สถิติค่าสุดขีด

ปิยภัทร บุษบาบดินทร์* และ อรุณ แก้วมั่น
อาจารย์ ภาควิชาคณิตศาสตร์ คณะวิทยาศาสตร์ มหาวิทยาลัยมหาสารคาม

* ผู้นิพนธ์ประสานงาน โทรศัพท์ 08-9542-6396 อีเมล: piyapatr.b@msu.ac.th
รับเมื่อ 15 กรกฎาคม 2557 ตอบรับเมื่อ 15 มกราคม 2558 เผยแพร่ออนไลน์ 15 พฤษภาคม 2558
DOI: 10.14416/j.kmutnb.2015.01.003

บทคัดย่อ

เป้าหมายของการวิเคราะห์แบบจำลองคือ การได้แบบจำลองที่ดีที่สุดสำหรับข้อมูลที่ศึกษา แต่เมื่อข้อมูลที่ศึกษามีค่าสุดขีดเกิดขึ้น นักวิเคราะห์ส่วนใหญ่มักจะตัดข้อมูลนั้นทิ้งไปไม่นำมาพิจารณาเพื่อหาแบบจำลอง เนื่องจากมีความซับซ้อนและยุ่งยากในการวิเคราะห์ แต่ในความเป็นจริง ถ้านักวิเคราะห์ต้องการทราบถึงความน่าจะเป็นในการเกิดขึ้นของเหตุการณ์ที่มีค่าสูงสุดหรือต่ำสุดซึ่งอยู่ในส่วนของปลายหางซึ่งมีค่าน้อยมาก เครื่องมือทางสถิติที่มีบทบาทเกี่ยวข้องในเรื่องนี้คือ "ทฤษฎีค่าสุดขีด" บทความนี้มีจุดประสงค์เพื่อยกตัวอย่างการประยุกต์ใช้ทฤษฎีค่าสุดขีดกับข้อมูลจริงในสาขาวิชาต่าง ๆ พร้อมทั้งกล่าวถึงแนวความคิดและการพัฒนาทฤษฎีค่าสุดขีด และทำการสรุปสถิติอนุมานของค่าสุดขีด การแจกแจงของค่าสุดขีด ได้แก่ การแจกแจงค่าสุดขีดวางนัยทั่วไป และการแจกแจงพาเรโตวางนัยทั่วไป เป็นต้น พร้อมทั้งตรวจสอบความเหมาะสมของแบบจำลองค่าสุดขีด คาบเวลาการเกิดซ้ำ และการหาค่าระดับการเกิดซ้ำในรอบปีที่สนใจ

คำสำคัญ: ทฤษฎีค่าสุดขีด การแจกแจงค่าสุดขีดวางนัยทั่วไป การแจกแจงพาเรโตวางนัยทั่วไป ระดับการเกิดซ้ำ คาบเวลาการเกิดซ้ำ

การอ้างอิงบทความ: ปิยภัทร บุษบาบดินทร์ และ อรุณ แก้วมั่น, "สถิติค่าสุดขีด," *วารสารวิชาการพระจอมเกล้าพระนครเหนือ*, ปีที่ 25, ฉบับที่ 2, หน้า 315 - 324, พ.ค. - ส.ค. 2558. http://dx.doi.org/10.14416/j.kmutnb.2015.01.003

Extreme Values Statistics

Piyapatr Busababodhin* and Arun Keawmun
Lecturer, Department of Mathematics, Faculty of Science, Mahasarakham University, Maha Sarakham, Thailand

* Corresponding Author, Tel. 08-9542-6396, E-mail: piyapatr.b@msu.ac.th
Received 15 July 2014; Accepted 15 January 2015; Published online: 15 May 2015
DOI: 10.14416/j.kmutnb.2015.01.003

Abstract

One of the greatest achievements of modeling is to find an optimal model for the data. If the extreme value is included, an analyst usually cuts them out from the data because of its complexity. In practice, if an analyst wants to know the extreme event's probability which there is extreme values in the tailed . The statistical tool which is very popular which is called "Extreme Value Theory." This article aims to give a sample, self-contained introduction to the motivations and basic ideas behind the development of extreme value theory. Also briefly covered are the inference statistics of extreme value and its distribution such as generalized extreme value distribution and generalized Pareto distribution theory, how to check and find optimal extreme value modeling, and return period and return level.

Keywords: Extreme Value Theory, Generalized Extreme Value Distribution, Generalized Pareto Distribution, Return Level, Return Period

Please cite this article as: P. Busababodhin and A. Keawmun, "Extreme Values Statistics," *J. KMUTNB*, Vol. 25, No. 2, pp. 315 - 324, May. - Aug. 2015 (in Thai). http://dx.doi.org/10.14416/j.kmutnb.2015.01.003

1. บทนำ

ทฤษฎีค่าสุดขีด (Extreme Value Theorem) เป็นทฤษฎีที่กล่าวถึงคุณสมบัติของเหตุการณ์ที่มีตัวแปรสุ่มซึ่งจัดอยู่ในลักษณะที่เรียกว่า "ค่าสุดขีด" อาจจะเป็นค่าสูงสุดหรือต่ำสุดก็ได้ พร้อมทั้งศึกษารูปแบบการแจกแจงความน่าจะเป็นของตัวแปรสุ่มเหล่านี้ การวิเคราะห์ข้อมูลเมื่อข้อมูลมีค่าสุดขีด (Extreme Value) เกิดขึ้น นักวิเคราะห์ส่วนใหญ่จะตัดข้อมูลส่วนนั้นทิ้งไปไม่นำมาพิจารณา เนื่องจากมีความซับซ้อนและยุ่งยากในการวิเคราะห์ แต่ในความเป็นจริงถ้าต้องการทราบถึงความน่าจะเป็นในการเกิดขึ้นของเหตุการณ์ที่มีค่าสูงสุดหรือต่ำสุด ซึ่งอยู่ในส่วนของปลายหางซึ่งมีค่าน้อยมาก อาทิเช่น ปริมาณน้ำฝนสูงสุด-ต่ำสุดในแต่ละวัน ความเร็วลมสูงสุดในรอบเดือน อุณหภูมิสูงสุด-ต่ำสุดในแต่ละวัน เป็นต้น ก็สามารถทำได้

การศึกษาทฤษฎีค่าสุดขีดเริ่มต้นมาตั้งแต่ทศวรรษที่ 19 และได้ถูกพัฒนามาอย่างต่อเนื่องโดยนักคณิตศาสตร์ ในปี ค.ศ. 2000 Kotz และ Nadaraja [1] ได้กล่าวว่าการแจกแจงค่าสุดขีดได้ถูกค้นพบครั้งแรกในปี ค.ศ. 1709 โดย Bernulli และถูกประยุกต์ใช้ครั้งแรกโดย Fuller ในปี ค.ศ. 1914 ซึ่งในช่วงระยะเวลาสิบปีที่ผ่านมามีงานวิจัยที่เกี่ยวข้องกับสถิติของค่าสุดขีดถูกตีพิมพ์เผยแพร่เป็นจำนวนมาก ซึ่งงานวิจัยเหล่านี้ได้สรุปความรู้พื้นฐานและเครื่องมือที่ใช้สำหรับวิเคราะห์ค่าสุดขีดในเบื้องต้น เพื่อใช้ประกอบการตัดสินใจและหาแนวทางในการป้องกันและแก้ไขสถานการณ์ต่าง ๆ เช่น ด้านเศรษฐศาสตร์ ได้นำทฤษฎีค่าสุดขีดช่วยประเมินค่าและราคาของการทำประกัน (Insurance) เมื่อพิจารณาโอกาสของเหตุการณ์ที่ก่อให้เกิดความหายนะอย่างใหญ่หลวงที่อาจจะเกิดขึ้น ถึงแม้ในความเป็นจริงแล้วเหตุการณ์เหล่านี้มีโอกาสจะเกิดขึ้นได้ยากก็ตาม หรือใช้ทฤษฎีนี้เพื่อประมาณมูลค่าของความเสี่ยง (Value at Risk: VaR) ในสถาบันการเงินและบริษัทที่มีการลงทุนด้านหลักทรัพย์หรือสินทรัพย์ (รายละเอียดเพิ่มเติม [2] - [5]) ด้านอุทกวิทยา ใช้ทฤษฎีค่าสุดขีดเพื่อตรวจสอบความเสี่ยงของเหตุการณ์รุนแรงทางสิ่งแวดล้อมที่จะเกิดขึ้น เช่น การหาความสูงของคลื่นในทะเลเพื่อป้องกันการเกิดน้ำท่วมของพื้นที่ชายฝั่ง การหาปริมาณน้ำฝนสูงสุดเพื่อป้องกันการเกิดน้ำท่วม (รายละเอียดเพิ่มเติม [6] - [8]) ด้านวิศวกรรม ประยุกต์ใช้ทฤษฎีค่าสุดขีดเพื่อหาขอบเขตของความต้านทานของวัตถุเพื่อคำนวณการเสื่อมสภาพของผลิตภัณฑ์อันเนื่องจากการขยายตัวอย่างช้า ๆ ของรอยแตกตามระยะเวลาการใช้งานในเชิงพลวัต (Dynamic) และความเชื่อถือได้ในการก่อสร้างอาคาร เช่น การสร้างสะพาน อุปกรณ์ขุดเจาะน้ำมัน และใช้ในการประมาณค่าระดับมลพิษที่เกิดขึ้น (รายละเอียดเพิ่มเติม [9] - [11])

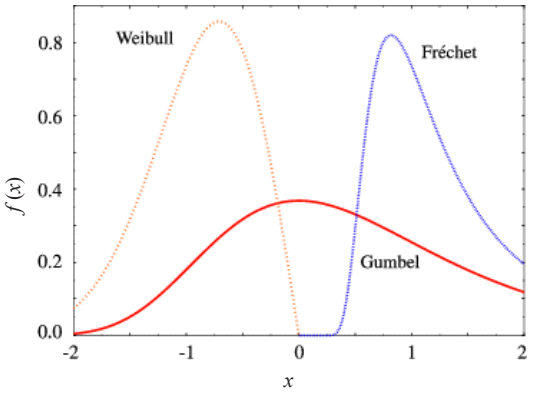
โดยหัวข้อถัดไปผู้เขียนได้กล่าวถึงความเป็นมาของค่าสุดขีด การแจกแจงของค่าสุดขีด การประมาณค่าพารามิเตอร์ การตรวจสอบความเหมาะสมของแบบจำลองค่าสุดขีด และคาบเวลาการเกิดซ้ำและการหาระดับการเกิดซ้ำ ตามลำดับ

2. ความเป็นมาของค่าสุดขีด

เหตุการณ์สุดขีด (Extreme Events) คือเหตุการณ์ของตัวแปรสุ่มที่มีโอกาสเกิดขึ้นมากกว่าค่าคาดหวังของตัวแปรสุ่ม สมมติให้ X_i เมื่อ $i = 1,2,...,n$ เป็นตัวแปรสุ่มที่อิสระต่อกัน และมีฟังก์ชันความน่าจะเป็นสะสม $F(x;\theta) \equiv \Pr(X_i \le x)$ เดียวกัน กำหนดให้ค่าสูงสุดของตัวแปรสุ่ม คือ $X_{(n)} = Max(X_1, X_2,..., X_n)$ ที่มีฟังก์ชันการแจกแจงเป็น $F_n(x)$ ซึ่งมีความสัมพันธ์กับ $F(x;\theta)$ โดยมีเงื่อนไขดังต่อไปนี้

$$\begin{aligned} F_n(n) &\equiv \Pr(X_{(n)} \le x) \\ &\equiv \Pr(X_1 \le x, X_2 \le x,... X_n \le x) \\ &\equiv \Pr(X_1 \le x) \cdot \Pr(X_2 \le x) \cdot ... \cdot \Pr(X_n \le x) \equiv F_n(x) \cdot \end{aligned}$$

เมื่อ $n \to \infty$ พบว่า $\Pr(X_{(n)} \le x)$ ลู่เข้าสู่ $F(x;\theta)$ และ $x_{(n)} = (X_{(n)} - b_n)/a_n$ เป็นฟังก์ชันการแจกแจงความน่าจะเป็นแบบนอนดีเจนเนอเรต (Non-degenerate Distribution) ของ $F(x;\theta)$ โดยที่ a_n และ b_n เป็นค่าคงที่ ดังนั้น



รูปที่ 1 การแจกแจงกัมเบลฟรีเชทและไวล์บูลเมื่อกำหนด $\alpha = 0$

$$\begin{aligned}\Pr(X_{(n)} \le x) &= \Pr((X_{(n)} - b_n)/a_n \le x) \\ &= \Pr(X_{(n)} \le a_n x + b_n) \\ &= F^n(a_n x + b_n) \to F(x) \text{ เมื่อ } n \to \infty .\end{aligned}$$

อาจจะเรียกทฤษฎีค่าสุดขีดว่า “ทฤษฎีสามแบบ” (Three Types Theorem) ตามลักษณะของลิมิตการแจกแจงของฟังก์ชัน $F(x;\theta)$ ตามรูปที่ 1 [12] ดังนี้

แบบที่ 1 กัมเบล (Gumbel Type)

$F(x) = \exp(-\exp(-x))$

เมื่อ $-\infty < x < \infty$;

แบบที่ 2 ฟรีเชท (Fréchet Type)

$F(x) = \exp(-x^{-a})$

เมื่อ $-\infty < x < \infty$ และ $a > 0$

พบว่าถ้า $F(x) = 0$ เมื่อ $x < 0$;

แบบที่ 3 ไวล์บูล (Weibull Type)

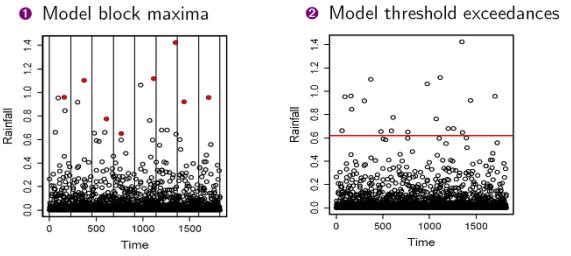
$F(x) = \exp(-(-x)^{-a})$

เมื่อ $-\infty < x < \infty$และ $a > 0$

พบว่าถ้า $F(x) = 1$ เมื่อ $x > 0$.

3. การแจกแจงของค่าสุดขีด

การวิเคราะห์แบบจำลองค่าสุดขีดด้วยทฤษฎีค่าสุดขีดสามารถแบ่งลักษณะการแจกแจงของค่าสุดขีดได้เป็น 2 ประเภทตามลักษณะของการเลือกข้อมูลค่าสุดขีดที่นำมาวิเคราะห์ ดังแสดงในรูปที่ 2 ได้แก่ การแจกแจงค่าสุดขีดวางนัยทั่วไป (Generalized Extreme Value Distribution: GEV) และการแจกแจงพาเรโตวางนัยทั่วไป (Generalized Pareto Distribution: GPD) ซึ่งการสร้างแบบจำลองด้วย GEV เหมาะสำหรับวิเคราะห์ค่าสุดขีดในช่วงคาบเวลาที่สนใจ เช่น รายปี รายเดือน รายไตรมาส หรือรายสัปดาห์ เป็นต้น ซึ่งค่าสังเกตที่รวบรวมไว้ควรจะมีจำนวนมากกว่า 30 ปีขึ้นไป โดยจะเลือกข้อมูลที่สูงสุดในแต่ละช่วงคาบเวลาที่ผู้วิเคราะห์สนใจ แต่ถ้าต้องการวิเคราะห์การแจกแจงของปลายหางของข้อมูลเหล่านี้เมื่อข้อมูลมีจำนวนมาก หรือข้อมูลเก็บรวบรวมเป็นรายวัน การสร้างแบบจำลองด้วย GPD จะมีความเหมาะสมกว่า GEV เนื่อง GPD จะอธิบายลักษณะข้อมูลที่มีการแจกแจงแบบมีหางที่หนัก (Heavy-tailed Distribution) ได้ดีกว่า และจำนวนค่าสุดขีดที่นำมาวิเคราะห์ด้วย GPD จะมีจำนวนมากกว่าข้อมูลที่นำมาวิเคราะห์ด้วย GEV ซึ่งสามารถลดความไม่แน่นอนที่เกิดขึ้นจากการสุ่มตัวอย่างได้ การสร้างแบบจำลองด้วย GPD มีขั้นตอนสำคัญคือ การกำหนดค่าเกณฑ์ (Threshold) ที่เหมาะสมกับข้อมูลที่นำมาวิเคราะห์ และการพิจารณาความไม่เป็นอิสระของข้อมูลค่าสุดขีด ซึ่งสามารถแก้ไขได้โดยการจัดกลุ่มค่าสุดขีด (Declustering) ที่มีค่าเกินกว่าค่าเกณฑ์



รูปที่ 2 การเลือกค่าสุดขีดสำหรับแบบจำลอง GEV (A) และ GPD (B)

ดังนั้น การสร้างแบบจำลองด้วย GPD จึงได้รับความนิยมอย่างแพร่หลายและเหมาะสำหรับวิเคราะห์ข้อมูลด้านอุตุวิทยาและอุทกวิทยาเพื่อวิเคราะห์ความ

เสียหายของเหตุการณ์รุนแรงที่มีโอกาสเกิดขึ้นได้ยาก (รายละเอียดเพิ่มเติม [6], [7], [9], [10], [13], [14])

3.1 การแจกแจงค่าสุดขีดวางนัยทั่วไป (Generalized Extreme Value Distribution: GEV)

สมมติให้ X_i เมื่อ $i = 1, 2,..., n$ เป็นตัวแปรสุ่มที่อิสระต่อกันและมีฟังก์ชันหนาแน่นความน่าจะเป็น $F(x;\theta)$ เดียวกัน ค่าสูงสุดของตัวแปรสุ่มคือ $X_{(n)} = Max(X_1, X_2,..., X_n)$ ซึ่งจะประยุกต์ใช้ในเรื่องนี้ในรูปแบบของการแจกแจงค่าสุดขีดวางนัยทั่วไป ที่มีพารามิเตอร์ที่กำกับการเกิดขึ้นทั้งหมด 3 พารามิเตอร์คือ μ แสดงถึงตำแหน่ง (Location) σ แสดงถึงขนาด (Scale) และ ξ แสดงถึงรูปร่าง (Shape)

สำหรับ GEV ถูกพัฒนาขึ้นในค.ศ.1955 โดยJenkinson [15] สามารถเขียนฟังก์ชันการแจกแจงค่าสุดขีดได้ 3 การแจกแจง ได้แก่การแจกแจงกัมเบล (Gumbel Distribution) การแจกแจงฟรีเชท (Fréchet Distribution) และการแจกแจงไวล์บูล (Weibull Distribution) ต่อมาในปี ค.ศ. 1978 Galambos [16] ได้สร้างฟังก์ชันการแจกแจงสะสม (Cumulative Distribution Function: CDF) ของ GEV ดังนี้

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

และสามารถเขียนฟังก์ชันการแจกแจงความน่าจะเป็น (Probability Distribution Function: pdf) ของ GEV ดังนี้

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

สำหรับ $1+\xi\left(\frac{x-\mu}{\sigma}\right) > 0.$

จากสมการที่ (1) และ (2) พบว่า เมื่อ $\xi = 0$ เรียกการแจกแจงค่าสุดขีดวางนัยทั่วไปว่า “การแจกแจงกัมเบล” เมื่อ $\xi > 0$ เรียกการแจกแจงค่าสุดขีดวางนัยทั่วไปว่า “การแจกแจงฟรีเชท” และเมื่อ $\xi < 0$ เรียกการแจกแจงค่าสุดขีดวางนัยทั่วไปว่า “การแจกแจงไวล์บูล”

สำหรับความสัมพันธ์ระหว่างพารามิเตอร์ μ และ σ กับโมเมนต์ที่ k ของสมการที่ (1) เป็นดังนี้

ถ้ากำหนด $\xi < \frac{1}{k}$ สามารถคำนวณโมเมนต์ที่ 1 ของสมการที่ (1) ได้จาก

$$E(x) = \mu + \frac{\sigma}{\xi}\left(\Gamma(1-\xi)-1\right),$$

เมื่อ $\Gamma(x)$ แทนฟังก์ชันแกมมา

กรณีที่ $\xi \to 0$ สามารถคำนวณโมเมนต์ที่ 1 ได้จาก $E(x) = \mu + \sigma\xi$ เมื่อ $\xi = 0.577...$ เป็นค่าคงที่ออยเลอร์ (Euler Constant)

และสามารถคำนวณโมเมนต์ที่ 2 (ความแปรปรวน) ของสมการที่ (1) เมื่อกำหนด $\xi < \frac{1}{2}$ ดังนี้

$$E\left(\left(x-E(x)\right)^2\right) = \left(\frac{\sigma}{\xi}\right)^2\left(\Gamma(1-2\xi)-\Gamma^2(1-\xi)\right),$$

และเมื่อ $\xi \to 0$ พบว่า $E\left(\left(x-E(x)\right)^2\right) = \sigma^2\pi^2/6.$

3.2 การแจกแจงพาเรโตวางนัยทั่วไป (Generalized Pareto Distribution: GPD)

การวิเคราะห์แบบจำลองของค่าสุดขีดด้วย GEV ซึ่งเป็นวิธีที่พิจารณาข้อมูลที่สูงสุดในแต่ละช่วงคาบเวลาที่ผู้วิเคราะห์สนใจ แต่ถ้าต้องการวิเคราะห์การแจกแจงของปลายหางของข้อมูลเหล่านี้ หรือวิเคราะห์ $\bar{F}(x) = 1 - F(x)$ เมื่อข้อมูลมีจำนวนมาก ทำได้โดยประยุกต์ใช้อนุกรมเทเลอร์ (Taylor Series Extension) กับสมการที่ (1) ซึ่งสามารถเขียนฟังก์ชันใหม่ได้ดังนี้

$$1-F(x) = 1-\left(\exp\left(-\left(1+\xi\left(\frac{x-\mu}{\sigma}\right)\right)_+\right)^{-1/\xi}\right)^{1/n}$$

ซึ่งจะลู่เข้าสู่ 1 ถ้า $x \to X^F$ ดังนั้น

$$P\left(\bar{x} > u + x \middle| \bar{x} > u\right) = \frac{P(\bar{x} > u + x)}{P(\bar{x} > u)}$$

$$\cong \left(\frac{1+\xi(x+u-\mu)/\sigma}{1+\xi\left(\frac{u-\mu}{\sigma}\right)}\right)\left(\frac{n^{-1}}{n^{-1}}\right) = \left(1+\xi\left(\frac{x}{\sigma^*}\right)\right)^{1/\xi}$$

เมื่อ $\sigma^* = \sigma + \xi\left(\frac{u-\mu}{\sigma}\right)$ เรียกฟังก์ชันนี้ว่า "การแจกแจงพาเรโตวางนัยทั่วไป"

โดยปกติจะเห็นว่าเหตุการณ์ที่เกิดค่าสุดขีด ค่า X_i จะมีค่าสูงกว่าค่า u (Threshold) ที่กำหนด ถ้า u มีค่ามากจะทำให้ฟังก์ชันการแจกแจงของ X_i - u มีเงื่อนไขเมื่อ $X_i > u$ ดังนี้

$$H(x) = 1 - \left(1 + \xi\frac{x}{\tilde{\sigma}}\right)^{-1/\xi} \quad (3)$$

สำหรับ $x > 0, 1 + \xi x/\tilde{\sigma} > 0$ เมื่อ $\tilde{\sigma} = \sigma + (u - \mu)$

จากสมการที่ (3) เป็นการแจกแจงที่อยู่ในกลุ่มการแจกแจงเดียวกันกับการแจกแจงพาเรโต โดยค่า σ_u เป็นพารามิเตอร์แสดงขนาด เมื่อ $u > u_0$ พบว่า $\sigma_u = \sigma_{u_0} + \xi(u - u_0)$ ดังนั้นค่าพารามิเตอร์ขนาด (ξ) จะเปลี่ยนไป ยกเว้นเมื่อ $\xi = 0$ การแจกแจง GPD จะไม่มีการเปลี่ยนแปลง การปรับพารามิเตอร์ขนาดจะปรับโดยสมการ $\sigma^* = \tilde{\sigma} - \xi u$ เมื่อค่า u_0 คือค่าต่ำที่สุดของ u และตัวประมาณของ σ^* และ ξ เป็นค่าคงที่ ซึ่งสามารถเขียนฟังก์ชันความน่าจะเป็นของการแจกแจงพาเรโต ได้ดังนี้

$$h(x) = 1 + \left(\xi\left(\frac{x-u}{\sigma}\right)\right)^{-1/\xi} \quad (4)$$

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

4. การประมาณค่าพารามิเตอร์ (Parameter Estimation)

การประมาณพารามิเตอร์สำหรับ GEV และ GPD ในบทความนี้ผู้เขียนขอกล่าวถึงเฉพาะการประมาณค่าพารามิเตอร์แบบจุดด้วยวิธีความน่าจะเป็นสูงสุด (Maximum Likelihood Estimation: MLE) เท่านั้น

ขั้นตอนการประมาณค่าพารามิเตอร์ด้วยวิธีภาวะความน่าจะเป็นสูงสุด มีรายละเอียดดังนี้

ขั้นตอนที่ 1 พิจารณาฟังก์ชันการแจกแจงความน่าจะเป็นของ GEV ในสมการที่ (2)

ขั้นตอนที่ 2 สร้างฟังก์ชันไลค์ลิฮูด (Likelihood Function) ของฟังก์ชันการแจกแจงความน่าจะเป็นของ GEV ได้ดังนี้

$$L(f(x)) = \prod_{i=1}^{n}\left(\frac{1}{\sigma}\left(1+\xi\left(\frac{x-\varpi}{\sigma}\right)\right)^{-1/\xi+1}\exp\left(-\left(1+\xi\left(\frac{x-\mu}{\sigma}\right)\right)^{-1/\xi}\right)\right)$$

$$= \frac{1}{\sigma^n}\prod_{i=1}^{n}\left(1+\xi\left(\frac{x-\mu}{\sigma}\right)\right)^{-1/\xi+1}\exp\left(-\sum_{i=1}^{n}\left(1+\xi\left(\frac{x-\mu}{\sigma}\right)\right)^{-1/\xi}\right)$$

ขั้นตอนที่ 3 สร้างฟังก์ชันล็อคไลค์ลิฮูด (Log-likelihood Function) ของฟังก์ชันการแจกแจงความน่าจะเป็นของ GEV ที่ได้จากขั้นตอนที่ 2 ได้ดังนี้

$$\ln\left(L\left(f\left(x\right)\right)\right) = n\ln\left(\frac{1}{\sigma}\right) - \frac{1}{\xi} + \ln\prod_{i=1}^{n}\left(1+\xi\left(\frac{x-\mu}{\sigma}\right)\right)\left(-\frac{1}{\xi}\right)$$
$$\ln\left(\exp\left(-\sum_{i=1}^{n}\left(1+\xi\left(\frac{x-\mu}{\sigma}\right)\right)\right)\right)$$

หรือ

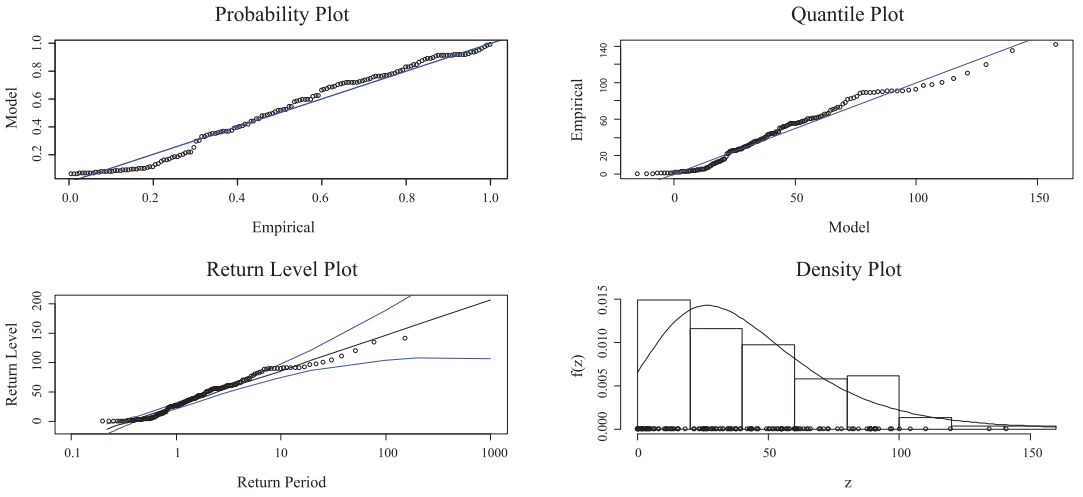
$$l(\mu,\sigma,\xi) = n\log\sigma - \left(1-\frac{1}{\xi}\right)\sum_{i=1}^{n}\log\left(1-\xi\left(\frac{x_i-\mu}{\sigma}\right)\right)$$
$$-\sum_{i=1}^{n}\left(1+\xi\left(\frac{x_i-\mu}{\sigma}\right)\right)_{+}^{-1/\xi}$$

ขั้นตอนที่ 4 ประมาณค่าพารามิเตอร์ด้วยการหาอนุพันธ์ย่อย (Partial Derivative) จากฟังก์ชันที่ได้ในขั้นตอนที่ 3 และการประมาณพามิเตอร์สำหรับ GPD ด้วยวิธี MLE จะทำลักษณะเดียวกันกับการแจกแจง GEV จากฟังก์ชันการแจกแจงความน่าจะเป็นของ GPD ในสมการที่ (4) สามารถเขียนฟังก์ชันล็อคไลค์ลิฮูดได้ดังนี้

$$l(\sigma,\xi) = -k\log\sigma - \left(1+\frac{1}{\xi}\right)\sum_{i=1}^{k}\log\left(1+\xi\frac{x_i}{\sigma}\right)_{+} \quad (5)$$

เมื่อกำหนดให้ k แทน จำนวนข้อมูลที่มีค่ามากกว่าค่า u

สำหรับการประมาณช่วงความเชื่อมั่นสำหรับพารามิเตอร์นั้น มีวิธีประมาณที่นิยมใช้กันอย่างแพร่หลาย มี 3 วิธี คือ วิธีเชิงเส้นกำกับปกติของตัวประมาณวิธีประมาณภาวะน่าจะเป็นสูงสุด (Asymptotic Normality of MLES) วิธีช่วงความเชื่อมั่นแบบควรจะเป็นโปรไฟล์ (Profile Likelihood Confidence Interval) และวิธีเดลต้า (Delta Method) ซึ่งผู้เขียนไม่ได้กล่าวรายละเอียดไว้ในบทความนี้ สามารถศึกษารายละเอียดเพิ่มเติมได้จาก [17]



รูปที่ 3 ตัวอย่างการพิจารณากราฟต่าง ๆ

5. การตรวจสอบความเหมาะสมของแบบจำลองค่าสุดขีด

ในทางปฏิบัติหากข้อมูลที่นำมาวิเคราะห์มาจากกระบวนการคงที่ ผู้วิเคราะห์สามารถประมาณค่าพารามิเตอร์และนำมาอธิบายความเป็นไปของตัวแปรสุ่มได้อย่างถูกต้อง แต่ถ้าหากข้อมูลที่นำมาวิเคราะห์มาจากกระบวนการที่มีการเปลี่ยนแปลงขึ้นอยู่กับเวลาแล้วนั้น การจะนำค่าพารามิเตอร์ไปใช้ย่อมเกิดข้อผิดพลาดขึ้นได้ ดังนั้นการหาแบบจำลองที่เหมาะสมและหาระดับการเกิดซ้ำของข้อมูลสุดขีดโดยใช้การแจกแจงค่าสุดขีดจึงเป็นอีกแนวทางในการบริหารจัดการและตัดสินใจของนักวิเคราะห์

เมื่อนักวิเคราะห์ได้แบบจำลองที่เหมาะสมกับข้อมูลมากกว่าหนึ่งตัวแบบ การตรวจสอบความเหมาะสมของแบบจำลองจึงมีความสำคัญอย่างมาก สำหรับวิธีการตรวจสอบวิธีหนึ่งซึ่งนักวิเคราะห์มักใช้การทดสอบตัวแบบ คือ การทดสอบอัตราส่วนควรจะเป็น (Likelihood Ratio Test) โดยสถิติทดสอบดังกล่าวมีการแจกแจงไคกำลังสอง (Chi-square Distribution) ด้วยองศาอิสระ ν (Degree of Freedom: ν) เมื่อ ν เท่ากับผลต่างของจำนวนพารามิเตอร์ของตัวแบบที่ต้องการนำมาเปรียบเทียบ

เมื่อได้ตัวแบบที่ดีที่สุดสำหรับข้อมูลแล้ว ขั้นตอนถัดมาคือ การทดสอบภาวะสารูปสนิทดี (The Goodness of Fit) ของแบบจำลองที่ดีที่สุด สามารถพิจารณาจากกราฟ (Diagnostic Plots) ดังแสดงในรูปที่ 3 ซึ่งประกอบด้วย กราฟควอนไทล์ (Quantile Plot) กราฟความน่าจะเป็น (Probability Plot) กราฟความหนาแน่น (Density Plot) และกราฟระดับการเกิดซ้ำ (Return Level Plot) หรือพิจารณาจากสถิติทดสอบสารูปสนิทดีด้วยสถิติทดสอบโคโมโกรอฟสเมอรนอฟ (Kolmogorov-smirnov Test) หรือสถิติทดสอบไคกำลังสอง (Chi-square Test) ได้เช่นกัน

6. คาบเวลาการเกิดซ้ำ (Return Period) และการหาระดับการเกิดซ้ำ (Return Level)

เมื่อนักวิเคราะห์ได้แบบจำลองที่เหมาะสมที่สุดของข้อมูลแล้ว สิ่งที่ต้องวิเคราะห์ต่อมาคือ การหาระดับการเกิดซ้ำของค่าสุดขีด โดยกำหนดให้ x_T แทนระดับค่าข้อมูลที่สูงกว่าค่าเฉลี่ยในรอบ T ปี (คาบเวลา) หมายถึงการวิเคราะห์ความถี่ของการเกิดเหตุการณ์ค่าสุดขีดในรูปของความน่าจะเป็น หรือโอกาสที่จะเกิดเหตุการณ์นั้น ๆ ซึ่งจะใช้ข้อมูลในอดีตมาวิเคราะห์ โดยใช้หลักการของทฤษฎีค่าสุดขีดเพื่อหาขนาดหรือ ปริมาณของค่าสุดขีด

ที่จะเกิดขึ้นที่คาบเวลาการเกิดซ้ำ เช่น ค่าสูงสุดหรือค่าต่ำสุดของปริมาณน้ำฝน ค่าความเร็วลมสูงสุด เป็นต้น

ความถี่ของการเกิดเหตุการณ์ค่าสุดขีดคำนวณจากความถี่สัมพัทธ์ของเหตุการณ์ เช่น ในจำนวนชุดข้อมูลค่าสุดขีดจำนวน N ค่า ถ้ามีจำนวนข้อมูลที่อยู่ในช่วงเกิดเหตุการณ์ใดๆ อยู่ n ค่า จะมีความถี่สัมพัทธ์ของเหตุการณ์นั้นเท่ากับ $\frac{n}{N}$ และถ้ามีจำนวนชุดข้อมูลที่วัดได้มากพอ ความถี่สัมพัทธ์จะมีค่าเข้าใกล้ความน่าจะเป็นของการเกิดเหตุการณ์ $P(E) = \lim_{n\to\infty} \frac{n}{N}$

ใน ค.ศ. 1958 Gumbel [18] ได้เสนอทฤษฎีการแจกแจงความถี่ในรูปสมการทั่วไปสำหรับโอกาสที่จะเกิดเหตุการณ์ X ที่มีค่าน้อยกว่าหรือเท่ากับ x หรือ $P(X \le x)$ นั่นคือ

$$\begin{aligned} F(X) &= P(X \le x) \\ &= 1 - P(X \ge x) \\ &= 1 - \frac{1}{T} \end{aligned}$$

เมื่อ T คือคาบเวลาหรือรอบปีการเกิดซ้ำ

ตัวอย่าง การหาระดับการเกิดซ้ำสำหรับการออกแบบโครงการทางวิศวกรรมแหล่งน้ำต่างๆ นิยมเรียกกันว่าระดับการเกิดซ้ำ x_T คือตำแหน่งของข้อมูล (Quantiles) เมื่อ p คือความน่าจะเป็นของเหตุการณ์ที่ $x > x_T$ โดยเฉลี่ย 1 ครั้งในทุกๆ T ปี ซึ่ง T คือรอบปีหรือคาบเวลาการเกิดซ้ำที่มีความสัมพันธ์กับ p โดยที่ $T = \frac{1}{p}$ และเขียนฟังก์ชันสะสมของระดับการเกิดซ้ำได้ว่า $F(x_T) = 1 - \frac{1}{T}$ จะเห็นได้ว่ารอบปีการเกิดซ้ำ T แท้จริงแล้วคือจำนวนรอบปีที่เกิดเหตุการณ์ภัยพิบัติ $x > x_T$ เกิดขึ้นโดยเฉลี่ย 1 ครั้งนั่นเอง

6.1 ระดับการเกิดซ้ำสำหรับการแจกแจงค่าสุดขีดวางนัยทั่วไป

สามารถคำนวณระดับการเกิดซ้ำ ณ คาบเวลา T ที่สนใจของ GEV $\left(x_T^{GEV}\right)$ ดังนี้

$$x_T^{GEV} = \mu - \frac{\sigma}{\xi}\left(1 - \left(-\log\left(1 - \frac{1}{T}\right)\right)^{-\xi}\right)$$

6.2 ระดับการเกิดซ้ำสำหรับการแจกแจงพาเรโตวางนัยทั่วไป

สำหรับ GPD ซึ่งมีพารามิเตอร์ σ และ ξ เมื่อมีข้อมูลที่มีค่าสูงกว่าค่า u นั่นแสดงว่า $X > u$ ซึ่งสามารถเขียนสมการทั่วไปสำหรับโอกาสที่จะเกิดเหตุการณ์ดังกล่าวได้ดังนี้

$$\Pr(X > u) = \zeta_u \left(1 + \xi\left(\frac{x-u}{\sigma}\right)\right)^{-1/\xi}$$

โดยกำหนดให้ $\zeta_u = \Pr(X > u)$ ดังนั้น ระดับการเกิดซ้ำหมายถึง ค่าเฉลี่ยของค่าที่สูงเกินกว่าค่า u ทุกๆ ค่าสังเกตณ คาบเวลา T ที่สนใจ นั่นคือ

$$\zeta_u \left(1 + \xi\left(\frac{x-u}{\sigma}\right)\right)^{-1/\xi} = \frac{1}{T}$$

และสามารถจัดรูปสมการระดับการเกิดซ้ำสำหรับการแจกแจงพาเรโตวางนัยทั่วไป $\left(x_T^{GPD}\right)$ ได้ดังนี้

$$x_T^{GPD} = u + \frac{\sigma}{\xi}\left(\left(T\zeta_u\right)^{\xi} - 1\right), \text{ ถ้า } \xi \ne 0$$

สำหรับ $T = N \times n_y$ เมื่อ n_y คือจำนวนค่าสังเกตต่อปี และ N เป็นจำนวนปี

7. สรุป

จากที่ผู้เขียนได้กล่าวถึงสถิติของค่าสุดขีดโดยได้กล่าวถึงทฤษฎีค่าสุดขีด ผ่านการแจกแจงของค่าสุดขีดทั้งการแจกแจง GEV และ GPD และการเลือกค่าสุดขีดที่ถูกนำมาวิเคราะห์ในการแจกแจงทั้งสองแบบ รวมทั้งการหาระดับการเกิดซ้ำสำหรับการแจกแจง GEV และ GPD ซึ่งเป็นหัวใจสำคัญในการศึกษาแบบจำลองค่าสุดขีด อย่างไรก็ตามยังมีรายละเอียดของทฤษฎีค่าสุดขีดอีกหลายประเด็นที่ผู้เขียนไม่ได้กล่าวในบทความนี้ อาทิ เช่น วิธีการแก้ปัญหาการเลือกค่าเกณฑ์ที่เหมาะสมสำหรับการวิเคราะห์ GPD การประมาณค่าพารามิเตอร์ด้วยวิธีของเบย์ (Bayesian Methods) และการเปรียบเทียบ

แบบจำลองต่าง ๆ เพื่อหาแบบจำลองที่เหมาะสมสำหรับข้อมูลที่นำมาวิเคราะห์ด้วยเกณฑ์สารสนเทศของ Akaike (Akaike's Information Criterion: AIC) และเกณฑ์สารสนเทศเบส์ (Bayesian Information Criterion: BIC) เป็นต้น

ในปัจจุบันมีนักวิจัยพัฒนาทฤษฎีค่าสุดขีดอย่างต่อเนื่อง พร้อมทั้งเปรียบเทียบแบบจำลองของการแจกแจงค่าสุดขีดกับการแจกแจงแบบมีหางที่หนักอื่น ๆ เช่น การแจกแจงแคปปา (Kappa Distribution) และการแจกแจงเวคบีย์ (Wakeby Distribution) ซึ่งมีพารามิเตอร์ที่กำกับการเกิดขึ้น 4 และ 5 พารามิเตอร์ ตามลำดับ อย่างไรก็ตามมีงานวิจัยหลายเรื่องได้ข้อสรุปว่าแบบจำลองที่เหมาะสมที่สุดสำหรับข้อมูลที่มีค่าสุดขีดรวมอยู่ โดยเฉพาะข้อมูลด้านอุตุนิยมวิทยาและอุทกวิทยา คือแบบจำลองของการแจกแจงค่าสุดขีด (รายละเอียดเพิ่มเติม [1], [6], [7], [9]) ดังนั้นจึงเป็นประเด็นที่น่าสนใจสำหรับนักวิเคราะห์ที่จะทราบเกี่ยวกับสถิติค่าสุดขีดเพื่อเป็นความรู้เบื้องต้นในการประยุกต์ใช้ทฤษฎีต่าง ๆ เพื่อวิเคราะห์ข้อมูลได้อย่างมีประสิทธิภาพและถูกต้อง

เอกสารอ้างอิง

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

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

[3] B. Erik and H. Rootzen, "Univariate and bivariate GPD methods for predicting extreme wind storm losses," *Insurance Mathematics and Economics*, vol. 44, pp. 345-356, 2009.

[4] B. Finkenstadt and H. Rootzen, *Extreme values in finance, telecommunications, and the environment*, London: Chapman and Hall/CRC Press, 2004.

[5] J. Beirlant et. al, *Statistics of Extremes: Theory and Applications*, New York: John Wiley & Sons, 2004.

[6] 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.

[7] R.W. Katz, M.B. Parlange, and P. Naveau, "Statistics of extremes in hydrology," *Advanced Water Resources*. vol. 25, pp. 1287-1304, 2002.

[8] S. Nadarajah and D. Choi, "Maximum daily rainfall in South Korea," *Journal of Earth System Science*. vol. 116, no. 4, pp. 311-320, 2007.

[9] L. Rajaram, *Statistical Models in Environmental and Life Sciences*, Florida: University of South Florida, 2006.

[10] V. Storch and F.W. Awiers, *Statistical analysis in climate research*, Cambridge: Cambridge University Press, 2001.

[11] M.R. Leadbetter, G. Lindgren, and H. Rootzen, *Extremes and related properties of randorn sequences and processes*, New York: Springer Verlag, 1983.

[12] S. Coles, *An introduction to statistical modeling of extreme values*, London: Springer Verlag, 2001.

[13] B.B. Brabson and J.P. Palutikof, "Tests of the generalized Pareto distribution for predicting extreme wind speeds," *Journal of Applied Meteorology*, vol. 39, pp. 1627-1640, 1999.

[14] R.D. Reiss and M. Thomas, *Statistical analysis of extreme values: with applications to insurance, finance, hydrology, and other fields*, Basel: Birkhauser, 2001.

[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, pp. 158-171, 1955.

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

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

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