 t)$

และ $\sigma(t) = \exp(\beta_0 + \beta_1 t)$

โดยที่ ξ เป็นค่าคงที่

หากผลการวิเคราะห์ค่าพารามิเตอร์เป็นรูปแบบที่ 1 สามารถกล่าวได้ว่าข้อมูลที่นำมาวิเคราะห์เป็นข้อมูลที่มีความคงที่ (Stationary) แต่ถ้าผลเป็นรูปแบบอื่นๆ จะได้ว่าข้อมูลที่นำมาวิเคราะห์เป็นข้อมูลที่มีความไม่คงที่ (Non-stationary)

4. การประมาณค่าพารามิเตอร์

วิธีในการประมาณค่าพารามิเตอร์มีอยู่ด้วยกันหลายวิธี แต่ในบทความนี้จะนำเสนอการประมาณค่าพารามิเตอร์ด้วยวิธีประมาณค่าภาวะน่าจะเป็นสูงสุด (Maximum Likelihood Estimation; MLE) ซึ่งเป็นวิธีประมาณค่าที่มีความแกร่งมากที่สุด

ขั้นตอนการหาตัวประมาณพารามิเตอร์ของการแจกแจง GEV ด้วยวิธี MLE มีดังต่อไปนี้

1. สร้างฟังก์ชันภาวะน่าจะเป็น (Likelihood Function) ของฟังก์ชันการแจกแจงความน่าจะเป็นของ GEV จะได้

$$L(\beta) = \prod_{t=1}^{m} g(x_t; \mu(t), \sigma(t), \xi(t))$$

เมื่อ β เป็นเวกเตอร์ของพารามิเตอร์ β_i และ $g\left(x_t; \mu(t), \sigma(t), \xi(t)\right)$ เป็นฟังก์ชันการแจกแจงความน่าจะเป็นของ GEV ที่มี $\mu(t)$, $\sigma(t)$ และ $\xi(t)$ เป็นพารามิเตอร์ที่ x_t

2. สร้างฟังก์ชันล็อกภาวะน่าจะเป็นของฟังก์ชันการแจกแจงความน่าจะเป็นของ GEV ที่ได้จากขั้นตอนที่ 1 และสมการที่ (3) สำหรับ t = 1, 2, .., m ดังนี้

2.1 กรณี $\xi \neq 0$ ดังสมการที่ (4)

$$l(\beta) = -\sum_{t=1}^{m} \left\{ \begin{array}{l} \log\sigma(t) + \left(1 + \frac{1}{\xi(t)}\right) \log\left[1 + \xi(t)\left(\frac{x_t - \mu(t)}{\sigma(t)}\right)\right] \\ + \left[1 + \xi(t)\left(\frac{x_t - \mu(t)}{\sigma(t)}\right)\right]^{-1/\xi(t)} \end{array} \right\} \quad (4)$$

นิยามบน $1 + \xi(t)\left(\frac{x_t - \mu(t)}{\sigma(t)}\right) > 0$

2.2 กรณี $\xi = 0$ ดังสมการที่ (5)

$$l(\beta) = -\sum_{t=1}^{m} \left\{ \log\sigma(t) + \left(\frac{x_t - \mu(t)}{\sigma(t)}\right) + \exp\left[-\left(\frac{x - \mu}{\sigma}\right)\right] \right\} \quad (5)$$

ตัวอย่างการแจกแจง GEV ภายใต้กระบวนการไม่คงที่ เมื่อ $\mu(t) = \beta_0 + \beta_1 \exp(-\beta_2 t)$, $\sigma(t)$ และ $\xi(t)$ เป็น ค่าคงที่ จะได้

$$l\left(\mu(t), \sigma, \xi\right) = -\sum_{t=1}^{m} \left\{ \begin{array}{l} \log\sigma + \left(1 + \frac{1}{\xi}\right) \log\left[1 + \xi\left(\frac{x_t - (\beta_0 + \beta_1 \mathrm{e}^{-\beta_2 t})}{\sigma}\right)\right] \\ + \left[1 + \xi\left(\frac{x_t - (\beta_0 + \beta_1 \exp(-\beta_2 t))}{\sigma}\right)\right]^{-1/\xi} \end{array} \right\}$$

3. ประมาณค่าพารามิเตอร์ $\mu(t)$, $\sigma(t)$ และ $\xi(t)$ $(\hat{\mu}(t), \hat{\sigma}(t), \hat{\xi}(t))$ ด้วยการหาอนุพันธ์ย่อย (Partial Derivative) จากฟังก์ชันที่ได้ในขั้นตอนที่ 2 ตามลำดับ ดังต่อไปนี้

$$\frac{\partial l}{\partial \mu(t)}\left(\mu(t), \sigma(t), \xi(t)\right) = 0,$$

$$\frac{\partial l}{\partial \sigma(t)}\left(\mu(t), \sigma(t), \xi(t)\right) = 0$$

และ $$\frac{\partial l}{\partial \xi(t)}\left(\mu(t), \sigma(t), \xi(t)\right) = 0$$

5. การเลือกแบบจำลอง

วัตถุประสงค์ของการเลือกแบบจำลอง คือ ได้แบบจำลองที่เหมาะสมที่สุดสำหรับข้อมูลที่นำมาวิเคราะห์ ในกรณีแบบจำลองที่ได้จากการวิเคราะห์มีมากกว่า 1 แบบจำลองนั้น ขั้นตอนการตรวจสอบความเหมาะสมของแบบจำลองมีความสำคัญอย่างมาก

ในกรณีที่ข้อมูลอยู่ภายใต้กระบวนการคงที่กระบวนการเดียว การเปรียบเทียบแบบจำลองต่างๆ เพื่อหาแบบจำลองที่เหมาะสม วิธีการที่ใช้กันอย่างแพร่หลาย คือ พิจารณา

กราฟ (Diagnostic Plots) ซึ่งประกอบด้วย กราฟควอนไทล์ (Quantile Plot) กราฟความน่าจะเป็น (Probability Plot) กราฟความหนาแน่น (Density Plot) และกราฟระดับการเกิดซ้ำ (Return Level Plot) หรือพิจารณาจากเกณฑ์ข้อสนเทศของอะกะอิเกะ (Akaike's Information Criterion; AIC) หรือเกณฑ์ข้อสนเทศของเบส์ (Bayesian Information Criterion; BIC)

แต่สำหรับการเลือกแบบจำลองที่เหมาะสมกับข้อมูลที่มีแบบจำลองทั้งแบบจำลองภายใต้กระบวนการคงที่และแบบจำลองภายใต้กระบวนการไม่คงที่ แบบจำลองแบบใดเหมาะสมมากกว่ากัน ซึ่งโดยทั่วไปแล้วจากวิธีการประมาณค่าโดยวิธีประมาณค่าภาวะน่าจะเป็นสูงสุดที่กล่าวมาแล้วในหัวข้อที่ผ่านจะได้แบบจำลองมากกว่า 1 แบบจำลองและแบบจำลองที่ได้มีความเกี่ยวเนื่องกัน ซึ่งแบบจำลองที่มีลักษณะนี้ว่า "แบบจำลองซ้อนใน (Nested Model)" สถิติที่ใช้ในการทดสอบในลักษณะนี้คือ สถิติดีเวียนซ์ (Deviance Statistic) ซึ่งจะแทนด้วย D

กำหนดให้ M_0 และ M_1 เป็นแบบจำลองตั้งต้นและแบบจำลองที่ต้องการเปรียบเทียบตามลำดับ และมีเงื่อนไขว่า $M_0 \subset M_1$ โดยมีสมมุติฐานในการทดสอบดังนี้

H_0: แบบจำลองตั้งต้นมีความเหมาะสม

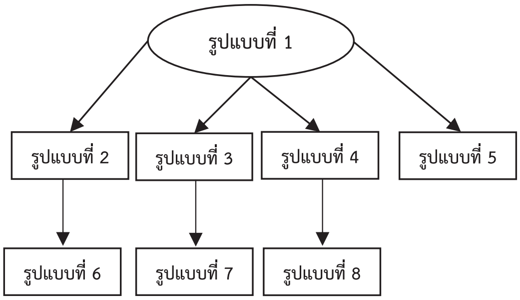
H_1: แบบจำลองที่ต้องการเปรียบเทียบมีความเหมาะสม

โดยใช้ตัวสถิติ D ซึ่ง D นิยามดังนี้

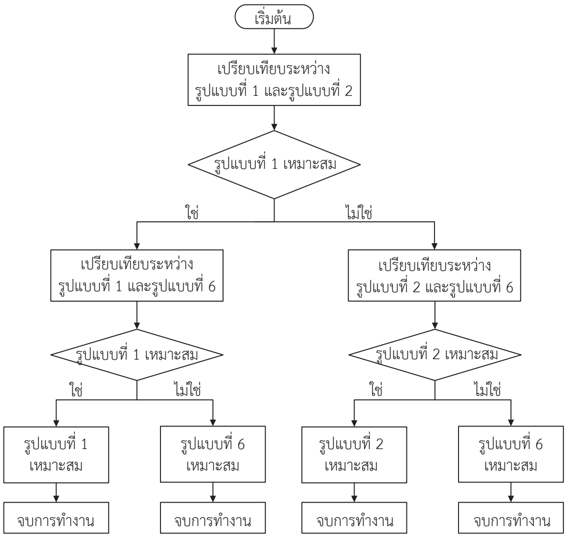
$$D(0,1) = 2\{l_1(M_1) - l_0(M_0)\} \tag{6}$$

เมื่อ $l_0(M_0)$ และ $l_1(M_1)$ คือ ค่าล็อกภาวะน่าจะเป็นสูงสุดของแบบจำลอง M_0 และ M_1 ตามลำดับ

จากสมการที่ (6) จะได้ว่าตัวสถิติ D จะมีการแจกแจงลู่เข้าการแจกแจงไคกำลังสอง (Chi-square Distribution) ที่มีองศาเสรีเท่ากับ k (χ_k^2) เมื่อ k คือ ผลต่างระหว่างมิติของพารามิเตอร์ในแบบจำลอง M_0 และ M_1 โดยจะปฏิเสธแบบจำลอง M_0 ที่ระดับนัยสำคัญ (Level of Significance; α) ถ้า $D > c_\alpha$ เมื่อ c_α คือ ควอนไทล์ (Quantile) ที่ $(1-\alpha)$ ของ



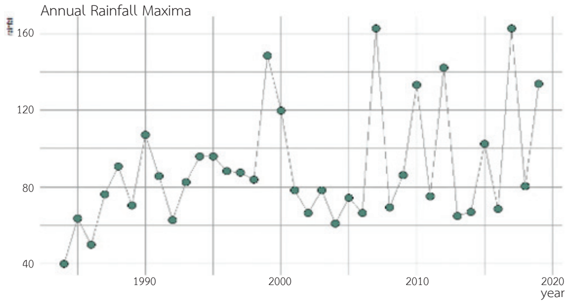
รูปที่ 2 ขั้นตอนการเปรียบเทียบรายคู่ตามลำดับ



รูปที่ 3 ผังงานขั้นตอนการเปรียบเทียบรายคู่ของกรณีที่ 1

χ_k^2 ดังนั้นถ้าปฏิเสธแบบจำลอง M_0 แล้วจะได้ว่าแบบจำลอง M_1 สามารถอธิบายความผันแปรของข้อมูลได้ดีกว่าแบบจำลอง M_0 โดยการทดสอบด้วยตัวสถิติ D จะทำได้ครั้งละ 1 คู่ เมื่อได้แบบจำลองที่เหมาะสมในการทดสอบแล้ว จึงนำแบบจำลองที่ได้ไปทดสอบกับแบบจำลองอื่นๆ จนกระทั่งได้แบบจำลองที่เหมาะสมที่สุด

ดังนั้น จากตัวอย่างรูปแบบพารามิเตอร์ข้างต้น เมื่อให้ M_i แทนรูปแบบที่ i จะได้ว่า $M_1 \subset M_3 \subset M_7$, $M_1 \subset M_2 \subset M_6$, $M_1 \subset M_4 \subset M_8$ และ $M_1 \subset M_5$ โดยขั้นตอนการเปรียบเทียบรายคู่ตามลำดับแสดงดังรูปที่ 2 และผังงาน (Flowchart) แสดงขั้นตอนการเปรียบเทียบรายคู่ของกรณีแรก ($M_1 \subset M_2 \subset M_6$) แสดงดังรูปที่ 3



รูปที่ 4 ปริมาณน้ำฝนสูงสุดรายปีของจังหวัดชัยภูมิ ตั้งแต่ พ.ศ. 2527–2562 [15]

เมื่อได้แบบจำลองที่เหมาะสมที่สุดแล้ว จะนำแบบจำลองที่ได้ไปประมาณระดับการเกิดซ้ำ (Return Level Estimation) ณ เวลา T ปี $(\hat{R}_T)$ ได้จาก

$$\hat{R}_T = \begin{cases} \hat{\mu}(t) - \dfrac{\hat{\sigma}(t)}{\hat{\xi}(t)}\left\{1 - \left[-\log\left(1 - \dfrac{1}{T}\right)\right]^{-\hat{\xi}(t)}\right\} & ;\xi \neq 0 \\ \hat{\mu}(t) - \hat{\sigma}(t)\log\left\{-\log\left(1 - \dfrac{1}{T}\right)\right\} & ;\xi = 0 \end{cases} \quad (7)$$

ในหัวข้อถัดไปจะกล่าวถึงตัวอย่างการวิเคราะห์แบบจำลองภายใต้สมมติฐานกระบวนการไม่คงที่

6. ตัวอย่างการวิเคราะห์ข้อมูล

6.1 ข้อมูล

ข้อมูลที่ใช้เป็นตัวอย่างในการวิเคราะห์เป็นข้อมูลปริมาณน้ำฝนสูงสุดรายปีของจังหวัดชัยภูมิ ตั้งแต่ พ.ศ. 2527–2562 โดยนำข้อมูลไปพล็อตกราฟเพื่อดูลักษณะของข้อมูล ได้ดังรูปที่ 4

จากรูปที่ 4 พบว่า ปริมาณน้ำฝนสูงสุดรายปีของจังหวัดชัยภูมิเป็นข้อมูลที่มีความเหมาะสมที่จะนำมาวิเคราะห์ด้วยกระบวนการไม่คงที่ เนื่องจากข้อมูลมีแนวโน้มที่เพิ่มขึ้นซึ่งอาจเกิดจากอิทธิพลจากตัวแปรอื่นๆ

6.2 ค่าประมาณพารามิเตอร์และรูปแบบของพารามิเตอร์ที่เหมาะสม

เมื่อนำรูปแบบทั้ง 8 รูปแบบ ที่กล่าวมาข้างต้นมาวิเคราะห์ จากผลวิเคราะห์จะได้จำนวนพารามิเตอร์ ค่าประมาณพารามิเตอร์ที่เหมาะสม และค่าล็อกภาวะน่าจะเป็นเชิงลบน้อยที่สุด (Negative log-likelihood; nllh) ของแต่ละรูปแบบแสดงในตารางที่ 1

จากตารางที่ 1 เมื่อนำค่า nllh มาเรียงจากน้อยไปหามากจะได้ รูปแบบที่ 4 (162.9838) รูปแบบที่ 3 (165.0347) รูปแบบที่ 2 (166.6437) รูปแบบที่ 8 (167.5585) รูปแบบที่ 5 (167.8340) รูปแบบที่ 1 (169.2635) รูปแบบที่ 6 (175.7565) และรูปแบบที่ 7 (195.4657) ตามลำดับ

ทำการเปรียบเทียบรายคู่ตามลำดับ จากความสัมพันธ์ $M_1 \subset M_2 \subset M_6$, $M_1 \subset M_3 \subset M_7$, $M_1 \subset M_4 \subset M_8$ และ $M_1 \subset M_5$ ด้วยสถิติดีเวียนซ์ (D) โดยที่ค่า $l_i(M_i)$ คือ ค่า –nllh ของแบบจำลองที่ i และที่ระดับนัยสำคัญ (α) 0.05 จะได้ผลการวิเคราะห์ แสดงในตารางที่ 2

ตารางที่ 1 จำนวนพารามิเตอร์ ค่าประมาณพารามิเตอร์ และค่าล็อกภาวะน่าจะเป็นสูงสุด (nllh) ของแต่ละรูปแบบ

รูปแบบ	จำนวนพารามิเตอร์	ค่าประมาณพารามิเตอร์	nllh
รูปแบบที่ 1	3	$\hat{\mu} = 75.4553,\ \hat{\sigma} = 21.8695,\ \hat{\xi} = 0.0607$	169.2635
รูปแบบที่ 2	4	$\hat{\mu} = 62.7384 + 0.6438t,\ \hat{\sigma} = 18.8038,\ \hat{\xi} = 0.2008$	166.6437
รูปแบบที่ 3	5	$\hat{\mu} = 54.6965 + 2.1794t - 0.0474t^2,\ \hat{\sigma} = 16.8224,\ \hat{\xi} = 0.3253$	165.0347
รูปแบบที่ 4	5	$\hat{\mu} = 72.1178 - 90.9835\exp(-0.1983t),\ \hat{\sigma} = 15.8318,\ \hat{\xi} = 0.3291$	162.9838
รูปแบบที่ 5	4	$\hat{\mu} = 74.4703,\ \hat{\sigma} = \exp(3.4944 - 0.0348t),\ \hat{\xi} = 0.4923$	167.8340l
รูปแบบที่ 6	5	$\hat{\mu} = 91.0330 - 0.4908t,\ \hat{\sigma} = \exp(3.2002 + 0.0167t),\ \hat{\xi} = -0.3066$	175.7565
รูปแบบที่ 7	6	$\hat{\mu} = -9.4116 - 0.6145t + 0.0883t^2,\ \hat{\sigma} = \exp(4.9012 - 0.0057t),\ \hat{\xi} = -0.8424$	195.4657
รูปแบบที่ 8	6	$\hat{\mu} = 60.8735 + 13.5237\exp(-0.8294t),\ \hat{\sigma} = \exp(2.8956 + 0.0081t),\ \hat{\xi} = 0.0466$	167.5585

ตารางที่ 2 ค่าสถิติดีเวียนซ์ (D) ผลต่างระหว่างมิติของพารามิเตอร์ (k) และรูปแบบที่เลือก

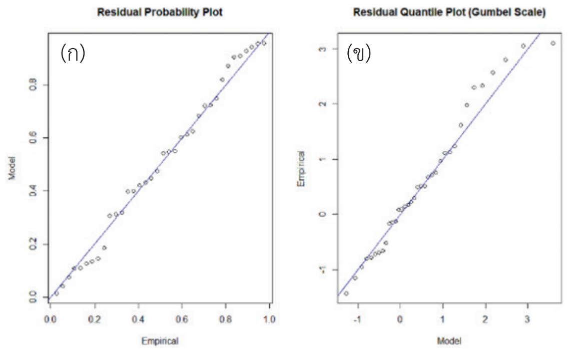
เปรียบเทียบรูปแบบ (i, j)	$D(I, j)$	k	χ_k^2	รูปแบบที่เลือก
1 กับ 2	5.2394	1	3.841	2
2 กับ 6	–18.2256	1	3.841	2
1 กับ 3	8.4575	2	5.991	3
3 กับ 7	–60.8620	1	3.841	3
1 กับ 4	12.5593	2	5.991	4
4 กับ 8	–9.1493	1	3.841	4
1 กับ 5	2.8590	1	3.841	1
2 กับ 3	3.2181	1	3.841	2
2 กับ 4	**7.3198**	**1**	**3.841**	**4**

จากตารางที่ 2 พบว่า ในการเปรียบเทียบรูปแบบที่รูปแบบที่ 1 กับรูปแบบที่ 2 จะได้ $D(1, 2) = 5.2394 > \chi_1^2 = 3.841$ ดังนั้นปฏิเสธรูปแบบที่ 1 นั่นคือ รูปแบบที่ 2 มีความเหมาะสม จากนั้นนำรูปแบบที่ 2 ไปเปรียบเทียบกับรูปแบบที่ 6 จะได้ $D(2, 6) = -18.2256 < \chi_1^2 = 3.841$ ดังนั้นยอมรับรูปแบบที่ 2 นั่นคือ รูปแบบที่ 2 มีความเหมาะสม และในทำนองเดียวกันกับความสัมพันธ์อื่นๆ จะได้ว่า ความสัมพันธ์ที่ 1 รูปแบบที่ 2 เหมาะสม ความสัมพันธ์ที่ 2 รูปแบบที่ 3 เหมาะสม ความสัมพันธ์ที่ 3 รูปแบบที่ 4 เหมาะสม และความสัมพันธ์ที่ 4 รูปแบบที่ 1 เหมาะสม แต่ในกรณีความสัมพันธ์ที่ 4 ที่รูปแบบที่ 1 เหมาะสมจะไม่ถูกนำมาเปรียบเทียบเนื่องจากรูปแบบที่ 2 รูปแบบที่ 3 และรูปแบบที่ 4 เหมาะสมกว่ารูปแบบที่ 1 จากนั้นนำรูปแบบที่ 2 รูปแบบที่ 3 และรูปแบบที่ 4 มาทำการเปรียบเทียบรายคู่ต่อไป ซึ่งพบว่า รูปแบบที่ 4 เป็นรูปแบบเหมาะสมที่สุด ดังนั้นสามารถเขียนแบบจำลองที่เหมาะสำหรับข้อมูลปริมาณน้ำฝนสูงสุดรายปีของจังหวัดชัยภูมิได้ดังนี้

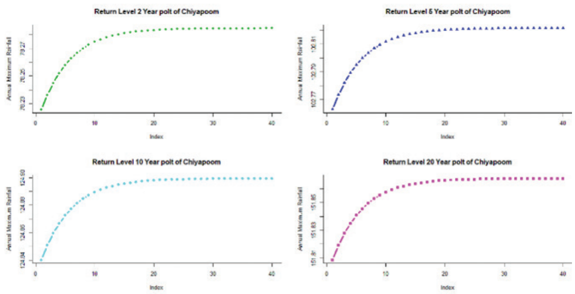
$X \sim \mathrm{GEV}\big(\mu(t), \sigma, \xi\big)$

เมื่อ $\mu(t) = 72.1178 - 90.9835\exp(-0.1983t)$, $\sigma = 15.8318$ และ $\xi = 0.3291$

เมื่อนำแบบจำลองที่เหมาะสมที่ได้จากการเลือกรูปแบบ 4 ไปพล็อตกราฟความน่าจะเป็น (Probability Plot) และกราฟควอนไทล์ (Quantile Plot) เพื่อดูความเหมาะสมของแบบจำลอง ซึ่งได้ดังรูปที่ 5



รูปที่ 5 (ก) กราฟความน่าจะเป็น และ (ข) กราฟควอนไทล์ของแบบจำลองรูปแบบที่ 4



รูปที่ 6 กราฟระดับการเกิดซ้ำ (Return Level Plot)

จากรูปที่ 5 พบว่า แบบจำลองที่ได้มีความเหมาะสมกับข้อมูล

นำแบบจำลองที่ได้มาประมาณระดับการเกิดซ้ำ 2 ปี 5 ปี 10 ปี และ 20 ของปริมาณน้ำฝนสูงสุดรายปีของจังหวัดชัยภูมิ ซึ่งผลการประมาณที่ได้แสดงดังรูปที่ 6

จากรูปที่ 6 พบว่า ค่าประมาณระดับการเกิดซ้ำของปริมาณน้ำฝนสูงสุดรายปีของจังหวัดชัยภูมิ ในรอบปีการเกิดซ้ำ 2 ปี จะมีปริมาณน้ำฝนสูงสุดรายปีประมาณ 78 มิลลิเมตร ด้วยความน่าจะเป็น 0.5 สำหรับรอบปีการเกิดซ้ำ 5 ปี จะมีปริมาณน้ำฝนสูงสุดรายปีประมาณ 102 มิลลิเมตร ด้วยความน่าจะเป็น 0.2 สำหรับรอบปีการเกิดซ้ำ 10 ปี จะมีปริมาณน้ำฝนสูงสุดรายปีประมาณ 124 มิลลิเมตร ด้วยความน่าจะเป็น 0.1 และรอบปีการเกิดซ้ำ 20 ปี จะมีปริมาณ

น้ำฝนสูงสุดรายปีประมาณ 151 มิลลิเมตร ด้วยความน่าจะเป็น 0.05

7. สรุป

เนื่องจากข้อมูลในปัจจุบันไม่ได้มีแค่ข้อมูลที่มีความคงที่ แต่ยังมีข้อมูลที่มีลักษณะที่ขึ้นอยู่กับปัจจัยอื่นๆ ซึ่งจะเรียกข้อมูลลักษณะนี้ว่า ข้อมูลที่มีความไม่คงที่ ดังนั้นเพื่อให้ได้แบบจำลองที่เหมาะสมกับข้อมูลที่สุด บทความนี้ผู้เขียนจึงได้กล่าวถึงวิธีการวิเคราะห์ข้อมูลที่มีความไม่คงที่ โดยใช้ทฤษฎีค่าสุดขีดผ่านการแจกแจงค่าสุดขีดนัยทั่วไป (GEV) วิธีการเลือกแบบจำลองที่ดีที่สุดโดยใช้สถิติดีเวียนซ์ใช้ในการทดสอบ และรวมไปถึงการประมาณระดับการเกิดซ้ำ (Return Level Estimation) ณ เวลาต่างๆ ของแบบจำลองที่ได้ พร้อมทั้งยกตัวอย่างการวิเคราะห์ข้อมูล โดยใช้ข้อมูลปริมาณน้ำฝนสูงสุดรายปีของจังหวัดชัยภูมิ ตั้งแต่ พ.ศ. 2527–2562 ซึ่งผลการวิเคราะห์พบว่า แบบจำลองที่เหมาะสมสำหรับข้อมูลปริมาณน้ำฝนสูงสุดรายปีของจังหวัดชัยภูมิคือ แบบจำลองภายใต้กระบวนการไม่คงที่ ดังนี้

$$X \sim \mathrm{GEV}\left(\mu(t), \sigma, \xi\right)$$

เมื่อ $\mu(t) = 72.1178 - 90.9835\exp(-0.1983t)$, $\sigma = 15.8318$ และ $\xi = 0.3291$ และมีระดับการเกิดซ้ำเพิ่มขึ้นเมื่อรอบปีการเกิดซ้ำเพิ่มขึ้น

เอกสารอ้างอิง

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

[2] J. Beirlant, Y. Goegebeur, J. Segers, and J. L. Teugels, *Statistics of Extremes: Theory and Applications*. New York: John Wiley & Sons, 2004.

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

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

[5] P. Embrecht, C. Kluppelberg, and T. Mikosch, *Modeling Extremal Events for Insurance and Finance*. Berlin: Springer Verlag, 1997.

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

[7] B. Finkenstadt and H. Rootzen, *Extreme Values in Finance, Telecommunications, and the Environment*. London: Chapman and Hall/CRC Press, 2004.

[8] 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 (in Thai).

[9] 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.

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

[11] 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, pp. 535–545, 2005.

[12] 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, pp. 158–171, 1955.

[13] J. Galambos, *The Asymptotic Theory of Extreme*

Order Statistics. New York: Wiley, 1978.

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

[15] Meteorological Department of Thailand. (2020, May). Weather forecast. Thai Meteorological Department. Bangkok, Thailand [Online]. (in Thai). Available: http://www.tmd.go.th