การวิเคราะห์ค่าความคลาดเคลื่อนของการพยากรณ์กำลังผลิตไฟฟ้าจากพลังงานแสงอาทิตย์ของโรงไฟฟ้าในประเทศไทย

Error Analysis of Solar Power Forecasting in Thailand's Solar Farms

ผู้ช่วยศาสตราจารย์ ดร.จิตโกมุท ส่งศิริ

Smart Grid Research Unit (SRGU)

ภาควิชาวิศวกรรมไฟฟ้า คณะวิศวกรรมศาสตร์

จุฬาลงกรณ์มหาวิทยาลัย

August 25, 2020

สารบัญ

สารบัญรูป

สารบัญตาราง

บทที่ 1

บทนำ

ค่าพยากรณ์ที่ได้จากแบบจำลองพยากรณ์หนึ่งๆ นั้น เป็นที่ทราบกันดีว่า จะเป็นฟังก์ชันของค่าวัดของตัวแปรนั้นของเวลาในอดีต และเนื่องจากเรามักจะมีสมมุติฐานว่าค่าวัดของตัวแปรที่สนใจใดๆ นั้นมีความไม่แน่นอนอยู่ เราจึงสรุปได้ว่า ค่าพยากรณ์นั้นก็ถือเป็นตัวแปรสุ่มหนึ่งที่มีความไม่แน่นอนเช่นกัน หากจำแนกวิธีตามแนวทางที่จัดการกับความไม่แน่นอนของค่าพยากรณ์นั้น เราจึงสามารถแบ่งออกได้เป็น

1. deterministic forecasting (การพยากรณ์เชิงกำหนด) เป็นการให้ค่าพยากรณ์ที่เป็นค่าหนึ่ง และไม่มีข้อมูลเพิ่มเติมเกี่ยวกับความไม่นอนของค่าพยากรณ์นั้นๆ การพยากรณ์นี้เป็นแนวทางที่ปฏิบัติกันมาแต่เดิม เช่น การใช้ statistical models, การใช้ machine learning tools เช่น Neural network, Support vector regression เป็นต้น

2. probabilistic forecasting (การพยากรณ์เชิงน่าจะเป็น) เป็นการให้ค่าพยากรณ์และคุณสมบัติเชิงสถิติของค่าพยากรณ์นั้น วิธีนี้ยังสามารถแบ่งออกได้เป็น 2 ประเภทคือ

   (a) การวิเคราะห์หาคุณสมบัติเชิงสถิติของค่าพยากรณ์จากวิธี deterministic: วิธีนี้หมายถึง การนำค่า deterministic forecasts มาวิเคราะห์ความไม่แน่นอนของค่าความผิดพลาดในการพยากรณ์ ว่ามีการกระจายตัวอย่างไร มีค่าพารามิเตอร์พื้นฐานทางสถิติ เช่น mean, variance, skewness, kurtosis อย่างไร หรือเป็นการวิเคราะห์ช่วงความเชื่อมั่นของค่าความคลาดเคลื่อน การวิเคราะห์ลักษณะนี้ จะใช้วิธี parametric หรือ non-parametric ทางสถิติ

   (b) การวิเคราะห์หาคุณสมบัติเชิงสถิติของค่าพยากรณ์จากวิธี probabilistic โดยตรง: วิธีนี้หมายถึง แบบจำลองการพยากรณ์ที่ใช้ จะคำนวณหรือวิเคราะห์คุณสมบัติเชิงสถิติออกมาพร้อมกัน เทคนิคที่พบใช้งานวิจัยที่ผ่านมา เช่น การใช้ quantile regression หรือ extreme learning machine (ELM) นอกจากนี้ งานวิจัยในอดีตของกลุ่มการพยากรณ์โหลดเชิงสถิติจะพบได้มากกว่า ดังเช่น review paper [HF16] หรือ การพยากรณ์ราคาไฟฟ้าเชิงสถิติใน review paper [NW18]

ในรายงานฉบับนี้ จะบรรยายงานวิจัยที่ผ่านมาเกี่ยวกับ *การวิเคราะห์หาคุณสมบัติเชิงสถิติของค่าพยากรณ์จากวิธี deterministic* เป็นหลัก

การวิเคราะห์ค่าพยากรณ์จะอยู่บนหลักของการประมาณการกระจายตัวของค่าพยากรณ์ หรืออาจจะเป็นการกระจายตัวของค่าความคลาดเคลื่อนของค่าพยากรณ์ จาก point forecasts ที่ได้รับมาจากแบบจำลองหนึ่งๆ พารามิเตอร์ที่จะใช้บ่งชี้คุณภาพของ deterministic forecasts มักจะใช้ ช่วงของค่าพยากรณ์ (prediction interval) ซึ่งเป็นพารามิเตอร์ที่ derived ได้ หากเราทราบ distribution ของค่าพยากรณ์ ในงานวิจัยที่ผ่านมา จึงสามารถแบ่งแนวทางออกได้เป็น 2 แนวทางหลัก คือ

การประมาณช่วงของค่าพยากรณ์โดยตรง. ในแนวทางนี้มีหลายวิธีที่พิจารณาในงานวิจัย ใน [LHHB09] มีสมมติฐานว่าค่าความผิดพลาดในการพยากรณ์นั้นมีการกระจายตัวแบบ Guassian ทำให้สามารถกำหนดช่วงความเชื่อมั่น (CI) ได้ในรูป

$$\hat{I}(t) \in (\hat{I}_{\text{point}}(t) \pm 2\sigma(\cos\theta(t), k(t))$$

ในที่นี้ [LHHB09] ได้วิเคราะห์ว่า σ ซึ่งคือ standard deviation ของค่าผิดพลาดในการพยากรณ์ ว่าขึ้นกับตัวแปรหลักที่สำคัญ คือ solar zenith angle $(\cos\theta(t))$ และดัชนีฟ้าใส $(k(t))$ ในรูปแบบของฟังก์ชันพหุนาม ดังนั้น งานวิจัยนี้จึงมีการพัฒนาแบบจำลองพหุนาม ของ σ ในตัวแปรสองตัวดังกล่าว ผลการวิเคราะห์ด้วยวิธีดังกล่าวกับ single station ได้แสดงว่าช่วงของ CI นั้นจะแคบในวันที่ฟ้าใส และช่วง CI จะกว้างมากขึ้นในวันที่มีเมฆ และพบว่าค่ากำลังไฟฟ้าที่ผลิตได้จริงก็ตกอยู่ในช่วง CI ด้วยความเชื่อมั่นที่ตั้งไว้

เราพบว่า การประมาณช่วงของค่าพยากรณ์ (PI) จากแบบจำลอง neural network มีบทความ review ใน [KNCA11, KNC13] ที่มีใช้หลักว่า total variance สามารถแบ่งได้เป็น $\sigma^2 = \sigma_{\hat{y}}^2 + \sigma_\epsilon^2$ โดยที่ total variance σ^2 คือ ความแปรปรวนที่เทียบระหว่างค่าพยากรณ์กับ target (ที่เป็นค่าวัด), $\sigma_{\hat{y}}^2$ คือความแปรปรวนของค่าพยากรณ์ที่เทียบกับ true regression mean (ของแบบจำลองนั้นๆ) ส่วน σ_ϵ^2 เป็นความแปรปรวนของ noise หรือในงานการพยากรณ์โหลด [SM00] ที่ใช้ neural network เป็นแบบจำลองการพยากรณ์ และมีสมมติฐานว่าหาก activation function เป็นฟังก์ชันเชิงเส้นแล้วนั้น ความคลาดเคลื่อนการพยากรณ์จะกระจายตัวแบบ t แล้วจึงคำนวณหา PI จากการกระจายตัวนั้น ในบทความวิจัยนี้ ทุกวิธีที่กล่าวถึง คือการประมาณ $\hat{y}$ และการประมาณ $\sigma_{\hat{y}}^2$ แล้วนำค่าแปรปรวนไปคำนวณหา PI วิธีเหล่านี้ได้แก่ Delta method, Bayesian method, MVE method, Lower upper bound estimation method และ Bootstrap method (ที่ใช้ ensemble of NNs) นอกจากนี้ยังมีวิธี Hybrid Intelligence Algorithm (HIA) เป็นวิธีที่นำมาเสนอใน [WXP+14] และนำมาเปรียบเทียบในงานของ [GPG16] ที่ใช้ ELM (Extreme learning machine) เป็น algorithm ในการ train Neural network ที่มี target เป็น lower bound และ upper bound ของ PI ที่ค่า nominal coverage ต่างๆ

วิธีกลุ่มใหญ่ที่ใช้คำนวณ PI คือ Bootstrap-based method หรือ resampling method เช่น [WXP+14] ในการพยากรณ์ลม หรือใน [ABCP16, DPP16, FH11] ที่เป็นการพยากรณ์โหลด สิ่งที่เสริมใน [ABCP16] คือการพิจารณา family-wise error (FWE) จากการพยากรณ์หลาย horizon ตัวอย่างคือเช่น ผู้พยากรณ์อาจจะต้องการควบคุมเหตุการณ์ที่มีความคลาดเคลื่อนติดกันเป็นจำนวน k จุด ในงาน [FH11] ได้พิจารณาการใช้ block bootstrap กล่าวคือ คำนวณ forecasting residuals จากข้อมูลในช่วงเวลาใกล้กัน ด้วยเหตุผลที่ว่าแบบจำลองพยากรณ์ในรายครึ่งชั่วโมง จำนวน 48 แบบจำลองนั้น จะมี correlation กัน

การประมาณ distribution ของค่าคลาดเคลื่อนของการพยากรณ์. ในแนวทางนี้ จะประมาณ distribution ของความคลาดเคลื่อนก่อน หลังจากนั้น พารามิเตอร์เช่น prediction interval หรือ probability intervals ใดๆ ก็จะเป็น by-product parameter ตัวอย่างแนวทางนี้พบได้จากงานวิจัยของการพยากรณ์พลังงานลม ดังเช่นใน [HM11, BDNL08] ซึ่งต้องวิเคราะห์พารามิเตอร์เชิงสถิติ เช่น kurtosis ของค่าความผิดพลาดการพยากรณ์ ซึ่งพบว่ามีค่าสูงในช่วง $3 < \kappa < 6$ และได้แสดงให้เห็นว่าไม่เหมาะสมที่จะใช้ Guassian ในการประมาณการกระจายตัว ในงาน [BDNL08] เป็นการวิเคราะห์ wind power error ที่ได้เลือก Beta distribution แทน จากนั้นใช้หลักการประมาณฟังก์ชันการกระจายตัวเพื่อประมาณค่าพารามิเตอร์ของฟังก์ชันดังกล่าว ซึ่งทำให้หาช่วง quantiles ได้จากฟังก์ชันการกระจายตัวของค่าความผิดพลาด ทั้งนี้ horizon ของการพยากรณ์ที่ต่างกัน ทำให้จำนวนค่าพยากรณ์ ซึ่งจะเป็นจำนวน samples ที่มาใช้ในการประมาณหา pdf นั้น ไม่เท่ากัน จึงมีการประมาณข้อมูลใน Power bin resolution ตามต้องการด้วย สำหรับ [HM11] เป็นการวิเคราะห์ wind power error เช่นกัน ได้คำนวณค่า kurtosis และ skewness ที่เป็นพารามิเตอร์เชิงสถิติเพื่อดู shape ในช่วง peak และความสมมาตรของ histogram จากนั้นได้พิจารณาเลือก Cauchy, Weibull, Beta distributions ในการ fit distribution ซึ่งผลพบว่า cauchy function นั้นให้ค่า fit loglikelihood ที่สูงที่สุด การ fit distribution นั้นใช้คำสั่ง `fitdist` จาก software มาตรฐานและวิเคราะห์หา prediction interval ประกอบ

วิธีการประมาณ distribution แบบ non-parametric ยังพบได้ในงาน [GPG16] สำหรับพลังงานแสงอาทิตย์ คือการใช้ kernel density estimation (หรือที่เรียกอีกเทอมว่า Parzen–Rosenblatt window method) ซึ่งพบว่าการกระจายตัวแบบ Gaussian และ beta นั้น ไม่เหมาะสมกับค่าความคลาดเคลื่อนจากการพยากรณ์พลังงานแสงอาทิตย์ ที่สรุปผลจากการทำ one-sample Kolmogorov-Smirnov (KS) test นอกจากนี้ การกระจายตัวแบบ Rayleigh, Beta, Gamma, Weibull ไม่ได้นำมาพิจารณาเนื่องจากตัวแปรสุ่มเหล่านี้ต้องมีค่าเป็นบวกเท่านั้น ซึ่งไม่เหมือนกับค่าความคลาดเคลื่อนพยากรณ์ที่อาจเป็นได้ทั้งบวกและลบ

กล่าวโดยสรุป จากงานวิจัยที่ผ่านมาดังข้างต้นนั้น เป็นทั้งงานวิจัยที่ศึกษาในการพยากรณ์พลังงานแสงอาทิตย์ หรือพลังงานลม แนวทางที่จะมาวิเคราะห์ความคลาดเคลื่อนการพยากรณ์ของแบบจำลองที่ Provider ใช้ จะทดลองจากวิธีที่มีอยู่ในงานวิจัยดังกล่าว อันได้แก่

1. การประมาณค่า quantiles จากวิธี bootstrap
2. การประมาณหา distribution ที่เหมาะสมกับความคลาดเคลื่อนการพยากรณ์พลังงานแสงอาทิตย์ ที่จะใช้ทั้งวิธี non-parametric

(ที่จะเรียกว่า kernel density estimation) และวิธี parametric (อันได้แก่ การเลือกหา known distribution เช่น Gaussian, t-location scale, Stable และอื่นๆ ว่าการกระจายตัวแบบใดจะเหมาะสม)

ทั้งสองแนวทางนั้น จะทำให้เราคำนวณ prediction interval จากข้อมูลได้ สำหรับในรายงานฉบับนี้ จะอธิบายและแสดงผลการทดลองที่มาจากวิธีการประมาณฟังก์ชันการกระจายตัว (distribution function) ด้วยวิธี kernel density estimation และใช้ distribution ที่มีรูปแบบต่างๆ และใช้วิธี bootstrap ที่ไม่อิงกับ รูปแบบของ distribution เลย

บทที่ 2

หลักการทางสถิติที่เกี่ยวข้อง

หากใช้วิธีการพยากรณ์เชิงความน่าจะเป็น หรือมีการวิเคราะห์เชิงสถิติจากผลพยากรณ์ที่ได้นั้น (ซึ่งอาจจะมาจาก point forecasts ก็ได้) การแสดงคุณสมบัติเชิงสถิติของการพยากรณ์ดังกล่าว เรากำหนดให้ $F(y)$ คือ cumulative density distribution ของตัวแปร y (ซึ่งจะใช้ y เป็นค่าตัวแปรที่สนใจ เช่น ในที่นี้ คือ solar power หรือ solar irradiance) และกำหนดให้ $f(y)$ คือ probability density distribution ของตัวแปร y ด้วยการกำหนดดังกล่าว เราสามารถนิยาม **quantiles** ของตัวแปรสุ่ม y ที่ค่าความน่าจะเป็น α นั้น ว่าเป็น

$$q(\alpha) = F^{-1}(\alpha)$$

หรือนั่นคือ quantile ที่ค่าความน่าจะเป็น α คือ ค่าของ y ที่ทำให้ $F(y) = P(y \leq q(\alpha)) = \alpha$ โดยที่ $\alpha \in [0, 1]$ ค่าความน่าจะเป็น α นั้น ในหลายครั้งจะใช้คำว่า coverage rate ในทางปฏิบัติ เนื่องจากเราไม่ทราบ distribution ที่แท้จริงของตัวแปร y หนึ่งๆ ที่สนใจ (เช่น ไม่ทราบว่า พลังงานลม มีการกระจายตัวแบบ Gaussian หรือ Gamma หรือฟังก์ชันอื่นๆ) เราจะใช้สัญลักษณ์ $\hat{f}$ เพื่อแทนฟังก์ชันประมาณของ f การประมาณดังกล่าวนั้น อาจจะอยู่ในรูปแบบ closed-form expression เช่น $\hat{f}(y)$ เป็นฟังก์ชันกระจายตัวแบบ Gaussian ก็จะมีฟังก์ชันในรูป $\hat{f}(y) = \frac{1}{\sqrt{2\pi\sigma}} e^{-(y-\mu)^2/2\sigma^2}$ หรือ $\hat{f}$ อาจจะอยู่ในรูปของ collections of quantiles นั่นคือ ช่วงของ y ที่เป็นไปได้ จะถูกแบ่งเป็นจุดย่อย y_k หลายจุด และ $\hat{f}$ คือการประมาณค่า density function ที่จุด y_k นั้นๆ ในกรณีนี้ เราจึงใช้สัญลักษณ์ในเชิงเซตกับ $\hat{f}$

$$\hat{f} = \{q(\alpha_i) \mid 0 \leq \alpha_1 \leq \alpha_2 \leq \cdots \leq 1\} \tag{2.1}$$

การทราบค่าประมาณของ quantiles ที่ค่าความน่าจะเป็นต่างๆ นั้น จึงเป็นเรื่องเดียวกันกับการหา $\hat{f}$ (ไม่ว่าจะได้มาด้วยวิธีใดในการประมาณ) ทำให้เรานำไป quantiles ต่างๆ หรือรูปแบบฟังก์ชันของ $\hat{f}$ นำไปคำนวณปริมาณทางสถิติใดๆ ที่สนใจได้ ตัวอย่างเช่น สมมติให้ $\hat{y}$ คือ ความคลาดเคลื่อนของการพยากรณ์ในหน่วย p.u. เราอาจจะมีคำถามว่า

- ด้วยวิธีการพยากรณ์ที่เราใช้อยู่นั้น จะทำให้ error มีค่าสูงมากกว่า 0.8 p.u. ด้วยความน่าจะเป็นเท่าใด นั่นคือ เราต้องคำนวณ $P(y \geq 0.8) \approx 1 - \hat{F}(0.8)$
- ค่าความคลาดเคลื่อนค่าใด ที่มีโอกาสเกิดสูงที่สุด นั่นคือ เราต้องการหา mode ซึ่งคำนวณได้จากค่าในแกน y ที่ทำให้เกิด peak ใน $\hat{f}(y)$ (note: ค่า mode ของการกระจายตัวแบบ Gaussian คือค่าเดียวกันกับค่า mean แต่ผลลัพธ์นี้ อาจจะไม่จริงสำหรับการกระจายตัวชนิดอื่นๆ ในกรณีทั่วไป)

ปริมาณทางสถิติที่สนใจดังตัวอย่างข้างต้นนั้น อาจจะคำนวณได้โดยตรงจากข้อมูล (โดยไม่ต้องหา $\hat{f}$ มาก่อน) เช่น ต้องการทราบค่าความแปรปรวนของความคลาดเคลื่อน เราก็สามารถคำนวณ sample variance จากข้อมูลได้โดยตรง แต่การประมาณ $\hat{f}$ มาก่อน ก็ทำให้เราสามารถตั้งคำถามที่เกี่ยวกับประมาณทางสถิติที่สนใจได้หลายคำถาม พร้อมทั้งคำนวณปริมาณทางสถิติดังกล่าว ที่เป็นผลพลอยได้จาก $\hat{f}$ ได้หลายปริมาณพร้อมกัน หนึ่งในปริมาณทางสถิติที่มักจะนำมาพิจารณา คือ **ช่วงการทำนาย (prediction interval)** หรือ PI เราจะกล่าวว่า $[l, u]$ คือช่วงการทำนายของตัวแปรสุ่ม y ที่มี nominal coverage เป็น γ เมื่อ $P(l \leq y \leq u) = \gamma$ การหาช่วงดังกล่าวนั้น หากเราทราบ distribution และพารามิเตอร์ของ distribution นั้น (เช่น ทราบว่าเป็น Gaussian ที่มีค่าเฉลี่ยเป็น μ และความแปรปรวน σ^2) การหาช่วงดังกล่าว ก็ทำได้จากการใช้ quantile function

ของการกระจายตัวแบบนั้น เช่น $[l, u] = [\mu - 1.96\sigma, \mu + 1.96\sigma]$ คือช่วงการทำนายของตัวแปร Gaussian ที่จะมีความน่าจะเป็นของ y ที่ตกในช่วงนี้เท่ากับ 0.95 ที่คำนวณจาก $-1.96 = \Phi^{-1}(0.05/2)$ โดย Φ คือ cumulative density function of Gaussian และใช้ fact ที่ว่าการกระจายตัว Gaussian นั้น symmetric รอบ μ สำหรับการหา PI ในกรณีที่เราไม่ทราบ distribution หรือถึงแม้ว่าจะทราบ distribution แต่ไม่ทราบ parameters ของ distribution นั้น เราสามารถอธิบาย PI ในรูปของ quantile function ได้ดังนี้

$$\text{PI} = \text{Prediction Interval} = [q(\alpha_1), q(\alpha_2)], \quad \alpha_2 - \alpha_1 = \gamma \tag{2.2}$$

โดยที่ $q(\alpha)$ ใน (2.2) นั้นอาจจะประมาณได้มาจาก

- quantile function ที่มาจาก distribution ของ y และได้ค่าพารามิเตอร์มาจากการประมาณ (เช่น ใช้ $q(\alpha)$ เป็น $\Phi^{-1}(\alpha)$ เมื่อทราบว่าเป็น Gaussian แต่ว่า ค่า (μ, σ^2) ที่แท้จริงไม่ทราบ จึงใช้ sample mean และ sample variance ที่คำนวณมาจาก samples ใน y แทน) หรือ
- การประมาณ quantile function ด้วยวิธี non-parametric เช่น วิธี bootstrap ที่จะอธิบายต่อไปในหลักการและข้อจำกัดของวิธีดังกล่าว

โดยช่วงการทำนายดังใน (2.2) สำหรับ symmetric distribution จะมีจุดกลางที่ค่าเฉลี่ย แต่สำหรับ nonsymmetric distribution ใดๆ ช่วงการทำนายจะมีจุดกลางที่ค่า median เราจะกล่าวว่าช่วงการทำนาย (หรือที่เราจะใช้ในบริบทการพยากรณ์แสงอาทิตย์ว่า คือ ช่วงของการพยากรณ์) จะเป็นพารามิเตอร์พื้นฐานที่สำคัญ ในการให้ข้อมูลเชิงสถิติเพิ่มเติมของค่าพยากรณ์

2.1 ปัจจัยที่ส่งผลต่อการวิเคราะห์การพยากรณ์

ในการวิเคราะห์ PI ของค่าพยากรณ์ เราควรมีสมมติฐานหรือทราบก่อนว่า การพยากรณ์ที่ใช้พิจารณานั้น มีปัจจัยใดจะส่งผลถึงสมบัติการกระจายตัวของค่าพยากรณ์ โดยทั่วไปนั้น จะมีปัจจัยดังนี้

1. กลุ่มของตัวแปรขาเข้าของแบบจำลองพยากรณ์ (feature inputs): โดยทั่วไป feature ดังกล่าว อาจจะประกอบไปด้วยค่าวัดของตัวแปรสภาพภูมิอากาศ หรือ เป็นค่าพยากรณ์ของตัวแปรสภาพอากาศจากวิธีอื่น ค่าเหล่านี้ ต่างก็มีความไม่แน่นอนในตัวเอง และมีการกระจายตัวที่ต่างกันไป

2. แบบจำลองการพยากรณ์: แบบจำลองหนึ่งๆ สามารถอธิบายในรูปคณิตศาสตร์ได้ว่า เป็นรูปแบบความสัมพันธ์ระหว่าง input และ output ด้วยสมการคณิตศาสตร์หนึ่งๆ $y = g(x)$ ดังนั้น หาก x เป็นตัวแปรสุ่มที่มีการกระจายตัวเป็น f_x ในตัวแปร y นั้นก็ย่อมเป็นตัวแปรสุ่มเช่นเดียวกัน แต่การกระจายตัวของ y จะขึ้นอยู่กับ f_x และรูปแบบของฟังก์ชัน g

3. เวลาของค่าพยากรณ์ที่พิจารณา: เป็นที่ทราบกันดีว่า ตัวแปรภูมิอากาศมีความเกี่ยวข้องกับพลังงานแสงอาทิตย์ และตัวแปรเหล่านี้มีความไม่แน่นอนที่อาจจะเปลี่ยนไปตามเวลา (เช่น อุณหภูมิ มีค่าเฉลี่ยสูงขึ้น เมื่อหลายปีผ่านไป) ในทางสถิติ สมบัติเหล่านี้ เรียกว่า non-stationarity ของ process นั่นหมายความว่า การวิเคราะห์การกระจายตัวของค่าพยากรณ์ในปีนี้กับอีก 10 ปีข้างหน้า อาจจะไม่เหมือนกัน ในอีกนัยหนึ่งของเวลา เราพบว่า ความเข้มแสงอาทิตย์ในช่วงเช้า 6:00 นั้น มีการกระจายตัวที่ต่างกันกับเวลาเที่ยง 12:00 เพราะตอนเช้ามีแดดน้อย ความเข้มแสงเกือบเป็นศูนย์ในเกือบทุกวัน จึงมีความแปรปรวนต่ำ แต่ช่วงเที่ยง ค่าความเข้มแสงจะต่างกันไปตามสภาพอากาศในแต่ละวัน ดังนั้น ในบริบทนี้ เราจึงพิจารณาว่า การวิเคราะห์การพยากรณ์ควรทำแยกกันในแต่ละเวลา

4. lead time ของการพยากรณ์: เมื่อกำหนดให้ k ใน k-step ahead prediction คือ lead time เราพบว่า เมื่อ k ยิ่งมาก โดยทั่วไปนั้น การพยากรณ์จะทำได้แย่ลง (ข้อสรุปนี้ ขึ้นกับแบบจำลองการพยากรณ์ที่ใช้) ยกตัวอย่างเช่น เมื่อใช้ time series model การหา $\hat{y}(t+k|t)$ (การพยากรณ์ค่าที่เวลา $t+k$ โดยใช้ข้อมูลจากในอดีตถึงเวลา t) จะต้องใช้การประมาณ error of prediction จาก step $k-1, k-2, \ldots, 1$ มาคำนวณ ดังนั้น การสะสมของ error จากหลาย steps จะมีมากกว่า หากเราจะคำนวณค่า $\hat{y}(t+1|t)$ ผลของ lead time อาจจะอธิบายในมุมของแบบจำลองที่ต่างกัน

ได้ เช่น การใช้แบบจำลองกลุ่ม neural networks (NN) การพยากรณ์แบบ 1 ชั่วโมงล่วงหน้า (horizon เท่ากับ 1 ชั่วโมง) คือการ set NN target เป็นค่า scalar ที่ค่ากำลังไฟฟ้าที่ $t+1$ ส่วนการพยากรณ์แบบ 4 ชั่วโมงล่วงหน้า (หมายถึงมี horizon เป็น 4 ชั่วโมงล่วงหน้า และทำรายชั่วโมง) คือการ set NN target เป็น $(y(t+1), y(t+2), y(t+3))$ ซึ่งเป็น target vector ด้วยการทำเช่นนี้ เราจะเห็นว่า descriptions ของแบบจำลองจะไม่เหมือนกัน (feature inputs ที่ใช้อาจจะเหมือนกัน แต่ตอนฝึกสอนหรือเทรนแบบจำลองก็จะให้ผลการเทรนที่ไม่เหมือนกัน) เพื่อ accommodate การพยากรณ์ค่าที่ lead time ต่างกัน ดังนั้น กล่าวโดยสรุป lead time ที่ต่างกัน ก็จะส่งผลต่อการกระจายตัวของค่าพยากรณ์

การวิเคราะห์ค่าพยากรณ์เชิงสถิติจึงควรวิเคราะห์แยกกันตามปัจจัยดังข้างต้น กล่าวคือ วิเคราะห์การกระจายตัวของค่าพยากรณ์เมื่อปัจจัย A เปลี่ยนไปตามค่าต่างๆ และปัจจัยอื่นที่เหลือนั้น ถูกกำหนดให้คงที่ ในงานวิจัยที่ผ่านมา จึงพบได้ว่า การประมาณ $\hat{f}$ อาจจะตามมาด้วยสัญลักษณ์ subscript $\hat{f}_{t+k|t}$ เพื่อบ่งชี้ว่า เป็นการประมาณการกระจายตัวเมื่อ lead time เท่ากับ k หรือใช้เป็นสัญลักษณ์ $\hat{f}_{\mathrm{ANN}}$ เพื่อบ่งชี้ว่า เป็นการประมาณการกระจายตัวเมื่อใช้แบบจำลอง ANN เป็นต้น

สำหรับผลการวิเคราะห์ของ Provider นั้น เมื่อพิจารณา description แบบจำลองแล้ว (รายละเอียดใน section 3) เราจะวิเคราะห์ตามปัจจัยของ*เวลาค่าพยากรณ์* เนื่องจาก Provider ใช้แบบจำลองพยากรณ์ที่แยกกันในแต่ละเวลา

2.2 การประมาณ probability density function

กำหนดให้ Y เป็นตัวแปรสุ่มที่มีการกระจายตัวด้วย density function f และ cumulative density function F และกำหนดให้ $\mathbf{y} = (y_1, y_2, \ldots, y_N)$ เป็น samples ขนาด N ของตัวแปรสุ่ม Y เราจะนิยาม $\hat{F}(y)$ ว่าเป็น **empirical distribution function** ของ F ในรูปดังนี้

$$\hat{F}(y) = \frac{1}{N}\sum_{k=1}^{N} I\{y_k \leq y\}$$

โดยที่ $I\{y \in C\}$ เป็น indicator function ที่มีค่าฟังก์ชัน นิยามดังนี้

$$I\{y \in C\} = \begin{cases} 1, & y \in C \\ 0, & y \notin C \end{cases}$$

กล่าวคือ $\hat{F}$ ได้จากการนับจำนวน samples y_k's ที่มีค่าน้อยกว่า y เราจะเห็นว่า $\hat{F}$ นั้นขึ้นกับ N และถือเป็นค่าประมาณชนิดหนึ่งของ F

ปัญหาที่เราสนใจ คือการประมาณ f (หรือประมาณ F) ด้วยค่าตัวอย่างของ Y เหล่านั้น ในหลักการแล้ว การประมาณดังกล่าว สามารถทำได้หลายวิธี อันได้แก่ วิธีเชิง parametric ที่มีสมมติฐานของฟังก์ชันการกระจายตัว และประมาณค่าพารามิเตอร์ในฟังก์ชันนั้น หรือ วิธีเชิง non-parametric ที่ประมาณ cdf ด้วยค่า empirical จากข้อมูล หรือวิธี kernel density estimation ที่แบ่งช่วงข้อมูลเป็นช่วงย่อยๆ และประมาณ density function ในช่วงเล็กๆ นั้นด้วยฟังก์ชัน kernel ในเนื้อหาบทนี้ จะกล่าวถึงรายละเอียดของวิธีต่างๆ ดังที่กล่าวมา พร้อมตัวอย่างผลการประมาณเบื้องต้น เพื่อประกอบการอธิบาย

2.2.1 Maximum likelihood estimation

หากเรามีสมมติฐานว่า Y นั้นกระจายตัวด้วยฟังก์ชัน f ที่มีพารามิเตอร์เป็น θ เราสามารถเขียนฟังก์ชันลอการิทึมความเป็นไปได้ (log-likelihood function) ของ N samples ของ Y ได้เป็น

$$\log f(y_1, y_2, \ldots, y_N; \theta)$$

เราจะพบว่าฟังก์ชันลอการิทึมความเป็นไปได้นั้น หากมีค่าสูงจะแสดงถึงว่า ค่าพารามิเตอร์ θ ที่ใช้นั้นทำให้ f เหมาะสมหรือสอดคล้องกับค่าตัวอย่าง N ชุด มากเท่านั้น ดังนั้น หลักการของการประมาณค่าความเป็นไปได้สูงสุด (maxmimum likelihood estimation หรือ MLE) จึงกล่าวไว้ว่า ค่าพารามิเตอร์ θ ที่เหมาะสมสุดจะเลือกจาก

$$\hat{\theta}_{\mathrm{ml}} = \mathrm{argmax}_{\theta} \log f(y_1, y_2, \ldots, y_N; \theta) \tag{2.3}$$

การใช้ $\log(\cdot)$ มีเหตุผลว่าฟังก์ชันลอกริทึมเป็นฟังก์ชันเพิ่ม หาก f มีค่าสูง ค่า $\log(f)$ ก็มีค่าสูงตามไปด้วย และเหตุผลที่สำคัญ คือ ฟังก์ชันการกระจายตัวหลายแบบเป็น exponential family (เช่น Guassian, exponential, Rayleigh, และอื่นๆ) ดังนั้น การใช้ฟังก์ชันลอการิทึมจะทำให้ขั้นตอนทางคณิตศาสตร์นั้นยุ่งยากน้อยลง

เราสามารถแสดงให้เห็นได้ว่า เมื่อใช้วิธี MLE ประมาณพารามิเตอร์ของฟังก์ชันการกระจายตัวหลายฟังก์ชันนั้น ผลตอบสามารถแสดงให้อยู่ในรูปแบบปิดได้ (closed-form) ว่าขึ้นกับค่าตัวอย่าง $\{y_1, y_2, \ldots, y_N\}$ อย่างไร ตัวอย่างเช่น เมื่อ $Y \sim \mathcal{N}(\mu, \sigma^2)$ จะมีตัวประมาณ $\hat{\mu}_{\mathrm{ml}}$ เป็น sample mean และ $\hat{\sigma}^2_{\mathrm{ml}}$ เป็น sample variance ที่ normalized ด้วย N หรือ ตัวอย่างเช่น $Y \sim \mathrm{Poisson}(\lambda)$ จะมีตัวประมาณ $\hat{\lambda}_{\mathrm{ml}}$ เป็น sample mean เช่นกัน

เมื่อเราพิจารณาปัญหา MLE ใน (2.3) พบว่าเป็น unconstrained optimization แบบหนึ่ง ซึ่งเงื่อนไขค่าเหมาะสมที่สุด คือการกำหนดให้ gradient of log-likelihood function เมื่อเทียบกับ θ เท่ากับศูนย์ เงื่อนไขดังกล่าว เป็นที่ทราบกันดีว่า ในสำหรับบาง distribution อาจจะเป็นฟังก์ชันไม่เชิงเส้นของ θ ที่ไม่สามารถหาผลตอบในรูปแบบปิดได้ จึงจำเป็นต้องใช้เทคนิคด้าน optimization มาแก้หาผลตอบด้วย ตัวอย่างของฟังก์ชันการกระจายตัวในหมวดนี้ เช่น Gamma distribution ที่มี k และ θ เป็น shape และ scale parameters ตามลำดับ หรือ ฟังก์ชัน logistic เป็นต้น ปัญหา MLE นั้น ถือว่าเป็นปัญหาพื้นฐานในด้านทฤษฎีการประมาณ หลาย scientific softwares จึงมี library เพื่อหาคำตอบเชิงเลขสำหรับหลายฟังก์ชันการกระจายตัว ตัวอย่างใน MATLAB และ R คือการใช้คำสั่ง `fitdist` หรือ การเรียก `scipy stats` ใน python การใช้คำสั่งดังกล่าว ผู้ใช้ต้องมีสมมติฐานของฟังก์ชันการกระจายตัว (เช่น สมมติว่าข้อมูลมีการกระจายตัวแบบ Gamma) และมี sample data ส่วนผลลัพธ์ของคำสั่งคือ ให้ค่าประมาณพารามิเตอร์ของฟังก์ชันการกระจายตัวนั้น

ขั้นตอนการประมาณ distribution นั้น ประกอบไปด้วย

1. การเลือกการกระจายตัวในหมวดที่เหมาะสมกับตัวแปรที่สนใจ โดยทั่วไป ตัวแปรสุ่ม แบ่งได้เป็น ตัวแปรสุ่มแบบต่อเนื่อง (continuous) และแบบไม่ต่อเนื่อง (discrete) ในแต่ละฟังก์ชันการกระจายตัวก็มี support ของฟังก์ชัน (หรือ range of random variable) ที่ต่างกันไป ตัวอย่างเช่น Gaussian variables มีค่าที่เป็นไปได้ในช่วง $(-\infty, \infty)$ แต่ Gamma variables มีค่าในช่วง $(0, \infty)$ เป็นต้น ในขั้นตอนนี้ ผู้ใช้อาจจะมีตัวเลือกของ distributions อยู่เป็นจำนวนหนึ่ง

2. จากข้อ 1) หากผู้ใช้จะวิเคราะห์ต่อไปว่า distributions ใดจะเหมาะสมกับการกระจายตัวของข้อมูลที่มี หากเป็นวิธีพื้นฐาน เราอาจจะคำนวณค่าพารามิเตอร์เชิงสถิติพื้นฐานเช่น mean, variance, kurtosis, skewness เพื่อดูลักษณะรูปร่างความสมมาตร หรือ การลู่เข้าของหางการกระจายตัว นอกจากนี้ ยังมีวิธีที่ใช้ L-moment ซึ่งเป็นผลรวมเชิงเส้นของ order statistics ในการอธิบายรูปร่างของการกระจายตัว ในเบื้องต้น ไม่ว่าเราจะใช้วิธีใด สมมติว่าเราจะมี distributions in hypothesis อยู่ M distributions

3. การประมาณพารามิเตอร์ใน M distributions ด้วยเทคนิควิธีหนึ่งๆ ในชุดคำสั่ง fitdist นั้น โดยมากผลตอบจะใช้วิธี MLE ดังที่ได้อธิบายไป แต่สำหรับบางการกระจายตัว อาจจะใช้ค่า sample หรือใช้วิธีอื่นเช่น method of moment หรือ least-square estimator (ทั้งนี้ ต้องดูคำอธิบายในแต่ละการกระจายตัว)

4. การเปรียบเทียบฟังก์ชันการกระจายตัวที่ประมาณได้ ว่า distribution ใด ควรจะถูกเลือกมาอธิบายข้อมูลตัวอย่างที่เก็บมา ในขั้นตอนนี้ จะมีการทดสอบเชิงสถิติ เช่น Komogorov-Smirnov test ที่เป็นการทดสอบ ดังที่จะอธิบายใน หัวข้อที่ 2.2.4

ในเนื้อหาต่อไปนี้ จะกล่าวถึงสมบัติทางคณิตศาสตร์และสถิติของการกระจายตัว 4 แบบ อันได้แก่ logistic, location scale family, stable และ generalized extreme value เราจะแสดงให้เห็นในผลการทดลองต่อไปว่า การกระจายตัวดังกล่าวได้ถูกเลือกจากผลการทดลองว่า มีค่าความเป็นไปได้เข้ากันกับข้อมูลสูงที่สุด จึงเป็นสิ่งสำคัญที่จะศึกษาถึงสมบัติของการกระจายดังกล่าว

Logistic distribution. การกระจายตัวแบบ logistic มีฟังก์ชันหนาแน่นความน่าจะเป็น และ cumulative distribution function ในรูปทั่วไปคือ [JKB70]

$$\begin{aligned} f(x) &= \frac{e^{-(x-\mu)/\sigma}}{\sigma(1+e^{-(x-\mu)/\sigma})^2}, \quad -\infty < x < \infty \\ F(x) &= \frac{1}{1+e^{-(x-\mu)/\sigma}} \end{aligned}$$

โดยที่มี $\mu \in \mathsf{R}, \sigma > 0$ เป็น location และ scale parameters ตามลำดับ เราสามารถแสดงให้เห็นได้ว่า ทั้ง f และ F สามารถเขียนได้ในรูปของ hyperbolic secant และ hyperbolic tangent ดังนี้

$$f(x) = \frac{1}{4\sigma}\operatorname{sech}^2\left(\frac{x-\mu}{2\sigma}\right), \quad F(x) = \frac{1}{2} + \frac{1}{2}\tanh\left(\frac{x-\mu}{2\sigma}\right)$$

ดังนั้น logistic distribution จึงมีอีกชื่อเรียกว่าเป็น sech-squared distribution รูปร่าง S-shape ของ cdf เป็นลักษณะของการกระจายตัวแบบนี้ ซึ่งมีงานประยุกต์ในหลายด้าน เช่น logistic regression สำหรับปัญหา binary classification

Stable distribution [Nol18]. เราจะกล่าวว่า $X_1, X_2, \ldots, X_n$ เป็น i.i.d. stable random variables ถ้าหากว่าสำหรับทุกๆ n แล้วนั้น

$$X_1 + X_2 + \cdots + X_n \stackrel{d}{=} c_n X + d_n, \quad c_n > 0, d_n \in \mathsf{R}.$$

สำหรับนิยามที่สมมูลกัน เราจะกล่าวว่า X เป็นตัวแปรสุ่มแบบ stable in broad sense ถ้าหากว่าสำหรับ X_1 และ X_2 ที่เป็น independent copies ของ X และสำหรับ $a, b > 0$ นั้น

$$aX_1 + bX_2 \stackrel{d}{=} cX + d, \quad c > 0, d \in \mathsf{R} \tag{2.4}$$

โดยที่สัญลักษณ์ $\stackrel{d}{=}$ นั้นหมายถึงว่าตัวแปรทางซ้ายมือและขวามือของการเท่ากันใน distribution นั้นเป็นไปตาม probability law ที่เหมือนกัน เช่น รูปร่างของ distribution ของ X นั้นไม่เปลี่ยนไป (up to scale and shift) ภายใต้การบวก ดังนั้นการกระจายตัวใดๆ ที่สอดคล้องกับเงื่อนไข (2.4) ต่างก็เป็น stable distribution

จะเห็นว่าจากนิยามดังกล่าว เราอาจจะยังไม่สามารถแจกแจง stable distribution ว่ามีพารามิเตอร์อย่างไรได้ ดังนั้น การแจกแจงตัวแปร stable จะใช้สัญลักษณ์ $S(\alpha, \beta, \gamma, \delta; k)$ โดยจะมี 4 พารามิเตอร์แรกที่เป็นค่าที่ต้องการประมาณ ส่วน k เป็นค่าที่กำหนด การแจกแจงด้วยสัญลักษณ์ดังกล่าว จะอ้างอิงกับนิยามตัวแปร stable (ที่สมมูลกันกับ (2.4)) และที่ใช้ใน MATLAB ดังนี้ [Nol18]

นิยาม 1. *ตัวแปรสุ่ม X เป็นตัวแปร stable $S(\alpha, \beta, \gamma, \delta; 0)$ ถ้า X มี characteristic function เป็น*

$$\Phi_X(t) = \mathbf{E}[e^{itX}] = \begin{cases} \exp\left(-\gamma^\alpha |t|^\alpha [1 + i\beta\tan(\pi\alpha/2)\mathbf{sign}(t)(|\gamma t|^{1-\alpha} - 1)] + i\delta t\right), & \alpha \neq 1 \\ \exp\left(-\gamma|t|[1 + \frac{i2\beta}{\pi}\mathbf{sign}(t)\log(\gamma|t|)] + i\delta t\right), & \alpha = 1 \end{cases} \tag{2.5}$$

สำหรับพารามิเตอร์ของตัวแปร stable นั้น มีความหมายและมีค่าในช่วง ดังตาราง

พารามิเตอร์	ความหมาย	ค่า
α	first shape parameter	$0 < \alpha \leq 2$
β	second shape parameter	$-1 \leq \beta \leq 1$
γ	scale parameter	$0 < \gamma < \infty$
δ	location parameter	$-\infty < \delta < \infty$

ทั้ง α, β เรียกว่า shape parameters เพราะสองค่านี้บ่งชี้ถึงรูปร่างของการกระจายตัว โดย α นั้นจะบ่งชี้ถึงหางการกระจายตัว ส่วน β จะบ่งบอกถึง skewness ของการกระจายตัว หาก $\beta = 0$ นั้นหมายความว่า การกระจายตัวเป็นแบบสมมาตร หาก $\beta > 0$ นั้นการกระจายตัวเป็นแบบ right-skewed (or positively-skewed) นั่นคือ หางของการกระจายตัวทางด้านขวานั้นยาวกว่าทางด้านซ้าย (histogram จึงดูเบ้ซ้าย) และ หาก $\beta < 0$ นั้น การกระจายจะเป็นแบบ left-skewed

เราสามารถแสดงให้เห็นได้ว่า Normal, Cauchy, Levy distributions ต่างก็เป็น stable distribution ด้วยความสัมพันธ์ของพารามิเตอร์ดังนี้ [Nol18, §1.9]

- ตัวแปร normal

$$\mathcal{N}(\mu, \sigma^2) = S(2, 0, \sigma/\sqrt{2}, \mu; 0) \quad \text{และ} \quad S(2, 0, \gamma, \delta; 0) = \mathcal{N}(\delta, 2\gamma^2)$$

  นั่นคือ ตัวแปร Gaussian เป็นกรณีเฉพาะของ stable distribution เมื่อ $\alpha = 2$

- ตัวแปร cauchy

$$\text{Cauchy}(\gamma, \delta) = S(1, 0, \gamma, \delta; 0)$$

- ตัวแปร Levy

$$\text{Levy}(\gamma, \delta) = S(1/2, 1, \gamma, \delta + \gamma; 0)$$

ตัวแปร stable นั้น หาก $\alpha < 2$ จะมีลักษณะของ heavy tail (มีหางที่หนากว่าการกระจายตัวแบบ exponential) ใน [Nol18, §1.5] ได้แสดงให้เห็นว่า ค่า $P(X > x)$ เมื่อ x มีค่ามาก จะเป็นไปตาม power law ดังนี้

$$P(X > x) \sim \gamma^\alpha c_\alpha (1 + \beta) x^{-\alpha}, \quad f(x) \sim \alpha \gamma^\alpha c_\alpha (1 + \beta) x^{-(\alpha+1)}$$

t-Location scale distribution. ก่อนอื่นเราจะกล่าวว่า a location-scale family คือเซตของ distribution ที่สามารถแจกแจงได้ด้วย location parameter (μ) และ scale parameter (σ ที่มีค่าเป็นบวก) ตัวอย่างเช่น เมื่อกำหนดให้ $f(x)$ เป็น pdf เราจะเห็นว่า

$$f(x|\mu, \sigma) = \frac{1}{\sigma} f\left(\frac{x - \mu}{\sigma}\right)$$

ก็มีสมบัติเป็น pdf ด้วย เราจะเรียกว่า $f(x|\mu, \sigma)$ เป็น location-scale family สำหรับตัวอย่างการกระจายตัวที่เป็น location-scale family นั้น เช่น ตัวแปร Gaussian ที่มี $f(x|\mu, \sigma) = (1/\sqrt{2\pi\sigma^2}) e^{-(x-\mu)^2/2\sigma^2}$ หรือ ตัวแปร exponential ที่มี $f(x|\mu) = e^{-(x-\mu)}$ เมื่อ $x > \mu$ เป็นต้น ในทางคณิตศาสตร์ เราสามารถนิยาม location-scale family ได้ดังนี้

นิยาม 2. *เราจะกล่าวว่า X ที่มี pdf $f_X(x) = f(x; \mu, \sigma)$ ว่าเป็นสมาชิกของ location-scale family ก็ต่อเมื่อ มีตัวแปรสุ่ม Z ที่มี $f_Z(z) = f(z)$ (อยู่ใน family นั้นๆ) ที่ทำให้*

$$Z = \sigma Z + \mu$$

หรือกล่าวอีกนัยหนึ่งคือ เมื่อ X เป็น linear transformation ของตัวแปรสุ่ม Z

อีกนิยามหนึ่งของ location-scale family ที่สมมูลกัน [Lov11] สามารถอธิบายได้ดังนี้

นิยาม 3. *ตัวแปรสุ่ม X อยู่ใน location-scale family ถ้าหาก cdf เป็นฟังก์ชันของ $(x - \mu)/\sigma$ นั่นคือ*

$$F_X(x|\mu, \sigma) = F\left(\frac{x - \mu}{\sigma}\right), \quad \mu \in \mathsf{R}, \sigma > 0$$

โดยที่ $F(\cdot)$ ที่ต่างกัน ก็คือสมาชิกที่ต่างกันใน family นั้น

เราจะพบว่ามีหลายการกระจายตัวที่อยู่ใน location-scale family อาทิเช่น arc-sine, cauchy, normal, exponential, rayleigh, laplace, logistic distributions เป็นต้น

ในผลลัพธ์การทดลองเพื่อหาการกระจายตัวของค่าความคลาดเคลื่อนการพยากรณ์ เราจะพบว่า การกระจายตัวแบบ t-location scale นั้นถูกเลือกให้เป็นการกระจายตัวที่เหมาะสมกับข้อมูลในหลายกรณี ในหัวข้อนี้ จึงจะอธิบายเพิ่มเติมดังต่อไปนี้

ตัวแปรสุ่ม X อยู่ใน t-location scale family คือตัวแปรที่ได้จาก linear transformation $X = (T - \mu)/\sigma$ เมื่อ T เป็น student t distribution นั่นคือ X จะมี pdf เป็น

$$f(x) = \frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\sigma\sqrt{\nu\pi}\Gamma(\nu 2)}\left[\frac{\nu + \left(\frac{x-\mu}{\sigma}\right)^2}{\nu}\right]^{-\left(\frac{\nu+1}{2}\right)}, \quad -\infty < x < \infty \tag{2.6}$$

เมื่อ $-\infty < \mu < \infty$ คือ location parameter, $\sigma > 0$ คือ scale parameter และ $\nu > 0$ คือ shape parameter (เป็นตัวเดียวกับ degree of freedom ในตัวแปร t) เป็นที่ทราบกันดีว่า สำหรับตัวแปร student t นั้น ค่าเฉลี่ยจะนิยามเมื่อเมื่อ $\nu > 1$ และค่าความแปรปรวนจะ finite เมื่อ $\nu > 2$ ดังนั้น ภายใต้เงื่อนไขดังกล่าว ตัวแปร t location-scale จะมีค่าเฉลี่ยเป็น ν เมื่อ $\nu > 1$ และมีค่าความแปรปรวนเป็น $\sigma^2\nu/(\nu - 2)$ เมื่อ $\nu > 2$ ตัวแปร t location-scale มักจะใช้อธิบายตัวแปรที่มี heave tails

Extreme value distribution. การกระจายตัวชนิดนี้มักจะใช้อธิบายตัวแปรที่มีความหมายเป็นค่าสูงสุด หรือค่าน้อยสุดของกลุ่มข้อมูลที่เป็นอิสระต่อกันภายใต้การกระจายตัวเดียวกันหนึ่งๆ [Lov11] Extreme value distributions ใช้ในศาสตร์ด้าน extreme value theory (EVT) ที่ศึกษาแบบจำลองของเหตุการณ์ที่เกิดไม่บ่อย (rare event with small probability) เช่น งานในกลุ่ม risk management, insurance, hydrology, material science เป็นต้น เมื่อกล่าวในทางคณิตศาสตร์ หมายถึง การศึกษา asymptotic distribution ของ

$$Y_n = \max\{X_1, X_2, \ldots, X_n\}$$

เมื่อ $N \to \infty$ ทั้งนี้ เมื่อ ต้องการศึกษาค่าน้อยสุดของกลุ่ม samples ก็สามารถประยุกต์ใช้ได้ โดยการกลับเครื่องหมายของ Y_n เป็นค่าลบ ในกลุ่มการกระจายตัวนี้สามารถแบ่งออกได้เป็น type I, II, III ที่มีนิยามของ standard pdf ในรูปดังนี้

1. Gumbel (type I): มักใช้กับการจำลองเหตุการณ์ทางด้าน hydrology ที่เป็น extreme events (flood peak เป็นต้น)

$$f(x) = \frac{1}{\sigma}e^{-(x-\mu)/\sigma}e^{-e^{-(x-\mu)/\sigma}}, \quad x \in \mathsf{R}$$

ที่มีพารามิเตอร์เป็น $\mu \in \mathsf{R}, \sigma > 0$ ใน MATLAB จะอ้างถึง extreme value distribution ว่าเป็น type I โดยจะพิจารณาตัวแปรสุ่มที่จำลองค่า minimum แทน ดังนั้น รูปแบบ pdf ของ extreme value ใน MATLAB จะเขียนโดยใช้ $(x - \mu)/\sigma := -(x - \mu)/\sigma$

2. Fréchet (type II): ใช้กับงานประยุกต์ด้านการเงิน เช่น marget-returns ที่มีการกระจายตัวเป็น heavy-tails

$$f(x) = \begin{cases} 0, & x \le \mu, \\ \frac{\alpha}{\sigma}\left(\frac{x-\mu}{\sigma}\right)^{-(\alpha+1)} e^{-\left(\frac{x-\mu}{\sigma}\right)^{-\alpha}}, & x > \mu, \alpha > 0 \end{cases}$$

ที่มีพารามิเตอร์เป็น $\mu \in \mathsf{R}, \sigma > 0, \alpha > 0$ (location, scale, shape parameters ตามลำดับ)

3. Weibull (type III): ใช้กับงานด้าน fatigue analysis ของ material science หรือถ้าหาก X ใช้จำลองปริมาณ time-to-failure การกระจายตัว Weibull จะหมายถึง การอธิบาย failure rate ที่แปรผันกับฟังก์ชันยกกำลังของเวลา

$$f(x) = \begin{cases} \frac{k}{\sigma}\left(\frac{x}{\sigma}\right)^{k-1} e^{-(x/\sigma)^k}, & x > 0, k > 0, \\ 0, & x < 0. \end{cases}$$

ที่มีพารามิเตอร์เป็น $k > 0, \sigma > 0$ เป็น shape และ scale parameters ตามลำดับ กรณีเฉพาะของ Weibull distributions นั้นลดรูปเป็นการกระจายตัวอื่นอีกด้วย ตัวอย่างเช่น เมื่อ $k = 1$ จะกลายเป็น exponential distribution เมื่อ $k = 2$ จะกลายเป็น Rayleigh distribution (ที่มักจะใช้ประยุกต์ในงานด้าน telecommunication)

Generalized Extreme Value distribution. การกระจายตัวนี้ คือการรวม extreme value distributions ทั้ง 3 แบบ มาเขียนในรูปทั่วไป ที่มีพารามิเตอร์ 3 ตัวได้แก่ k, μ, σ (shape, location, scale parameters ตามลำดับ)

- เมื่อ $k \neq 0$ และสำหรับ $1 + k(x-\mu)/\sigma > 0$

$$f(x) = \frac{1}{\sigma} e^{-\left(1+\frac{k(x-\mu)}{\sigma}\right)^{-1/k}} \left(1 + \frac{k(x-\mu)}{\sigma}\right)^{-1-\frac{1}{k}}$$

  โดยที่ เมื่อ $k > 0$ จะเป็น extreme value distribution type II และเมื่อ $k < 0$ จะเป็น extreme value distribution type III ตามลำดับ

- เมื่อ $k = 0$ จะกลายเป็นกรณีของ Gumble (type I)

$$f(x) = \frac{1}{\sigma} e^{-(x-\mu)/\sigma} e^{-e^{-(x-\mu)/\sigma}}, \quad x \in \mathbf{R}$$

2.2.2 Kernel density estimation

กำหนดให้ $X_1, X_2, \ldots, X_n$ คือตัวอย่างข้อมูลที่สุ่มมาจาก distribution ที่มี pdf เป็น $f(x)$ การประมาณความหนาแน่นเคอร์เนลนั้น มีพื้นฐานจากฮิสโทแกรมที่นิยามไว้เพื่อประมาณ $f(x)$ ดังนี้ [Sil98, §2]

$$\hat{f}(x) = \frac{1}{nh} \text{ (จำนวนข้อมูล } X_i \text{ ที่ตกอยู่ในช่วง } [x-h, x+h])$$

โดยมีพารามิเตอร์ $h > 0$ คือความกว้างของ bin ตัวประมาณดังข้างต้นจะถูกเรียกว่า naive estimator หากเราจะเขียนตัวประมาณดังข้างต้นในรูปคณิตศาสตร์ โดยการนิยามฟังก์ชันถ่วงน้ำหนัก $w(x) = 1/2$ เมื่อ $|x| < 1$ และ $w(x) = 0$ เมื่อ $|x| \geq 1$ เราจะพบว่า การประมาณดังข้างต้น สามารถเขียนได้ในรูป

$$\hat{f}(x) = \frac{1}{nh} \sum_{i=1}^{n} w\left(\frac{x - X_i}{h}\right)$$

การใช้ $w(x)$ ซึ่งเปรียบเสมือนฟังก์ชันสี่เหลี่ยมนั้นพบว่า อาจไม่เหมาะสมกับการประมาณ pdf เนื่องจาก w ไม่ใช่ฟังก์ชันต่อเนื่อง (มีจุดไม่ต่อเนื่องตรงขอบของ bin) ดังนั้น จึงเป็นที่มาของการใช้ฟังก์ชันเคอร์เนลมาแทนฟังก์ชันสี่เหลี่ยม โดยที่เราจะนิยามว่าฟังก์ชันเคอร์เนล $K(x)$ นั้น เป็นฟังก์ชันสมมาตรที่สอดคล้องกับเงื่อนไข

$$\int K(t)dt = 1, \quad \int tK(t)dt = 0, \quad \int t^2 K(t)dt = k_2 \neq 0 \tag{2.7}$$

นิยามของตัวประมาณเคอร์เนล (kernel estimator) จึงเทียบเคียงกับ naive estimator ดังนี้

$$\hat{f}(x) = \frac{1}{nh} \sum_{i=1}^{n} K\left(\frac{x - X_i}{h}\right) \tag{2.8}$$

เราจะเรียก h ว่าเป็นความกว้างหน้าต่าง (window width) หรือเรียกว่าเป็น smoothing parameter หรือ bandwidth เราจะแสดงให้เห็นได้ว่า เมื่อ h มีค่าเล็กลง $\hat{f}(x)$ จะมีรายละเอียดเยอะใน bin ย่อยๆ จนเมื่อ h มีค่ามาก รายละเอียดดังกล่าวจะถูก smoothen out นอกจากนี้ เราจะเห็นจากนิยาม (2.8) และคุณสมบัติของ $K(x)$ ว่า $\hat{f}(x)$ ที่ได้นั้นมีคุณสมบัติเหมือน pdf (กล่าวคือ เป็นค่าบวก และมีพื้นที่ใต้กราฟเท่ากับ 1) นอกจากนี้ ความต่อเนื่องและการหาอนุพันธ์ได้ของ $\hat{f}(x)$ ก็จะถ่ายทอดมาจากคุณสมบัติของฟังก์ชันเคอร์เนลที่เลือกมาใช้

การเลือก width. จากทฤษฎีการประมาณ โดยทั่วไปนั้น จะใช้ตัวชี้วัด mean squared error (MSE): MSE = $\mathbf{E}[\hat{f}(x) - f(x)]^2$ ว่าตัวประมาณที่ใช้มีคุณภาพดีเพียงใด จากหลักการของการประมาณ เราสามารถแสดงให้เห็นได้ว่า MSE นั้น ประกอบไปด้วยสองเทอมเสมอ อันได้แก่ sum of squared bias และ variance ของตัวประมาณ ดังนี้

$$\mathrm{MSE} = \{\mathbf{E}[\hat{f}(x)] - f(x)\}^2 + \mathbf{var}\,\hat{f}(x)$$

ใน [Sil98, §3.2.1] ได้แสดงให้เห็นว่า ค่าเฉลี่ยและค่าความแปรปรวนของ $\hat{f}(x)$ นั้น อยู่ในรูป

$$\begin{aligned}\mathbf{E}[\hat{f}(x)] &= \int \frac{1}{h} K\left(\frac{x-y}{h}\right) f(y)dy,\\ n\,\mathbf{var}\,\hat{f}(x) &= \int \frac{1}{h^2} K\left(\frac{x-y}{h}\right)^2 f(y)dy - \left[\frac{1}{h}\int K\left(\frac{x-y}{h}\right) f(y)dy\right]^2.\end{aligned}$$

จากผลลัพธ์ข้างต้น เราจะเห็นว่า bias ของการประมาณ $f(x)$ นั้นคือ smoothed version ของฟังก์ชัน $f(x)$ ซึ่งขึ้นกับฟังก์ชันเคอร์เนลที่เลือก และไม่ได้ขึ้นกับจำนวน samples ที่ใช้ n เมื่อใช้การประมาณ $f(x)$ ด้วย Taylor series ค่า bias และความแปรปรวนของ $\hat{f}(x)$ นั้นจะประมาณได้ว่าเป็นฟังก์ชันของ h, n และฟังก์ชันเคอร์เนล ดังนี้ [Sil98, §3.3.1]

$$\mathrm{bias} \approx (1/2)Ch^2 f''(x), \quad \mathbf{var}\,\hat{f}(x) \approx \frac{1}{nh} f(x)\int K(x)^2 dt$$

จะเห็นว่า $\mathbf{var}\,\hat{f}(x)$ นั้นจะขึ้นกับทั้ง n และ h แบบผกผัน ด้วยผลลัพธ์นี้ จึงเกิดภาวะ bias-variance trade-off กล่าวคือ หากเลือกให้ h มีค่าน้อยเพื่อให้ bias ต่ำ แต่ variance ของการประมาณจะกลับมีค่าสูงขึ้น

ตัวอย่างฟังก์ชันเคอร์เนล. เมื่อพิจารณาผลของ h ต่อ MSE แล้วนั้น เราสามารถแสดงให้เห็นว่า เมื่อแทนค่า h ที่เหมาะสมที่สุด (minimize MSE) จะได้ว่าค่า MSE จะลดรูปเหลือเป็นฟังก์ชันของ

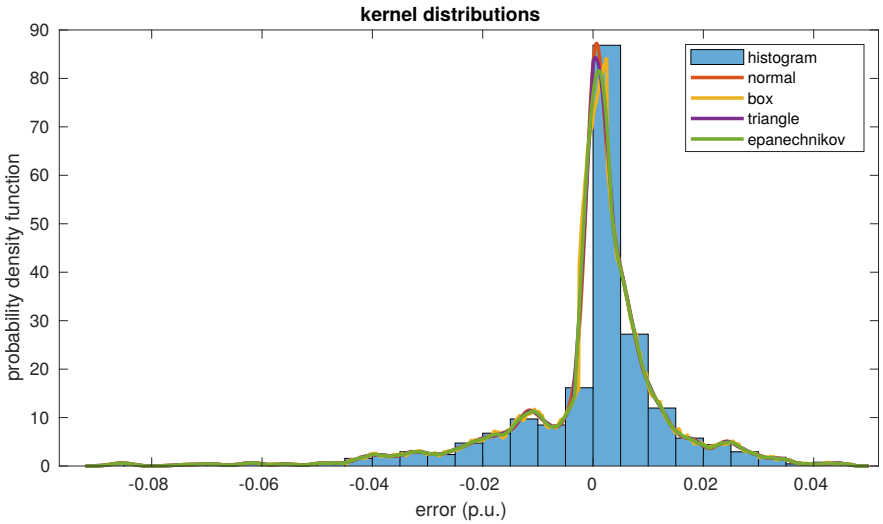
$$k_2^{2/5}\left(\int K(t)^2 dt\right)^{4/5} \tag{2.9}$$

นั่นคือ การเลือกฟังก์ชันเคอร์เนล ที่นอกจากจะต้องเป็นไปตามเงื่อนไข (2.7) จึงควรทำให้ค่า MSE ดัง (2.9) มีค่าน้อยอีกด้วย ตัวอย่างของฟังก์ชันในการใช้งานประยุกต์คือ

- Epanechnikov kernel: เป็นฟังก์ชันเคอร์เนลที่ทำให้ (2.9) น้อยที่สุด (optimal)
$$K(x) = \frac{3}{4\sqrt{5}}\left(1 - \frac{x^2}{5}\right), \quad |x| \le \sqrt{5}$$
- Gaussian kernel
$$K(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2}, \quad x \in \mathsf{R}$$
- Triangle kernel
$$K(x) = 1 - |x|, \quad |x| < 1$$
- Rectangular kernel (or box kernel)
$$K(x) = 1/2, \quad |x| < 1$$

การประมาณความหนาแน่นเคอร์เนล สามารถเรียกใช้คำสั่ง `fitdist` ได้เช่นกัน โดยการกำหนดว่าเป็น kernel distribution ที่ต้องเลือกฟังก์ชันเคอร์เนล และความกว้างหน้าต่างเป็น option หรือจะเรียกใช้ `ksdensity` ได้เช่นกัน ส่วนใน python สามารถเรียกใช้ `scipy statsmodels`

จากการทดลอง เราพบว่าผลของการเลือกฟังก์ชันเคอร์เนลหลายแบบ ไม่ได้ส่งผลต่างมากนักหากดูด้วยกราฟในรูป 2.1 แต่เราสามารถดูผลของการประมาณจากค่าฟังก์ชันความเป็นไปได้ (loglikelihood) รวมถึง komogorov-smirnov statistics ได้ ดังเช่นตัวอย่างในตาราง 2.1 ซึ่งในตัวอย่างนี้ การเลือกใช้ฟังก์ชันเคอร์เนลที่ต่างกัน ส่งผลให้ค่า statistics แทบไม่ต่างกัน (แต่ p-value มีค่าต่างบ้าง) และทุกฟังก์ชันก็ผ่าน hypothesis test ทุกตัว



รูป 2.1: ตัวอย่างการประมาณการกระจายตัวความคลาดเคลื่อนการพยากรณ์ ณ เวลา 6:30 น. ด้วยวิธี kernel density estimation

ตาราง 2.1: ตัวอย่างการประมาณด้วย kernel density estimation ของค่าความคลาดเคลื่อนที่เวลา 10:30 น.

Distribution	Negative loglikelihood	KS statistics	p-value	h-test
normal	-789.748	0.011	0.88186	0
box	-784.761	0.011	0.85485	0
triangle	-790.758	0.011	0.87178	0
epanechnikov	-787.550	0.011	0.86364	0

2.2.3 วิธี Bootstrap

ในหัวข้อนี้ จะอธิบายเทอมที่สำคัญทางสถิติ เพื่อใช้อธิบายหลักการ Bootstrap [ET93, JWHT13, HTF09] เราจะเรียก $\mathbf{y}^* = (y_1^*, y_2^*, \ldots, y_N^*)$ ว่าเป็น **a bootstrap sample** ที่สุ่มมาจาก $\hat{F}$ เมื่อเราสุ่ม N samples มาจาก $\mathbf{y}$ และการสุ่มนั้นเป็น *การสุ่มแบบคืน* (draw with replacement) ด้วยนิยามดังนี้ เราจะเห็นว่า ตัวอย่างของ bootstrap samples เมื่อ $N = 5$ อาจจะเป็นไปได้ดังนี้

$$\mathbf{y}^* = (y_1, y_1, y_2, y_3, y_5), \quad \mathbf{y}^* = (y_2, y_2, y_4, y_4, y_5), \quad \mathbf{y}^* = (y_1, y_2, y_3, y_4, y_5)$$

กล่าวคือ bootstrap samples อาจจะประกอบไปด้วยชุดของ samples เดิมทั้งหมด อาจจะประกอบไปด้วยค่า samples บางค่าเท่านั้น บาง samples อาจจะไม่โดนเลือกมาเลย ก็เป็นไปได้ทั้งหมด และเราจะเรียก F^* ว่าเป็น resampling distribution

เมื่อเรามี samples $\mathbf{y}$ ของตัวแปรสุ่ม y กำหนดให้ $s(\mathbf{y})$ เป็น statistic หนึ่งๆ นั่นคือฟังก์ชันใดๆ ที่รับค่า samples ของ y ออกมาเป็นค่า deterministic ค่าหนึ่ง ตัวอย่างเช่น $s(\mathbf{y}) = \bar{y} = (1/N)\sum_{k=1}^{N} y_k$ เป็น sample mean เราก็สามารถนิยามได้เช่นเดียวกันกับ bootstrap samples นั่นคือ $s(\mathbf{y}^*)$ เป็นการหา statistic ที่สนใจที่คำนวณจาก bootstrap samples การสุ่ม (แล้วคืน) เพื่อได้ bootstrap samples หลายๆ ชุด ก็ย่อมได้ค่า $s(\mathbf{y}^*)$ หลายชุดเช่นกัน ซึ่งจะเรียกว่าเป็น **bootstrap estimates**

Bootstrap algorithm. ขั้นตอนวิธี bootstrap นั้น ประกอบไปด้วย

1. กำหนด B ให้เป็นจำนวน bootstrap samples เราจะเลือก bootstrap samples ทั้งหมด B ชุด

$$\mathbf{y}^{*1}, \mathbf{y}^{*2}, \ldots, \mathbf{y}^{*B}$$

โดยที่แต่ละ bootstrap sample นั้น มีขนาด sample size เป็น N

2. คำนวณ bootstrap estimates (statistic ที่สนใจ) บนแต่ละ bootstrap sample นั้น

$$\hat{\theta}(b) = s(\mathbf{y}^{*b}), \quad b = 1, 2, \ldots, B$$

ด้วยการใช้ค่า B ที่มากพอ เราจะได้ sampling distribution ของ $\hat{\theta}(b)$

3. ขั้นตอนถัดไป ขึ้นกับจุดประสงค์ของการนำ bootstrap ไปใช้

หลักการของ bootstrap นั้น (in loose sense) กล่าวคือ เมื่อ B มีค่ามากพอ

A. $F^* \approx F$

B. การกระจายตัวของ $s(\mathbf{y})$ นั้นสามารถประมาณได้จากการกระจายตัวของ $s(\mathbf{y}^*)$

ตัวอย่าง 1. ตัวอย่างหนึ่งของการนำ bootstrap ไปใช้ [HTF09] อย่างเช่น การที่เราสนใจ statistic คือ sample mean $\bar{y}$ ซึ่งคำนวณมาจาก sample $\mathbf{y}$ ขนาด N ที่อาจจะ draw มาจาก population ที่เป็น distribution ใดๆ ที่เราไม่ทราบ ดังนั้น ค่า sample mean ที่คำนวณได้นั้น จะมีการกระจายตัวทางทฎษฎีเป็นอย่างไรก็ไม่ทราบได้ (กรณีเฉพาะคือ เมื่อ N มากพอ เราสามารถอ้าง central limit theorem ได้ว่า $\bar{y}$ จะเข้าใกล้ Gaussian แต่สมมติว่า N อาจจะมีค่าไม่มากได้) ดังนั้น หากคำถามที่สนใจคือ อยากทราบการกระจายตัวของ $\bar{y}$ หรืออยากทราบความแปรปรวนของ $\bar{y}$ เราจะได้ว่าในขั้นตอน bootstrap ที่ 2 เราสามารถคำนวณ

$$\hat{\theta}(b) = \bar{y}^* = \frac{1}{N}\sum_{k=1}^{N} y^{*k}, \quad b = 1, 2, \ldots, B$$

และในขั้นตอนที่ 3 ของ bootstrap เราสามารถคำนวณ*ค่าประมาณความแปรปรวน*ของ $\bar{y}$ ดังนี้

$$\widehat{\mathbf{var}}(\bar{y}) = \frac{1}{B-1}\sum_{b=1}^{B}\left(\hat{\theta}(b) - (1/B)\sum_{b=1}^{b}\hat{\theta}(b)\right)^2$$

ตัวอย่าง 2. สมมติว่าเราสนใจ statistic คือ median[1] นั่นคือ $\hat{\theta}(b)$ และเรามีคำถามว่า ต้องการหา confidence interval (CI) ของ sample median ค่านั้นว่าอยู่บนช่วงใด กล่าวคือ $\theta(y) = y_{\text{median}}$ และ $\hat{\theta}(b) = y^{*b}_{\text{median}}$ คือ bootstrap estimate บนการ resampling ครั้งที่ b ในกรณีนี้ หากในขั้นตอนที่ 2 เราคำนวณ

$$\delta^{*b} = \hat{\theta}(b) = y^{*b}_{\text{median}} - y_{\text{median}}, \quad b = 1, 2, \ldots, B$$

และเราสามารถเรียง $\delta^* = (\delta^{*1}, \delta^{*2}, \ldots, \delta^{*B})$ จากน้อยไปมาก เพื่อหา 95th และ 5th percentiles (เรียกว่า $\delta^*_{0.05}$ และ $\delta^*_{0.95}$ ตามลำดับ) จากหลักการของ bootstrap นั้น เราจะประมาณได้ว่า *ค่าประมาณแบบ bootstrap ของช่วง CI* ด้วยความเชื่อมั่น 0.9 นั้น สามารถหาได้จาก

$$[y_{\text{median}} - \delta^*_{0.05}, y_{\text{median}} - \delta^*_{0.95}]$$

จากตัวอย่างนี้ เราจะเห็นว่าในขั้นตอนที่ 3 ของ bootstrap จึงเป็นการประมาณการกระจายตัวของค่า sample median ด้วย bootstrap และการกระจายตัวนั้น คือการใช้ค่า percentiles เพื่อมาใช้อธิบายช่วงความเชื่อมั่นของค่า sample median

[1] จาก Jeremy Orloff and Jonathan Bloom lecture note

2.2.4 Komogorov-Smirnov test

Komogorov-Smirnov test (KS test) เป็น nonparametric test แบบหนึ่ง ที่ใช้ทดสอบสมมติฐานว่า distribution ที่สนใจนั้นใกล้เคียงกับ empirical distribution เพียงใด โดย distribution ที่สนใจ อาจจะมาจากสมมติฐานของผู้ใช้ ที่ผ่านขั้นตอนการประมาณ distribution มาแล้วด้วยวิธีหนึ่งๆ (เช่น kernel distribution, parametric distribution เป็นต้น) ส่วน empirical distribution นั้น นิยามดังนี้

$$F_n(x) = \frac{1}{n}\sum_{i=1}^{n} I_{\{X_i \geq x\}}(x)$$

โดยที่เราใช้สัญลักษณ์ indicator function: $I_C(x) = 1$ ถ้า $x \in C$ และมีค่าเป็นศูนย์ นอกเหนือจากนั้น เราจะเห็นว่า $F_n(x)$ จะแสดงถึงสัดส่วนความน่าจะเป็นของข้อมูลจากตัวอย่าง $X_1, X_2, \ldots, X_n$ ที่มีค่าน้อยกว่า x และ $F_n(x)$ เป็นตัวประมาณหนึ่งของ cdf

หากเรากำหนดให้ $F(x)$ หมายถึง cdf ของการกระจายตัวหนึ่งๆ ที่สนใจ ใน KS test นั้น จะใช้ statistics คือ

$$D_n = \sup_x |F_n(x) - F(x)| \tag{2.10}$$

โดยที่ $\sup$ คือ supremum ของฟังก์ชัน หาก $F(x)$ ที่เราสนใจมีค่าใกล้เคียงกับ $F_n(x)$ จะทำให้ค่า statistics D_n มีค่าน้อย เราก็อาจเชื่อได้ว่า ตัวอย่างข้อมูลที่สนใจนั้น เหมาะกับการกระจายตัวแบบ $F(x)$ การทดสอบสมมติฐานดังกล่าว จะใช้ null hypothesis ว่าคือ H_0 : ข้อมูลมาจากการกระจายตัว $F(x)$ และจะเรียกว่าเป็น one-sample test หากใช้ H_0 : ข้อมูลจากสองการกระจายตัวนั้นเหมือนกัน เราจะเรียกการทดสอบดังกล่าวว่าเป็น two-sample test

ในการทดลองเราจะใช้คำสั่ง `kstest` ใน MATLAB ที่มี input เป็น $F(x)$ ที่ต้องการทดสอบ และมี output เป็น KS statistics, p-value และผลการทดสอบ hypothesis ส่วนใน python ผู้ใช้สามารถใช้ `scipy.stats.kstest`

2.3 ดัชนีสมรรถนะของ probabilistic forecasts

การเปรียบเทียบสมรรถนะของการพยากรณ์เชิงสถิติ จะใช้ตัวชี้วัดดังที่จะกล่าวต่อไปนี้ เราจะใช้สัญลักษณ์ กำหนดให้ $\hat{f}(y)$ และ $\hat{q}(\alpha)$ คือ nonparametric density distribution และค่าประมาณของ quantiles ของตัวแปร y

1. Reliability [ZMAP14, GPG16]: นิยามคือ ค่าประมาณของความน่าจะเป็นที่ข้อมูล samples $\{y_i\}_{i=1}^N$ จะน้อยกว่าค่า quantiles ที่ค่า α หนึ่งๆ

   $$\hat{\alpha}_k = \frac{1}{N}\sum_{i=1}^{N} I\{y_i \leq \hat{q}(\alpha_k)\}$$

   ค่า $\hat{\alpha}_k$ นั้น จะค่าความน่าจะเป็นที่คิดเฉลี่ยจาก samples ซึ่งปกติมักจะเป็น time samples (นั่นคือ i เป็น time index) นอกจากนี้ Reliability นั้น ขึ้นกับ lead time ด้วย เพราะ lead time ของ forecasts คนละค่าก็จะมี nonparametric density function คนละชุดกัน Reliability diagram เป็นกราฟระหว่างค่า Reliability กับ nominal level (α) ซึ่งหากค่าวิธีการพยากรณ์แบบ probabilistic หนึ่งๆ มีสมรรถนะที่ดี จะได้ว่า ค่า reliability ที่ค่า α หนึ่งๆ ก็ควรไม่ต่ำว่า α นั้นๆ กล่าวคือ กราฟ reliability ควรเป็นกราฟไม่ต่ำกว่าเส้นตรงที่ความชัน 45°

2. PI coverage probability (PICP) [KNC13] เป็น score ที่บอกความน่าจะเป็นของข้อมูลวัดที่ตกอยู่ในช่วง prediction interval (PI) ที่ประมาณว่ามีมากเท่าใด (ยิ่งมากยิ่งดี)

   $$\text{PICP} = \frac{1}{N}\sum_{i=1}^{N} I\{y_i \in [l_i, u_i]\}$$

   โดยที่ $[l_i, u_i]$ คือ PI ที่ประมาณได้ (มักจะขึ้นกับจุดเวลา) และ y_i คือข้อมูลวัด (unseen) ที่นำมาทดสอบ และ N คือจำนวนข้อมูลที่มาทดสอบ โดยปกติแล้วนั้น การประมาณ PI ที่ได้แต่แรกจะขึ้นกับ nominal coverage (γ) ดังนั้น PICP ควรมีค่ามากกว่า $(1-\gamma)\%$ หากมีค่าน้อยกว่านั้น จึงถึงว่า PI ที่ประมาณได้ไม่น่าเชื่อถือ (unreliable)

3. Deviation [GPG16]: กำหนดให้นิยามคือ ความแตกต่างระหว่างค่า nominal level ที่กำหนดกับค่า nominal level ที่ประมาณได้

$$\text{Deviation}_k = |\alpha_k - \hat{\alpha}_k|$$

ค่านี้มักจะใช้ดู overall reliability ซึ่งจะดูว่า หากมีค่าต่ำ เราจะกล่าวว่า การพยากรณ์นั้นสามารถเชื่อถือได้มากกว่า (more reliable)

4. Sharpness [PNM+07] กำหนดให้เป็นความกว้างของช่วง quantiles ดังนี้

$$\text{Sharpness}(c) = \hat{q}(\alpha_2) - \hat{q}(\alpha_1), \quad c = \alpha_2 - \alpha_1$$

ซึ่งโดยปกติแล้ว Sharpness จะขึ้นกับ lead time และ เวลาที่พยากรณ์ ดังนั้น เราสามารถคำนวณค่าเฉลี่ยของค่าข้างต้นของข้อมูลจากทุกๆ เวลาพยากรณ์ได้

5. Normalized Mean PI width (NMPIW) [KNC13] คือค่าเฉลี่ยความกว้างของ PI ตามนิยามดังนี้

$$\text{NMPIW} = \frac{1}{RN}\sum_{i=1}^{N}(u_i - l_i)$$

โดย R คือค่าคงที่ที่แสดงถึง scale ของข้อมูลของ data set นั้น เรา normalize เพื่อที่จะทำให้ NMPIW สามารถนำไปเปรียบเทียบผลกันได้ในแต่ละ data set เราจะเห็นว่า PI ที่มีช่วงกว้าง ย่อมทำให้ PICP มีค่ามาก ดังนั้น width จึงเป็นตัวชี้วัดที่ใช้ดูประกอบกัน

เราจะพบว่า ตัวชี้วัด Reliability, PI coverage probability และ Deviation นั้น เป็นตัวชี้วัดที่มีความหมายคล้ายกัน คือเป็นการเทียบความน่าจะเป็นของข้อมูลใหม่ที่จะตกอยู่ในช่วง quantiles หรือช่วงความเชื่อมั่นที่สอดคล้องกับค่าความเชื่อมั่นหนึ่งๆ อย่างไรส่วนตัวชี้วัด Sharpness และ Normalized Mean PI width เป็นกลุ่มตัวชี้วัดที่บอกความกว้างของช่วงการทำนาย หรือช่วง quantiles ที่สนใจ หากช่วงกว้างมากไป หมายความว่าประสิทธิภาพการพยากรณ์เชิงสถิติวิธีนั้น ยังไม่ดีพอ เพราะให้ค่าพยากรณ์ที่มีการกระจายตัวมากเกินไป

บทที่ 3

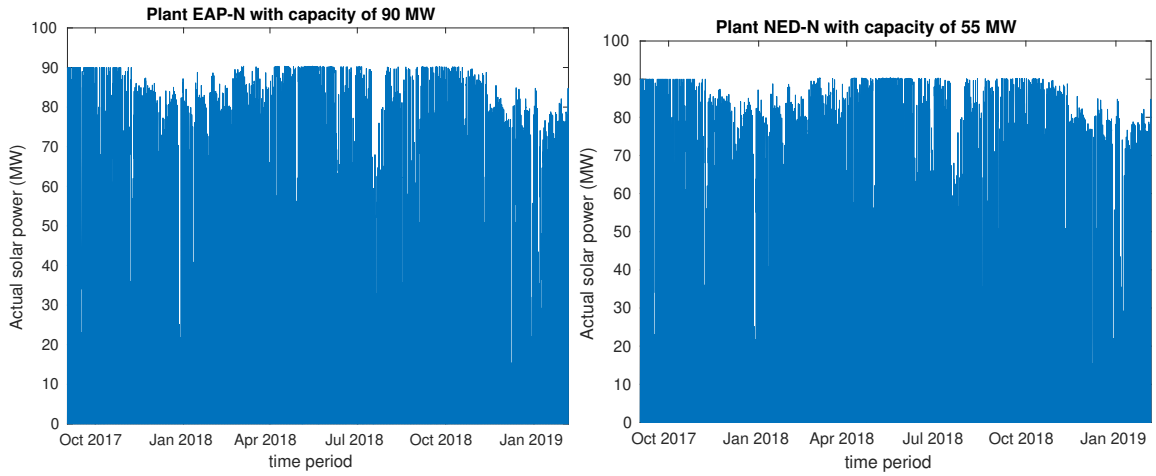
แบบจำลองการพยากรณ์และข้อมูลของ Provider

การวิเคราะห์ความคลาดเคลื่อนของค่าพยากรณ์นั้นจำเป็นจะต้องทราบถึงข้อกำหนดพยากรณ์ และโครงสร้างของแบบจำลองที่ใช้ รวมถึงรายละเอียดของข้อมูลและผลการพยากรณ์ที่ได้ ในเนื้อหาส่วนนี้จึงจะบรรยายรายละเอียดดังที่กล่าวมา โดยเป็นข้อมูลที่ได้รับจาก Provider ข้อกำหนดการพยากรณ์ของ Provider แสดงดังตาราง 3.1 ซึ่งเริ่มพยากรณ์วันที่ 9 มกราคม พ.ศ. 2562 จนถึงข้อมูลล่าสุดที่ทีมวิจัยได้รับคือค่าพยากรณ์ของวันที่ 29 กุมภาพันธ์ พ.ศ. 2563

ตาราง 3.1: ข้อกำหนดการพยากรณ์ของ Provider

Parameter	Hour-ahead forecasting	Day-ahead forecasting
horizon	1 วัน	7 วัน
Duration of forecasts	00:00-23:00	00:00-23:00
forecasting frequency	ทุกชั่วโมง	ทุกวัน
forecasting resolution	30 นาที	30 นาที
จำนวนค่าพยากรณ์ในแต่ละครั้ง	น้อยกว่าหรือเท่ากับ 48 จุด	72 จุด

จากการนำค่ากำลังไฟฟ้าที่ผลิตได้จาก 2 โรงไฟฟ้าในช่วงระหว่างกันยายน 2017 จนถึงกุมภาพันธ์ 2019 ในรูป 3.1 จึงพบว่าในหลายช่วงเวลานั้น ค่ากำลังไฟฟ้าที่ผลิตได้มีการโดนจำกัดค่าไว้ที่ installed capacity นั่นคือ 90 MW ของโรง A และ 55 MW ของโรง B



รูป 3.1: ค่ากำลังไฟฟ้าที่ผลิตได้จริงจากโรงไฟฟ้า A (ซ้าย) และ B (ขวา)

ข้อมูลที่ Provider ใช้ในการพยากรณ์จะมาจาก 2 แหล่งนั่นคือ โรงไฟฟ้า และ GWC ซึ่งเป็นบริษัทให้บริการข้อมูลพยากรณ์ของตัวแปรทางอากาศนำมาใช้ในโปรแกรม Nostradamus ของ ABB สำหรับการพยากรณ์ตัวแปรทางอากาศนั้นจะมีการพยากรณ์

ที่หลาย step ahead ทาง Provider จะใช้ค่าพยากรณ์ GWC สำหรับ step-ahead ที่สั้นที่สุดเพื่อเป็นตัวแทนของค่าจริงที่เก็บอยู่ในตัวแปรชื่อ `GWC actual value` ส่วนค่าวัดของตัวแปรที่ได้รับจากโรงไฟฟ้า เช่น solar power ไม่ได้ใช้ระหว่างพยากรณ์ แต่ใช้ในขั้นตอนการเรียนรู้แบบจำลองเท่านั้น

สัญลักษณ์ตัวชี้เวลาของแต่ละแบบจำลองนั้น จะใช้เป็น discrete-time index $t = 1, 2, \ldots$ ซึ่งจะขึ้นกับ forecasting resolution กล่าวคือ แบบจำลอง day-ahead เราใช้ $t = 1, 2, \ldots, 72$ เพื่อแทนเวลาในเซต

day 1: 00:00,01:00,..., 23:00, day 1: 00:00,01:00,..., 23:00 ,..., day 7: 00:00,01:00,..., 23:00

สำหรับแบบจำลอง hour-ahead เราจะใช้ $t = 1, 2, \ldots, 48$ เพื่อแทนเวลาในเซต

00:00, 00:30, 01:00,..., 22:30, 23:00

แบบจำลองพยากรณ์ของ Provider เป็นการใช้โปรแกรม Nostradamus ของบริษัท ABB ซึ่งเป็นโครงสร้าง ANN ที่มี target เป็น $P(t)$ และมีรายละเอียดการใช้ข้อมูลขาเข้า ดังนี้

1. Hour-ahead forecasting: ประกอบไปด้วย ANN $24 \times 2 = 48$ แบบจำลอง แต่ละแบบจำลองมีโครงสร้าง 1 hidden layer ที่มี 7 neurons และ 1 target สำหรับ ข้อมูลขาเข้านั้นมีดังนี้

   - $\mathrm{RH}(t)$
   - $\mathrm{T}(t-1), \mathrm{T}(t), \mathrm{T}(t+1)$
   - $I(t-1), I(t), I(t+1)$
   - $\mathrm{WS}(t)$
   - $P(t-4), P(t-5), P(t-6)$

   จะเห็นว่าในแบบจำลองนี้ ข้อมูลขาเข้าที่ต่างจากแบบจำลอง day-ahead คือการใช้ข้อมูลกำลังผลิตไฟฟ้าเมื่อ 2,2.5,3 ชั่วโมงในอดีต (4,5,6 lag)

2. Day-ahead forecasting: ประกอบไปด้วย ANN $24 \times 2 = 48$ แบบจำลอง แต่ละแบบจำลองให้ค่าพยากรณ์ที่เวลา 00:00, 01:00 จนถึง 23:00 เป็นจำนวน 48 ค่า แบบจำลองมีโครงสร้าง 1 hidden layer ที่มี 7 neurons และ 1 target และใช้ข้อมูลขาเข้าดังนี้

   - $\mathrm{RH}(t)$
   - $\mathrm{T}(t-1), \mathrm{T}(t), \mathrm{T}(t+1)$
   - $I(t-1), I(t), I(t+1)$
   - $\mathrm{WS}(t)$

โดยที่หากตัวแปรมีตัวชี้เวลา t หรือ $t-k$ ใดๆ นั้นหมายความว่าจะใช้ *ค่าวัดจริง* ของตัวแปรนั้นๆ ซึ่งมาจากค่าที่เก็บในชื่อ `GWC actual value` หากใช้ตัวแปรใดที่มีตัวชี้เวลา $t+1$ นั้นหมายความว่า จะใช้ *ค่าประมาณ* ของตัวแปรนั้นๆ ซึ่งจะมาจากค่าเก็บในชื่อ `GWC forecast value`

สำหรับแบบจำลอง hour-ahead นั้น ในรูป 3.2 แสดงให้เห็นตัวอย่างการพยากรณ์ว่าในแต่ละเวลาพยากรณ์ (forecasting time) นั้น window การพยากรณ์จะขยับไปอย่างไร ในแบบจำลอง Provider นั้น เวลาที่พยากรณ์จะอยู่ในรูป HH:15 และทำในราย*ชั่วโมง* แต่ให้ค่าพยากรณ์ที่ละเอียดทางเวลาในรูป HH:00, HH:30 (ราย 30 นาที) ส่วนในรูป 3.3 แสดงการเปรียบเทียบว่า หากเราจะนำค่าวัดจริงกับค่าพยากรณ์มาคำนวณหาความคลาดเคลื่อนนั้น จะต้องดูตัวชี้เวลาอย่างไร และเราเลือกที่จะเปรียบเทียบแค่เวลาในช่วงกลางวันเมื่อมีแสง ตั้งแต่ 6:00 - 17:30 น. ตัวอย่างเช่น หากจะเปรียบเทียบค่าวัดจริงในเวลา 06:00 (แถวที่เป็น 06:00) เราต้องเทียบกับค่าพยากรณ์ที่มาจาก 2-step, 4-step, 6-step, 8-step ณ ตอนที่พยากรณ์เมื่อเวลา 5:15, 4:15, 3:15, 2:15 ตามลำดับ จาก configuration ของแบบจำลอง hour-ahead ของ Provider ที่จะให้ค่าพยากรณ์ราย 30 นาที แต่ทำการพยากรณ์รายชั่วโมงนั้น ทำให้เกิดการ mismatch ระหว่าง ตัวชี้เวลาของค่าวัดจริง กับตัวชี้วัดของค่าพยากรณ์ กล่าวคือ

Execution Time	7:30	8:00	8:30	9:00	9:30	10:00	10:30	11:00	11:30	12:00	12:30	13:00	13:30	14:00	14:30	15:00	15:30	16:00	16:30	17:00	17:30	18:00
1/9/2019 0:15	6.2	18.2	22.6	35	45.8	49	60.5	60.3	59.4	60	64.7	62.3	59.1	66.1	60.3	59.9	58.9	50.3	35.7	17.8	5.8	1.6
1/9/2019 1:15	6.2	18.2	22.6	35	45.8	49	60.5	60.3	59.4	60	64.7	62.3	59.1	66.1	60.3	59.9	58.9	50.3	35.7	17.8	5.8	1.6
1/9/2019 2:15	6.2	18.2	22.6	35	45.8	49	60.5	60.3	59.4	60	64.7	62.3	59.1	66.1	60.3	59.9	58.9	50.3	35.7	17.8	5.8	1.6
1/9/2019 3:15	5.4	15.9	28.9	34.7	47	48.6	59.1	59.7	55.1	60.2	63.8	57.4	64.7	67.4	59.7	61.2	60	49.6	34.8	18.1	6.3	0.9
1/9/2019 4:15	5.2	15.7	21.6	33.5	45.6	46.6	61.7	62.1	58.9	60.7	65.7	58.1	59.4	66.6	62.5	62.7	58.8	50.6	35.8	18.9	6.1	1.1
1/9/2019 5:15	5.2	15.7	21.6	33.5	45.6	46.6	61.7	62.1	58.9	60.7	65.7	58.1	59.4	66.6	62.5	62.7	58.8	50.6	35.8	18.9	6.1	1.1
1/9/2019 6:15	5.2	15.7	21.6	33.5	45.6	46.6	61.7	62.1	58.9	60.7	65.7	58.1	59.4	66.6	62.5	62.7	58.8	50.6	35.8	18.9	6.1	1.1
1/9/2019 7:15	4.2	14.5	18.9	31.2	44.8	44.7	59.1	58	53	55.7	59.7	48.5	48.5	54.7	47.4	48	46.5	43.8	30.2	17.4	7	0.3
1/9/2019 8:15	4.2	14.5	18.9	31.2	44.8	44.7	59.1	58	53	55.7	59.7	48.5	48.5	54.7	47.4	48	46.5	43.8	30.2	17.4	7	0.3
1/9/2019 9:15	4.2	14.5	18.9	31.2	44.8	44.7	59.1	58	53	55.7	59.7	48.5	48.5	54.7	47.4	48	46.5	43.8	30.2	17.4	7	0.3
1/9/2019 10:15	4.2	14.5	18.9	31.2	44.8	44.7	59.1	58	53	55.7	59.7	48.5	48.5	54.7	47.4	48	46.5	43.8	30.2	17.4	7	0.3
1/9/2019 11:15	4.2	14.5	18.9	31.2	44.8	44.7	59.1	58	53	55.7	59.7	48.5	48.5	54.7	47.4	48	46.5	43.8	30.2	17.4	7	0.3
1/9/2019 12:15	4.2	14.5	18.9	31.2	44.8	44.7	59.1	58	53	55.7	59.7	48.5	48.5	54.7	47.4	48	46.5	43.8	30.2	17.4	7	0.3
1/9/2019 13:15	4.2	14.5	18.9	31.2	44.8	44.7	59.1	58	53	55.7	59.7	48.5	48.5	54.7	47.4	48	46.5	43.8	30.2	17.4	7	0.3
1/9/2019 14:15	4.2	14.5	18.9	31.2	44.8	44.7	59.1	58	53	55.7	59.7	48.5	48.5	54.7	47.4	48	46.5	43.8	30.2	17.4	7	0.3
1/9/2019 15:15	4.2	14.5	18.9	31.2	44.8	44.7	59.1	58	53	55.7	59.7	48.5	48.5	54.7	47.4	48	46.5	43.8	30.2	17.4	7	0.3
1/9/2019 16:15	4.2	14.5	18.9	31.2	44.8	44.7	59.1	58	53	55.7	59.7	48.5	48.5	54.7	47.4	48	46.5	43.8	30.2	17.4	7	0.3
1/9/2019 17:15	4.2	14.5	18.9	31.2	44.8	44.7	59.1	58	53	55.7	59.7	48.5	48.5	54.7	47.4	48	46.5	43.8	30.2	17.4	7	0.3
1/9/2019 19:15	4.2	14.5	18.9	31.2	44.8	44.7	59.1	58	53	55.7	59.7	48.5	48.5	54.7	47.4	48	46.5	43.8	30.2	17.4	7	0.3
1/9/2019 20:15	4.2	14.5	18.9	31.2	44.8	44.7	59.1	58	53	55.7	59.7	48.5	48.5	54.7	47.4	48	46.5	43.8	30.2	17.4	7	0.3
1/9/2019 21:15	4.2	14.5	18.9	31.2	44.8	44.7	59.1	58	53	55.7	59.7	48.5	48.5	54.7	47.4	48	46.5	43.8	30.2	17.4	7	0.3
1/9/2019 22:15	4.2	14.5	18.9	31.2	44.8	44.7	59.1	58	53	55.7	59.7	48.5	48.5	54.7	47.4	48	46.5	44.4	30	17	5.9	0.7
1/9/2019 23:15	4.2	14.5	18.9	31.2	44.8	44.7	59.1	58	53	55.7	59.7	48.5	48.5	54.7	47.4	48	46.5	44.4	30	17	5.9	0.7

รูป 3.2: ตัวอย่างของค่าพยากรณ์ในแต่ k-step ในแบบจำลอง hour-ahead เช่น 1-step คือสีเหลือง 2-step คือสีเขียว 3-step คือสีฟ้า 4-step คือสีขมพู 5-step คือสีเขียว 6-step คือสีฟ้าอ่อน 7-step คือสีเทา 8-step คือสีแดง เป็นต้น จะเห็นว่าเวลาการคำนวณค่าพยากรณ์จะขยับรายชั่วโมง

- กลุ่มข้อมูล ณ เวลาค่าวัดจริง ณ 6:00, 7:00, ..., 17:00 จะ **ไม่มี** ค่าพยากรณ์จาก step ที่เป็น **เลขคี่** มาให้เทียบ
- กลุ่มข้อมูล ณ เวลาค่าวัดจริง ณ 6:30, 7:30, ..., 17:30 จะ **ไม่มี** ค่าพยากรณ์จาก step ที่เป็น **เลขคู่** มาให้เทียบ

จากข้อมูลนี้ ทำให้พบว่า สำหรับการคำนวณค่าคลาดเคลื่อนสำหรับ k-step หนึ่งๆ นั้น จะมีจำนวนข้อมูล 12 จุด/วัน เท่านั้น

Actual Time	1-step	2-step	3-step	4-step	5-step	6-step	7-step	8-step
6:00		5:15		4:15		3:15		2:15
6:30	6:15		5:15		4:15		3:15	
7:00		6:15		5:15		4:15		3:15
7:30	7:15		6:15		5:15		4:15	
8:00		7:15		6:15		5:15		4:15
8:30	8:15		7:15		6:15		5:15	
9:00		8:15		7:15		6:15		5:15
9:30	9:15		8:15		7:15		6:15	
10:00		9:15		8:15		7:15		6:15
10:30	10:15		9:15		8:15		7:15	
11:00		10:15		9:15		8:15		7:15
11:30	11:15		10:15		9:15		8:15	
12:00		11:15		10:15		9:15		8:15
12:30	12:15		11:15		10:15		9:15	
13:00		12:15		11:15		10:15		9:15
13:30	13:15		12:15		11:15		10:15	
14:00		13:15		12:15		11:15		10:15
14:30	14:15		13:15		12:15		11:15	
15:00		14:15		13:15		12:15		11:15
15:30	15:15		14:15		13:15		12:15	
16:00		15:15		14:15		13:15		12:15
16:30	16:15		15:15		14:15		13:15	
17:00		16:15		15:15		14:15		13:15
17:30	17:15		16:15		15:15		14:15	

รูป 3.3: การเปรียบเทียบเวลาของค่าวัดกำลังไฟฟ้ากับเวลาของค่าพยากรณ์ในแต่ละ lead time. เวลาในคอลัมน์ซ้ายสุดคือเวลาของค่าวัดจริง ส่วนในแต่ละคอลัมน์ที่เหลือ คือเวลาที่ทำการพยากรณ์ที่จะทำให้ผลการพยากรณ์มาเทียบกับค่าวัดจริงในแถวเดียวกัน

ในการวิเคราะห์ distribution ของความคลาดเคลื่อนค่าพยากรณ์นั้น ประเด็นแรกที่ต้องคำนึงถึงคือ เราจะต้องใช้ข้อมูลเหล่านั้นผ่านขั้นตอนประมาณ (estimation process) ซึ่งเป็นที่ทราบกันดีว่า จำนวน samples ที่ใช้นั้นส่งผลโดยตรงต่อคุณภาพการปริมาณ เราจึงควรทราบจำนวนข้อมูลที่จะได้จาก Provider ในการวิเคราะห์นี้ ประเด็นที่สองคือ ความคลาดเคลื่อนการพยากรณ์นั้นมักจะมีคุณสมบัติการกระจายตัวที่ต่างกันในแต่ละ lead time (หรือ k-step ahead) หรือการกระจายตัวที่ต่างกันในแต่ละ

แบบจำลอง ตัวอย่างเช่น ค่าพยากรณ์ ณ เวลา 1 ชั่วโมงล่วงหน้าควรจะมีความแปรปรวนต่ำกว่าค่าพยากรณ์ ณ เวลา 4 ชั่วโมงล่วงหน้า หรือ ในกรณีของ Provider models ค่าพยากรณ์ของเวลา 6:00 น. น่าจะมีการกระจายตัวที่ต่างกันกันค่าพยากรณ์ของเวลา 12:00 เนื่องจากใช้คนละแบบจำลองกัน ประเด็นที่สามคือ การพยากรณ์พลังงานไฟฟ้าจากแสงอาทิตย์นั้น ควรพิจารณาข้อมูลเฉพาะช่วงเวลาที่ค่าความเข้มแสงไม่เป็นศูนย์ เนื่องจาก หากรวมเวลาตอนกลางคืน ค่าความผิดพลาดการพยากรณ์ของเวลากลางคืนนั้นแทบเป็นศูนย์ (พยากรณ์ได้ดีมาก แต่เป็นกรณีที่ trivial) และจะทำให้ค่าเฉลี่ยของสมรรถนะรวมของทุกเวลามีค่าสูงกว่า การที่พิจารณาเฉพาะข้อมูลเวลากลางวัน (non-trivial forecasting) ดังนั้น ในงานวิจัยนี้ จะพิจารณาข้อมูลพยากรณ์ในช่วงเวลา

6:00, 6:30, 7:00,..., 17:00,17:30

เป็นจำนวน 24 จุดเวลาต่อวัน สอดคล้องกับแบบจำลองย่อยจำนวน 24 แบบจำลอง

เพื่อเป็นการเตรียมการวิเคราะห์ความคลาดเคลื่อนการพยากรณ์ตาม 3 ประเด็นดังข้างต้น เราจึงสามารถประมาณจำนวนข้อมูลที่แบ่งแยกตาม*เวลาของค่าพยากรณ์*ดังนี้

ตาราง 3.2: จำนวนข้อมูลผลการพยากรณ์ของ Provider จากโปรแกรม Nostradamus (สำหรับแบบจำลองย่อยของเวลาหนึ่งๆ)

horizon	**จำนวนข้อมูลต่อวัน**	**จำนวนข้อมูลต่อเดือน**	**จำนวนข้อมูลต่อปี**	**จำนวนข้อมูลระหว่างม.ค.62 - ก.พ. 63**
hour-ahead	4	120	1460	**1680**
day-ahead	7	210	2555	**2940**

หมายเหตุ สำหรับ hour-ahead model

1. การวิเคราะห์ มีจำนวนข้อมูล 4 จุดต่อเวลาหนึ่งๆ ใน 1 วัน ด้วยเหตุผลที่อธิบายดังรูป 3.2

2. การสรุปแยกวิเคราะห์คุณสมบัติความคลาดเคลื่อนของการพยากรณ์ ตามช่วงเวลาพยากรณ์ดังที่อธิบายข้างต้น จะพบว่าสำหรับแบบจำลอง hour-ahead ณ เวลาพยากรณ์หนึ่งๆ จะมีค่าพยากรณ์มาให้เปรียบเทียบเพียง 4 ค่าเท่านั้น ตามที่อธิบายในรูป 3.3 (ไม่ใช่ 8 ค่า จาก 8-step ahead predictions) ด้วยเหตุนี้ ผลการทดลองที่จะแสดงต่อไป จึงจะเปลี่ยนการเรียกลำดับของ k-step ใหม่ ให้ $k = 1, 2, 3, 4$ เท่านั้น และหมายความว่า

    - กลุ่มแบบจำลองย่อยของการพยากรณ์ ณ เวลา 6:00, 7:00, ..., 17:00 คำว่า 1-step, 2-step, 3-step, 4-step predictions ในผลการทดลอง จะหมายถึง การใช้ 2-step, 4-step, 6-step, 8-step predictions จากแบบจำลองของ Provider ที่ตั้งไว้
    - กลุ่มแบบจำลองย่อยของการพยากรณ์ ณ 6:30, 7:30, ..., 17:30 คำว่า 1-step, 2-step, 3-step, 4-step predictions ในผลการทดลอง จะหมายถึง การใช้ 1-step, 3-step, 5-step, 7-step predictions จากแบบจำลองของ Provider ที่ตั้งไว้

บทที่ 4

แนวทางการวิเคราะห์เชิงสถิติของความคลาดเคลื่อนในการพยากรณ์

การทดลองศึกษาความคลาดเคลื่อนการพยากรณ์นั้น จะมีแนวทางดังนี้

การเตรียมข้อมูล

1. จัดเตรียมข้อมูลค่าวัดกำลังไฟฟ้าจริง และค่าพยากรณ์จาก 2 โรงไฟฟ้าของเดือนมกราคม พ.ศ. 2562 - กุมภาพันธ์ พ.ศ. 2563 มาจัดในรูปแบบที่จะเปรียบเทียบกันได้ ทั้งในเชิงเปรียบเทียบ lead-time และในเชิงวิเคราะห์ค่าพยากรณ์ ณ เวลาที่สนใจ

2. ข้อมูลที่เป็นค่าพยากรณ์มีขาดหายไปในบางเวลา ในงานทดลองนี้จะไม่เติมข้อมูล (impute) แต่จะตัด record ดังกล่าวออกไปจากการวิเคราะห์

3. กำหนดให้ y และ $\hat{y}$ คือค่ากำลังไฟฟ้าผลิตที่วัดได้จริง และค่าพยากรณ์กำลังไฟฟ้าผลิตตามลำดับ เราคำนวณค่าความคลาดเคลื่อนจาก

$$e(t) = \hat{y}(t) - y(t)$$

เพื่อให้เครื่องหมายบวกและลบของความคลาดเคลื่อนนั้น บ่งชี้ถึง การพยากรณ์เกิน (over-estimate) และการพยากรณ์ขาด (underestimate) ตามลำดับ ค่า e ดังกล่าวนั้น จะ normalize ด้วยค่า installed capacity และมีหน่วยเป็น p.u. ในทางปฏิบัติ เราหวังว่า การพยากรณ์ที่ดีไม่น่าจะให้ค่า e มาก ดังนั้น e ไม่ควรเกิน ± 1 p.u. (เพราะนั่นหมายถึงว่าพยากรณ์ผิด 100%)

4. การวิเคราะห์ความคลาดเคลื่อนจะทำแยกในแต่ละเวลาพยากรณ์ การเตรียมข้อมูลพร้อมทั้ง time index จึงแบ่งตามเวลา 6:00,6:30,..., 17:30

5. การตรวจสอบสมรรถนะของ prediction interval จะมีหลักการที่ควรตรวจสอบบนข้อมูลชุดใหม่ ในงานนี้จึงใช้วิธี cross-validation ในขั้นตอนการแบ่งข้อมูลเป็น train และ test data sets นั้น จะใช้ 10-fold cross validation ทำให้ข้อมูลในส่วนการ train นั้น จะมี 90% และข้อมูลในส่วนการ test จะมี 10%

การประมาณฟังก์ชันการกระจายตัว

1. แบ่งกลุ่มของฟังก์ชันการกระจายตัวสามารถเป็น 2 กลุ่มอันได้แก่

    - kernel distribution
    - distribution ที่มี support เป็นเซตจำนวนจริง อันได้แก่ tLocationScale, stable, normal, logistic, extreme value, generalized extreme value

2. ประมาณฟังก์ชันการกระจายตัว ด้วยชุดคำสั่ง `fitdist` กับข้อมูลความคลาดเคลื่อนการพยากรณ์ของแต่ละเวลาตั้งแต่ 6:00,6:30,...,17:30 รวมทั้งหมด 24 กรณี เก็บค่าฟังก์ชันความเป็นไปได้ Komogorov-Smirnov (KS) test statistics, p-value

3. สำหรับ ณ เวลาพยากรณ์หนึ่งๆ เลือกฟังก์ชันการกระจายตัวที่เหมาะสมที่สุด โดยดูจากค่าฟังก์ชันความเป็นไปได้ที่สูงที่สุด ดูผล KS test ประกอบว่าผ่าน hypothesis หรือไม่

การประมาณช่วง PI

1. สำหรับข้อมูล batch หนึ่งๆ เราทดสอบการประมาณ confidence interval (CI) ของ e ด้วยวิธี bootstrap ที่อธิบายไว้ใน 2.2.3 โดยการตั้งค่าจำนวน bootstramp samples ไว้ที่ 1000 และคำนวณ percentiles ที่ 5 และที่ 95 ทำให้ nominal coverage อยู่ที่ 90% (พารามิเตอร์นี้ สามารถปรับได้) ผลลัพธ์ในขั้นตอนนี้คือ การได้ค่าประมาณของ CI ที่ขึ้นกับเวลาค่าพยากรณ์ และขึ้นกับ nominal coverage ค่าต่างๆ สมมติให้ช่วงนี้คือ $[l, u]$

2. จากความเชื่อมั่น α ค่าหนึ่งๆ เราสามารถหาช่วง CI ที่คิดว่า $l \leq e(t) \leq u$ ดังนั้น เมื่อพิจารณาสมการของ e จึงได้ว่าด้วยความเชื่อมั่นดังกล่าวนั้น ทำให้

$$\hat{y}(t) - u \leq y(t) \leq \hat{y}(t) - l$$

และ prediction interval สำหรับค่าพยากรณ์กำลังไฟฟ้าจึงเป็น $[\hat{y}(t) - u, \hat{y}(t) - l]$ ช่วง PI ดังกล่าว คำนวณแยกกันตามเวลาของค่าพยากรณ์

การตรวจสอบสมรรถนะของการพยากรณ์และการประมาณ PI

1. สำหรับการวิเคราะห์สมรรถนะของการพยากรณ์นั้น จะใช้ตัวชี้วัดคือ NRMSE และ NMBE ที่คำนวณแยกในแต่ละเวลาค่าพยากรณ์ และแยกกันในแต่ละ lead time

2. สำหรับการประมาณ PI เราสามารถทดสอบคุณสมบัติของช่วงดังกล่าวได้ บนข้อมูลใน test data set ด้วยการใช้ตัวชี้วัด PI coverage probability, PI normalized averaged width, reliability diagram ที่อธิบายในหัวข้อ 2.3 ตัวชี้วัดเหล่านี้ จะคำนวณแยกกันตามเวลาของค่าพยากรณ์ และจะเป็นค่าเฉลี่ยจากทุก 5 fold ที่ทำ cross validation

3. ทำซ้ำกระบวนการเตรียมข้อมูล และการประมาณช่วง PI สำหรับข้อมูลบน 2 โรงไฟฟ้า และแบบจำลอง day-ahead และ hour-ahead

4. สรุปผลการทดลองและวิเคราะห์สมรรถนะของการประมาณ PI

ในการแสดงผลการวิเคราะห์นั้น บทที่ 5 จะแสดงผลการทดลองจากทุกแบบจำลอง และจากทุกโรงไฟฟ้า แยกกรณีกัน 4 กรณี ส่วนในบทที่ 6 จะนำผลการทดลองมาเปรียบเทียบในเชิง การเปรียบเทียบระหว่างสองโรงไฟฟ้า และการเปรียบเทียบระหว่าง 2 แบบจำลอง อันได้แก่ day-ahead และ hour-ahead models

บทที่ 5

ผลการวิเคราะห์ความคลาดเคลื่อนการพยากรณ์

ในบทนี้ จะแสดงผลการทดลองวิเคราะห์สมบัติของความคลาดเคลื่อนการพยากรณ์ ตามประเด็นดังนี้

- การดู profile time series เบื้องต้นของความคลาดเคลื่อน ว่ามีผลเชิงฤดูกาล อย่างไร
- ตัวชี้วัดสมรรถนะของการพยากรณ์ (แบบ point forecasts) อันได้แก่ RMSE (เพื่อดูค่าเฉลี่ยโดยรวม) และ MBE (mean bias error) เพื่อดูว่าความคลาดเคลื่อนนั้นเป็นการพยากรณ์ขาดหรือเกิน
- คุณสมบัติเชิงสถิติพื้นฐานของความคลาดเคลื่อน อันได้แก่ ค่าเฉลี่ย ค่าแปรปรวน kurtosis (เป็น 4th moment เพื่อดูลักษณะ tail ของ distribution ว่าลู่เข้าศูนย์เร็วเพียงใด), skewness (เพื่อดูความสมมาตรของ distribution) และ histogram
- ผลการประมาณ confidence interval ของความคลาดเคลื่อน ด้วยวิธี bootstrap เพื่อดูว่า ความคลาดเคลื่อนในแต่ละเวลาพยากรณ์นั้น มีช่วงความเชื่อมั่นประมาณเท่าใด และช่วงกว้างต่างกันหรือไม่
- ผลการประมาณ prediction interval ของค่ากำลังไฟฟ้าผลิต และแสดงผลคู่กันกับค่าวัดจริง และค่าพยากรณ์ เพื่อแสดงตัวอย่างการนำช่วงดังกล่าวไปประยุกต์ใช้ประกอบการตีความ หลังได้ค่า point forecast ออกมา
- การตรวจสอบสมรรถนะของการประมาณ prediction interval จากวิธี bootstrap ด้วยตัวชี้วัด PI coverage probability (ควรจะมีค่าใกล้เคียงกับ nominal coverage ที่ใช้คำนวณ PI), PI normalized averaged width (ความกว้างของช่วง PI ยิ่งแคบยิ่งดี), reliability diagram (กราฟไม่ควรต่ำกว่ากราฟเส้นตรงความชัน $45°$)
- การวิเคราะห์ ramp rate ของกำลังไฟฟ้าที่ผลิตได้จริง และที่พยากรณ์ได้ กำหนดให้ $r(t) = \frac{p(t)-p(t-\Delta t)}{\Delta t}$ ที่จะแสดงผลในหน่วย 100% P.U./minute และการแสดงค่าความคลาดเคลื่อนของ ramp rate ว่าแปรตามช่วงเวลาอย่างไร (แสดงเฉพาะแบบจำลอง day-ahead เนื่องจากแบบจำลอง hour-ahead ของ Provider มี execution time resolution ที่ต่างจาก resolution ของเวลาพยากรณ์)

รายการผลการทดลองดังข้างต้นนั้น จะแสดงผลตามปัจจัยในด้าน*เวลาของค่าพยากรณ์* ว่าจะมีสมรรถนะในด้านต่างๆ แปรเปลี่ยนไปตามแบบจำลองย่อยอย่างไร เวลาดังกล่าวจะพิจารณาเฉพาะ

6:00, 6:30, ..., 17:00,17:30

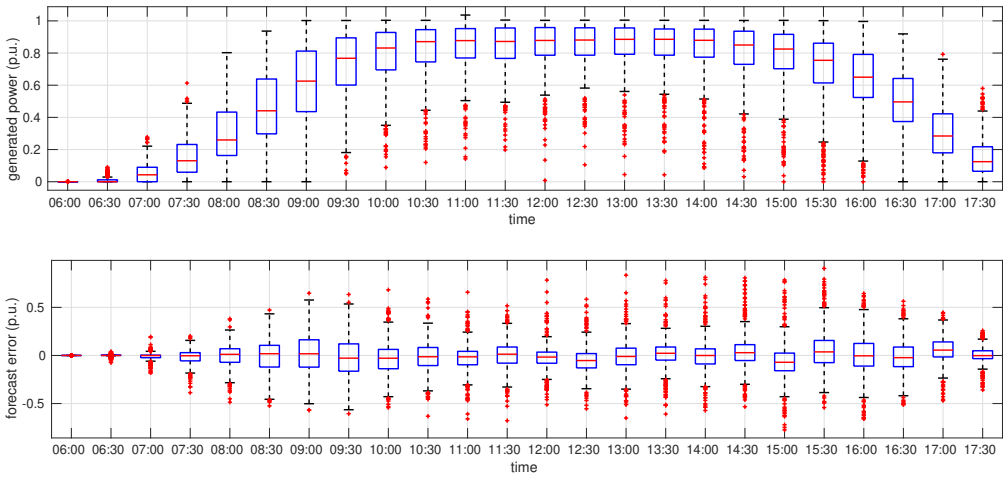
ผลการทดลองในบทนี้ แยกตามประเภทแบบจำลอง และโรงไฟฟ้า ดังนั้นจะมี 4 กรณีคือ

1. แบบจำลอง day-ahead ที่ใช้กับโรงไฟฟ้า A
2. แบบจำลอง day-ahead ที่ใช้กับโรงไฟฟ้า B
3. แบบจำลอง hour-ahead ที่ใช้กับโรงไฟฟ้า A

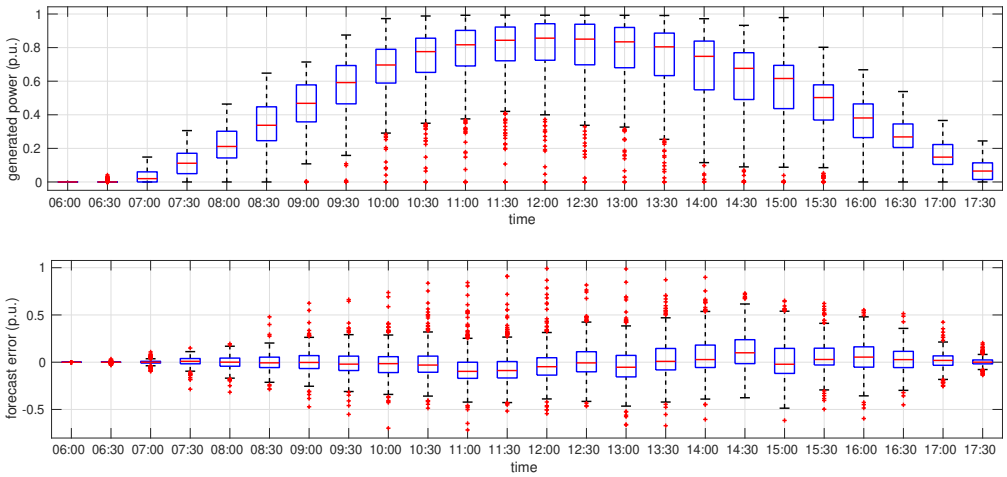
4. แบบจำลอง hour-ahead ที่ใช้กับโรงไฟฟ้า B

ในการแสดงผลการทดลองที่จะแยกตามแบบจำลอง และโรงไฟฟ้าดัง 4 กรณีข้างต้นนั้น เราจะแสดงผลการประมาณการกระจายตัวของค่าผิดพลาดพยากรณ์ ด้วยสองวิธี นั่นคือ 1) kernel density estimation 2) การมีสมมติฐานของฟังก์ชันการกระจายตัว และใช้วิธี MLE ในการประมาณพารามิเตอร์

ก่อนอื่นจะแสดงผลการกระจายตัวของค่ากำลังไฟฟ้าผลิตที่วัดได้จริงในแต่ละช่วงเวลาพยากรณ์ จากสองโรงไฟฟ้า ในรูป 5.1 พบว่าโรงไฟฟ้า A ค่ากำลังไฟฟ้าที่ผลิตได้ในช่วงเวลา 11:00 ถึง 14:30 มีการโดน saturate ไว้ที่ค่า installed capity (จนทำให้ solar generated power ใน boxplot มีช่วงถึงค่าเกือบ 1 p.u.) แต่ลักษณะการ saturation นี้ไม่พบในโรงไฟฟ้า B เราพบว่าค่ากำลังไฟฟ้าผลิตจริงของ A จะมีความแปรปรวนสูงในช่วงเวลา 8:00-9:00 และมี outliers เป็นจำนวนมากในช่วงเวลา 15:00-16:00 ส่วนค่ากำลังไฟฟ้าผลิตจริงของโรงไฟฟ้า B มีความแปรปรวนในช่วงเวลา 13:30-14:30 ทั้งสองโรงไฟฟ้ามีลักษณะร่วมกันคือ ในช่วงเวลาเช้าคือ 6:00-7:00 นั้น ค่ากำลังผลิตไฟฟ้าจะมีความแปรปรวนต่ำมาก เพราะส่วนใหญ่มีค่าเข้าใกล้ศูนย์



(a) โรงไฟฟ้า A

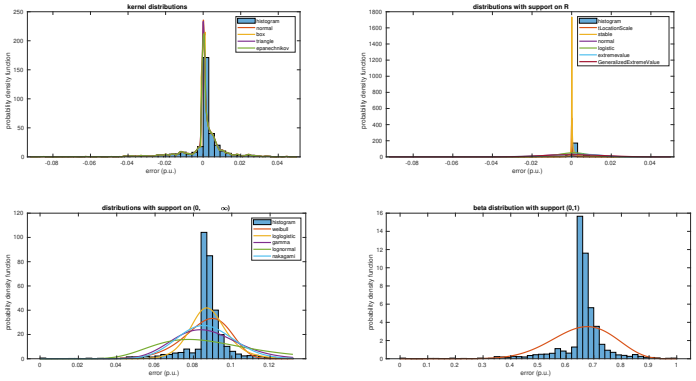


(b) โรงไฟฟ้า B

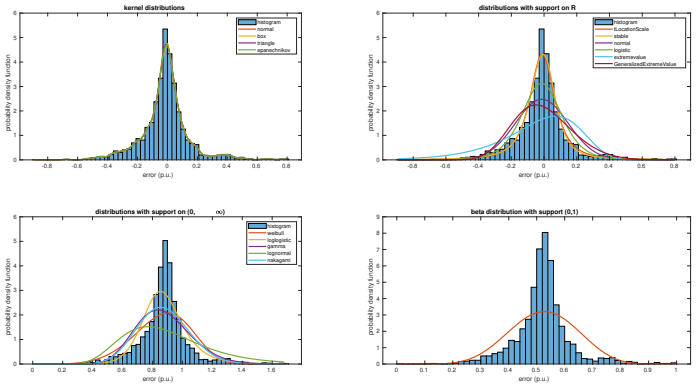
รูป 5.1: Box plot ของค่ากำลังไฟฟ้าที่ผลิตได้จากโรงไฟฟ้าทั้งสองแห่ง

5.1 ผลการประมาณฟังก์ชันการกระจายตัว

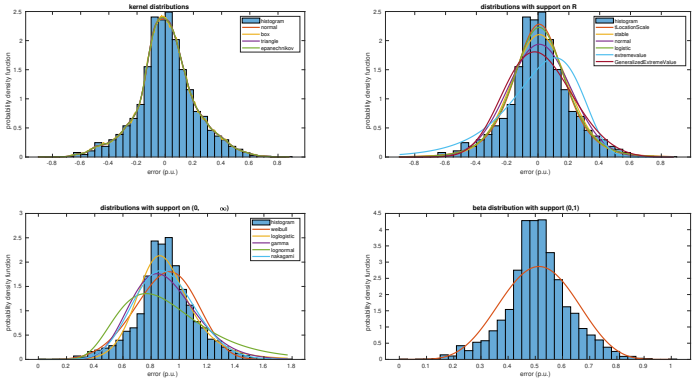
ในการประมาณว่าฟังก์ชันการกระจายตัวแบบใดจะเหมาะสมกับข้อมูลความคลาดเคลื่อนการพยากรณ์นั้น จะมีผลการทดลองที่แยกออกเป็น แบบจำลอง day-ahead/hour-ahead และแบบจำลองที่ใช้กับโรงไฟฟ้า A/B และแบบจำลองของการพยากรณ์ในแต่ละเวลาตั้งแต่ 6:00, 6:30, จนถึง 17:30 น. (จำนวนผลการประมาณเท่ากับ $24 \times 2 \times 2 = 96$ กรณี) ในหัวข้อนี้เราจึงแสดงตัวอย่างกราฟบางส่วนเท่านั้น จากแบบจำลอง day-ahead ที่ใช้กับสองโรงไฟฟ้า A สำหรับการพยากรณ์ 3 เวลา เราสามารถสรุปประเด็นจากผลการทดลองดัง รูปที่ 5.2 and 5.3 ได้ดังนี้ อันดับแรก การกระจายตัวของความคลาดเคลื่อนที่แต่ละเวลานั้น มีรูปร่างที่ต่างกัน ในช่วงเวลาเช้ามากเช่น 6:30 น. จะมีความคลาดเคลื่อนการพยากรณ์ต่ำ การกระจายตัวของข้อมูลจึงแคบ ส่วนข้อมูลช่วงเที่ยงถึงบ่าย การกระจายตัวของความคลาดเคลื่อนจะกว้าง



(a) การพยากรณ์ ณ เวลา 06:30 น.



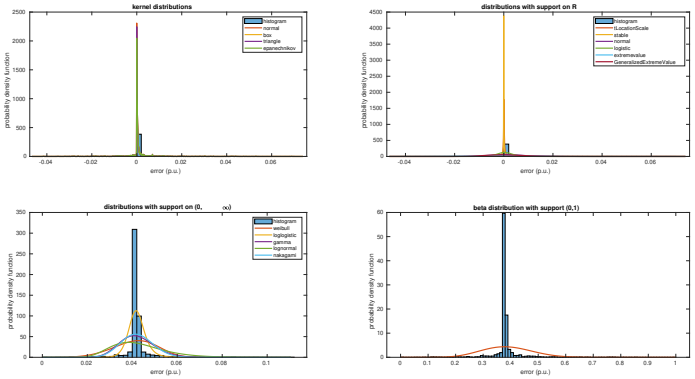
(b) การพยากรณ์ ณ เวลา 12:30 น.



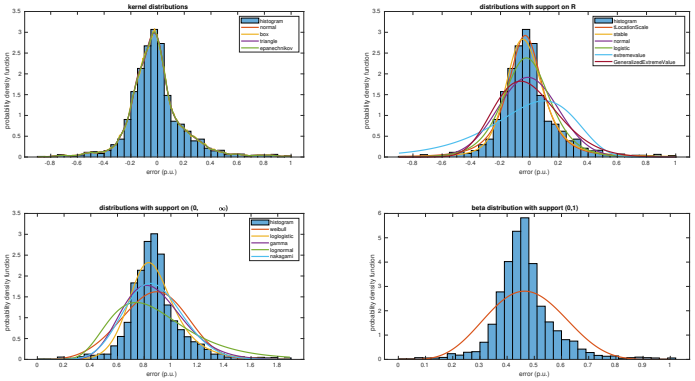
(c) การพยากรณ์ ณ เวลา 16:30 น.

รูป 5.2: ผลการประมาณฟังก์ชันการกระจายตัวของข้อมูลความคลาดเคลื่อนการพยากรณ์จากแบบจำลอง day-ahead ที่ใช้กับโรงไฟฟ้า A

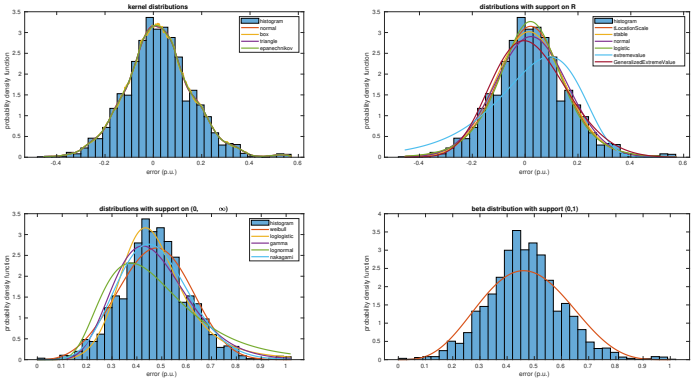
ประเด็นที่สอง ในรูปการประมาณของแต่ละเวลานั้น ได้ทดลองการประมาณ 4 กลุ่ม อันได้แก่ kernel density estimation (ซ้ายบน), distribution ที่มี support บนค่าจริง (ขวาบน), distribution ที่มี support บนค่าจริงบวก (ซ้ายล่าง) และ distribution ที่มี support บนช่วง $(0, 1)$ (ขวาล่าง) อันได้แก่ beta distribution ผลการทดลองพบว่า กรณีที่ให้ผลการประมาณที่ดี คือ kernel distribution และ distribution ที่มี support บนค่าจริง (เหมาะสมกับข้อมูลความคลาดเคลื่อนเนื่องจากมีค่าเป็นได้ทั้งบวกและลบ) ประเด็นที่สาม (ไม่ได้แสดงผลลัพธ์) kernel distribution ที่ประมาณได้ จะผ่านผล Komogorov-Smirnov test ในทุกกรณี แต่ fitted distribution ที่เป็นกลุ่มที่มี support บนค่าจริง นั้นไม่ผ่าน KS test จึงอาจกล่าวได้ว่า kernel distribution ให้ผลการประมาณที่ใกล้เคียงกับ empirical distribution มากกว่าวิธีอื่น อีกทั้งรูปผลการประมาณก็บ่งชี้ลักษณะเช่นเดียวกัน



(a) การพยากรณ์ ณ เวลา 06:30 น.



(b) การพยากรณ์ ณ เวลา 12:30 น.



(c) การพยากรณ์ ณ เวลา 16:30 น.

รูป 5.3: ผลการประมาณฟังก์ชันการกระจายตัวของข้อมูลความคลาดเคลื่อนการพยากรณ์จากแบบจำลอง day-ahead ที่ใช้กับโรงไฟฟ้า **B**

ประเด็นที่สี่ จากตาราง 5.1 จะพบว่า รูปแบบการกระจายตัวที่ให้ค่าความเป็นไปได้สูงสุดจะอยู่ในกลุ่มของ stable, t-location scale, generalized extreme value, logistic distributions ประเด็นที่สาม จากตาราง 5.2 ถึง 5.5 พบว่า

- ความคลาดเคลื่อนที่อธิบายด้วย stable distribution จะมีค่า shape parameter α ที่น้อยกว่า 2 บ่งชี้ ถึง heavy tails และมีค่า $\beta \neq 0$ อันบ่งชี้ถึง skewness (ความไม่สมมาตรของการกระจายตัว)
- ความคลาดเคลื่อนที่อธิบายด้วย logistic มักจะมีค่าพารามิเตอร์บ่งชี้ค่า mean ที่น้อย $\mu \in [-0.005, 0.09]$
- ความคลาดเคลื่อนที่อธิบายด้วย t-location scale จะมีค่าพารามิเตอร์ $\nu > 2$ (บ่งชี้ว่ามีค่าแปรปรวนที่จำกัด) มีค่า $\mu \in [-0.0001, 0.08]$ และมีค่า scale ที่อยู่ในช่วง $\sigma \in [0.01, 0.2]$ โดยที่ตัวแปร t-location scale มักจะใช้อธิบายข้อมูลที่มี heavy tails
- ความคลาดเคลื่อนที่อธิบายด้วย generalized extreme value distribution พบมากกว่าในข้อมูลที่มาจากโรงไฟฟ้า B โดยมีค่า shape parameter $k \neq 0$ ทุกกรณี ในกรณีย่อย $k < 0$ ซึ่งหมายถึง extreme value type III (หรือ Weibull) นั้น จะประมาณค่า μ ออกมาที่ติดลบ เกือบทุกกรณี เนื่องจากปกติ Weibull distribution ใช้อธิบายตัวแปรสุ่มที่เป็นบวก ส่วนกรณีย่อย $k > 0$ (ที่หมายถึง extreme value type II) นั้น พบเพียงกรณีเดียว และพารามิเตอร์ที่ประมาณได้มีค่า μ, σ ที่ต่ำกว่า เป็นกรณีของข้อมูลตอน 6:00 ซึ่งมีความแปรปรวนต่ำมาก

ตาราง 5.1: ฟังก์ชันการกระจายตัวของความคลาดเคลื่อนการพยากรณ์ที่ให้ค่าฟังก์ชันความเป็นไปได้สูงที่สุด เมื่อเปรียบเทียบจากทุกแบบจำลอง

Time	Day-ahead (A)	Day-ahead (B)	Hour-ahead (A)	Hour-ahead (B)
06:00	Stable	Generalized Extreme Value	Stable	Generalized Extreme Value
06:30	Stable	Stable	t Location-Scale	Stable
07:00	t Location-Scale	Stable	t Location-Scale	Stable
07:30	Stable	Stable	Stable	Stable
08:00	Logistic	Logistic	Stable	Stable
08:30	Logistic	Stable	Stable	Stable
09:00	t Location-Scale	t Location-Scale	t Location-Scale	Stable
09:30	Logistic	t Location-Scale	t Location-Scale	Stable
10:00	t Location-Scale	t Location-Scale	t Location-Scale	Stable
10:30	t Location-Scale	t Location-Scale	t Location-Scale	Stable
11:00	t Location-Scale	Stable	t Location-Scale	Stable
11:30	t Location-Scale	Stable	t Location-Scale	Stable
12:00	t Location-Scale	t Location-Scale	t Location-Scale	Stable
12:30	t Location-Scale	Stable	t Location-Scale	Stable
13:00	t Location-Scale	t Location-Scale	t Location-Scale	Stable
13:30	t Location-Scale	t Location-Scale	t Location-Scale	Stable
14:00	t Location-Scale	Stable	t Location-Scale	Stable
14:30	t Location-Scale	Logistic	Stable	Stable
15:00	t Location-Scale	Generalized Extreme Value	t Location-Scale	Generalized Extreme Value
15:30	Stable	Stable	t Location-Scale	Logistic
16:00	t Location-Scale	t Location-Scale	Logistic	t Location-Scale
16:30	Logistic	t Location-Scale	Logistic	Logistic
17:00	t Location-Scale	Logistic	t Location-Scale	Logistic
17:30	t Location-Scale	t Location-Scale	t Location-Scale	t Location-Scale

ตาราง 5.2: ฟังก์ชันการกระจายตัวของความคลาดเคลื่อนการพยากรณ์จากแบบจำลอง day-ahead ของโรงไฟฟ้า A

Time	Distribution	Parameters
06:00	Stable	$\alpha = 0.4, \beta = 1, \gamma = 4.1929e-06, \delta = 2.6019e-06$
06:30	Stable	$\alpha = 0.4, \beta = 0.51, \gamma = 0.00013669, \delta = 4.5094e-05$
07:00	t Location-Scale	$\mu = -0.002736, \sigma = 0.016723, \nu = 1.6052$
07:30	Stable	$\alpha = 1.6832, \beta = -0.63547, \gamma = 0.048378, \delta = -0.003194$
08:00	Logistic	$\mu = 0.01214, \sigma = 0.07782$
08:30	Logistic	$\mu = 0.03822, \sigma = 0.11175$
09:00	t Location-Scale	$\mu = 0.042076, \sigma = 0.1965, \nu = 10.0936$
09:30	Logistic	$\mu = -0.0011621, \sigma = 0.12637$
10:00	t Location-Scale	$\mu = -0.0053841, \sigma = 0.16088, \nu = 4.1296$
10:30	t Location-Scale	$\mu = -0.015545, \sigma = 0.15138, \nu = 5.2127$
11:00	t Location-Scale	$\mu = 0.0027938, \sigma = 0.10223, \nu = 2.1821$
11:30	t Location-Scale	$\mu = 0.009651, \sigma = 0.10911, \nu = 2.7875$
12:00	t Location-Scale	$\mu = -0.011254, \sigma = 0.081892, \nu = 1.9852$
12:30	t Location-Scale	$\mu = -0.041188, \sigma = 0.093347, \nu = 2.4136$
13:00	t Location-Scale	$\mu = -0.0017016, \sigma = 0.10935, \nu = 2.487$
13:30	t Location-Scale	$\mu = 0.012551, \sigma = 0.092978, \nu = 2.1178$
14:00	t Location-Scale	$\mu = -0.0091359, \sigma = 0.107, \nu = 2.3946$
14:30	t Location-Scale	$\mu = 0.037061, \sigma = 0.10622, \nu = 2.1486$
15:00	t Location-Scale	$\mu = -0.035211, \sigma = 0.16922, \nu = 4.2088$
15:30	Stable	$\alpha = 1.5774, \beta = 0.54817, \gamma = 0.13165, \delta = 0.041016$
16:00	t Location-Scale	$\mu = 0.016317, \sigma = 0.17267, \nu = 3.9078$
16:30	Logistic	$\mu = 0.007586, \sigma = 0.11222$
17:00	t Location-Scale	$\mu = 0.082743, \sigma = 0.13359, \nu = 4.7909$
17:30	t Location-Scale	$\mu = 0.012361, \sigma = 0.06258, \nu = 2.9983$

ตาราง 5.3: ฟังก์ชันการกระจายตัวของความคลาดเคลื่อนการพยากรณ์จากแบบจำลอง day-ahead ของโรงไฟฟ้า B

Time	Distribution	Parameters
06:00	Generalized Extreme Value	$k = 2.657, \sigma = 1.25e-10, \mu = 4.4875e-11$
06:30	Stable	$\alpha = 0.4, \beta = 0.12461, \gamma = 0.00010826, \delta = 9.113e-06$
07:00	Stable	$\alpha = 1.0313, \beta = -8.0854e-05, \gamma = 0.0085646, \delta = 0.00064233$
07:30	Stable	$\alpha = 1.7713, \beta = -0.72351, \gamma = 0.030546, \delta = 0.0074122$
08:00	Logistic	$\mu = 0.0039141, \sigma = 0.041712$
08:30	Stable	$\alpha = 1.8318, \beta = 0.5335, \gamma = 0.06343, \delta = -0.00066587$
09:00	t Location-Scale	$\mu = 0.011239, \sigma = 0.10297, \nu = 4.1055$
09:30	t Location-Scale	$\mu = -0.0095802, \sigma = 0.1098, \nu = 3.2389$
10:00	t Location-Scale	$\mu = -0.0036528, \sigma = 0.12546, \nu = 3.5117$
10:30	t Location-Scale	$\mu = -0.022799, \sigma = 0.10972, \nu = 2.6615$
11:00	Stable	$\alpha = 1.6256, \beta = 0.56355, \gamma = 0.11587, \delta = -0.077685$
11:30	Stable	$\alpha = 1.5167, \beta = 0.6453, \gamma = 0.10035, \delta = -0.08334$
12:00	t Location-Scale	$\mu = -0.030748, \sigma = 0.12419, \nu = 2.5968$
12:30	Stable	$\alpha = 1.677, \beta = 0.72208, \gamma = 0.10846, \delta = -0.011289$
13:00	t Location-Scale	$\mu = -0.042988, \sigma = 0.16228, \nu = 3.9468$
13:30	t Location-Scale	$\mu = 0.022442, \sigma = 0.17448, \nu = 4.2616$
14:00	Stable	$\alpha = 1.5104, \beta = 0.49687, \gamma = 0.1306, \delta = 0.027613$
14:30	Logistic	$\mu = 0.097059, \sigma = 0.11742$
15:00	Generalized Extreme Value	$k = -0.11229, \sigma = 0.19092, \mu = -0.06059$
15:30	Stable	$\alpha = 1.7531, \beta = 0.74898, \gamma = 0.10958, \delta = 0.020444$
16:00	t Location-Scale	$\mu = 0.051797, \sigma = 0.15616, \nu = 10.9065$
16:30	t Location-Scale	$\mu = 0.020734, \sigma = 0.1237, \nu = 10.4353$
17:00	Logistic	$\mu = 0.017723, \sigma = 0.051543$
17:30	t Location-Scale	$\mu = 0.00036933, \sigma = 0.032738, \nu = 3.0527$

ตาราง 5.4: ฟังก์ชันการกระจายตัวของความคลาดเคลื่อนการพยากรณ์จากแบบจำลอง hour-ahead ของโรงไฟฟ้า A

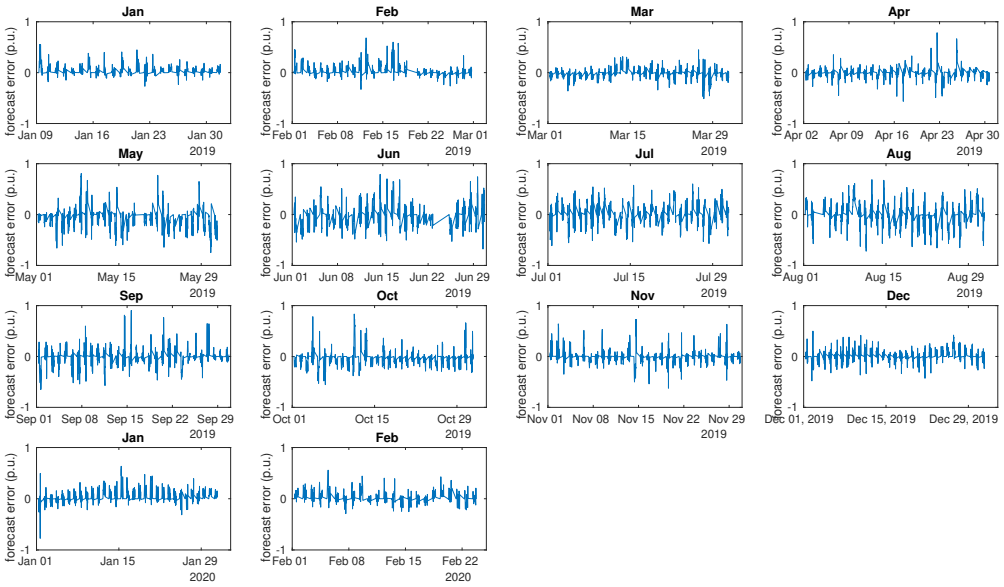
Time	Distribution	Parameters
06:00	Stable	$\alpha = 0.4, \beta = 0.99922, \gamma = 4.1929e-06, \delta = 2.6028e-06$
06:30	t Location-Scale	$\mu = -3.4669e-08, \sigma = 1e-06, \nu = 0.32255$
07:00	t Location-Scale	$\mu = 0.0011713, \sigma = 0.013186, \nu = 1.5616$
07:30	Stable	$\alpha = 1.6925, \beta = -0.69503, \gamma = 0.042989, \delta = 0.006884$
08:00	Stable	$\alpha = 1.7615, \beta = -0.87117, \gamma = 0.076399, \delta = 0.025006$
08:30	Stable	$\alpha = 1.8857, \beta = -0.79762, \gamma = 0.10127, \delta = 0.0089843$
09:00	t Location-Scale	$\mu = -0.024963, \sigma = 0.1634, \nu = 10.3832$
09:30	t Location-Scale	$\mu = -0.067168, \sigma = 0.13478, \nu = 4.4831$
10:00	t Location-Scale	$\mu = -0.066276, \sigma = 0.10902, \nu = 3.165$
10:30	t Location-Scale	$\mu = 0.012284, \sigma = 0.11345, \nu = 4.0833$
11:00	t Location-Scale	$\mu = -0.018589, \sigma = 0.084381, \nu = 2.5977$
11:30	t Location-Scale	$\mu = -0.017834, \sigma = 0.10548, \nu = 3.9808$
12:00	t Location-Scale	$\mu = -0.011573, \sigma = 0.082235, \nu = 2.6$
12:30	t Location-Scale	$\mu = -0.010036, \sigma = 0.08172, \nu = 2.6753$
13:00	t Location-Scale	$\mu = -0.049897, \sigma = 0.099623, \nu = 2.822$
13:30	t Location-Scale	$\mu = -0.024231, \sigma = 0.10587, \nu = 3.1706$
14:00	t Location-Scale	$\mu = -0.013598, \sigma = 0.087319, \nu = 2.2627$
14:30	Stable	$\alpha = 1.547, \beta = 0.39875, \gamma = 0.0894, \delta = -0.020802$
15:00	t Location-Scale	$\mu = -0.0036647, \sigma = 0.12343, \nu = 3.5149$
15:30	t Location-Scale	$\mu = 0.038626, \sigma = 0.15104, \nu = 3.5605$
16:00	Logistic	$\mu = 0.034604, \sigma = 0.11201$
16:30	Logistic	$\mu = 0.022527, \sigma = 0.10012$
17:00	t Location-Scale	$\mu = 0.027438, \sigma = 0.089397, \nu = 3.2188$
17:30	t Location-Scale	$\mu = 0.0075037, \sigma = 0.039633, \nu = 1.6578$

ตาราง 5.5: ฟังก์ชันการกระจายตัวของความคลาดเคลื่อนการพยากรณ์จากแบบจำลอง hour-ahead ของโรงไฟฟ้า B

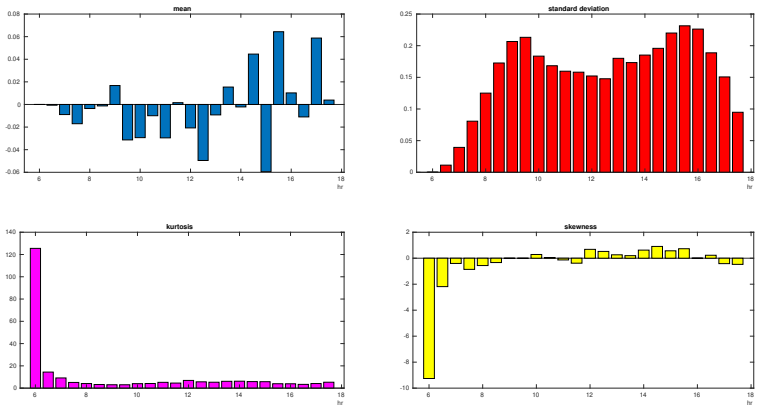
Time	Distribution	Parameters
06:00	Generalized Extreme Value	$k = 3.03, \sigma = 1.4467e-10, \mu = 4.6772e-11$
06:30	Stable	$\alpha = 0.41447, \beta = 0.52755, \gamma = 5.2911e-05, \delta = 1.8561e-05$
07:00	Stable	$\alpha = 0.98772, \beta = -0.053422, \gamma = 0.0064175, \delta = 0.0006938$
07:30	Stable	$\alpha = 1.8629, \beta = -0.51984, \gamma = 0.027727, \delta = -0.0044421$
08:00	Stable	$\alpha = 1.934, \beta = -1, \gamma = 0.043399, \delta = -0.0040867$
08:30	Stable	$\alpha = 1.7981, \beta = 0.44971, \gamma = 0.055816, \delta = -0.019139$
09:00	Stable	$\alpha = 1.6547, \beta = 0.6085, \gamma = 0.062638, \delta = -0.038595$
09:30	Stable	$\alpha = 1.6736, \beta = 0.68476, \gamma = 0.071383, \delta = -0.054522$
10:00	Stable	$\alpha = 1.4323, \beta = 0.53829, \gamma = 0.063347, \delta = -0.068151$
10:30	Stable	$\alpha = 1.7261, \beta = 0.63945, \gamma = 0.093826, \delta = -0.061399$
11:00	Stable	$\alpha = 1.6099, \beta = 0.64274, \gamma = 0.086384, \delta = -0.062391$
11:30	Stable	$\alpha = 1.5105, \beta = 0.77863, \gamma = 0.07771, \delta = -0.082769$
12:00	Stable	$\alpha = 1.7382, \beta = 1, \gamma = 0.098136, \delta = -0.048297$
12:30	Stable	$\alpha = 1.6945, \beta = 0.94777, \gamma = 0.10041, \delta = -0.080497$
13:00	Stable	$\alpha = 1.5477, \beta = 0.6227, \gamma = 0.095763, \delta = -0.048827$
13:30	Stable	$\alpha = 1.567, \beta = 0.78064, \gamma = 0.10769, \delta = 0.0043905$
14:00	Stable	$\alpha = 1.5341, \beta = 0.69512, \gamma = 0.10766, \delta = 0.004278$
14:30	Stable	$\alpha = 1.5585, \beta = 0.80379, \gamma = 0.10948, \delta = -0.00079484$
15:00	Generalized Extreme Value	$k = -0.13632, \sigma = 0.18053, \mu = -0.066836$
15:30	Logistic	$\mu = 0.02293, \sigma = 0.095512$
16:00	t Location-Scale	$\mu = 0.040057, \sigma = 0.14921, \nu = 79.3377$
16:30	Logistic	$\mu = 0.016541, \sigma = 0.055946$
17:00	Logistic	$\mu = 0.011632, \sigma = 0.039189$
17:30	t Location-Scale	$\mu = 0.012603, \sigma = 0.035333, \nu = 4.4815$

5.2 แบบจำลอง day-ahead ที่ใช้กับโรงไฟฟ้า A

เมื่อนำค่าความคลาดเคลื่อนของการพยากรณ์มาแสดงกราฟในแต่ละเดือนดังรูป 5.4 พบว่าในช่วงเดือนมกราคม-มีนาคม จะมีค่าไม่เกิน ± 0.5 p.u. และในเดือนพฤษภาคม-มิถุนายน ซึ่งเริ่มเข้าหน้าฝน ความคลาดเคลื่อนมีค่าสูงขึ้น กราฟในรูป 5.5 แสดงให้เห็นว่า สำหรับค่า mean ซึ่งบ่งชี้ว่าแบบจำลองย่อยของแต่ละเวลามี bias ต่างกันในแต่ละเวลา โดยจะ underestimate ในช่วงเช้าและเที่ยง และจะ overestimate ในตอนบ่าย ค่าความแปรปรวนนั้น จะมีค่าสูงในช่วง 8:30-10:00 และตอน 15:30-16:30 ค่า kurtosis มีค่าสูงมาก ที่ 6:00 และโดยรวมก็ยังมีค่าสูงกว่า 3 (ซึ่งเป็นค่า kurtosis ของ Gaussian distribution) ค่า skewness ที่ติดลบในช่วงเวลาเช้าบ่งชี้ว่า การกระจายตัวนั้นเป็นแบบ left-tailed

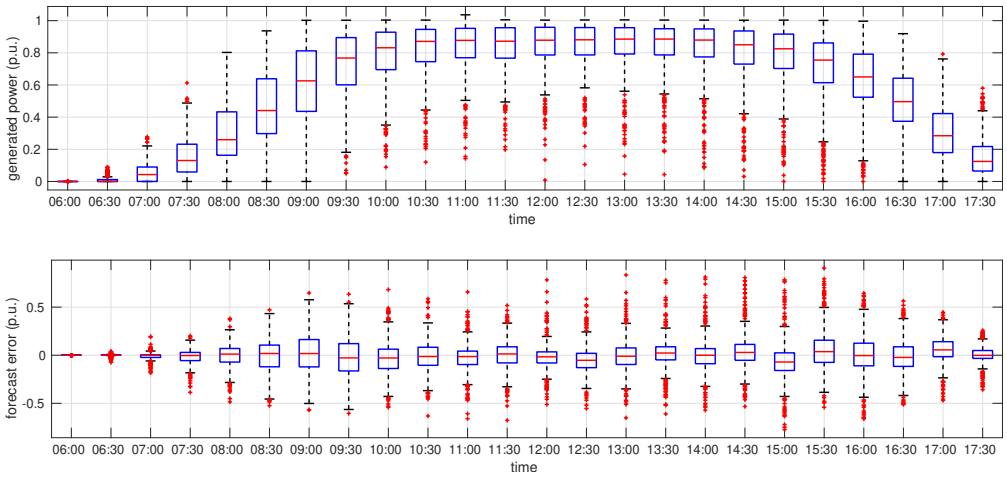


รูป 5.4: ความคลาดเคลื่อนในแต่ละเดือนของการพยากรณ์จากแบบจำลอง day-ahead ที่ใช้กับโรงไฟฟ้า A

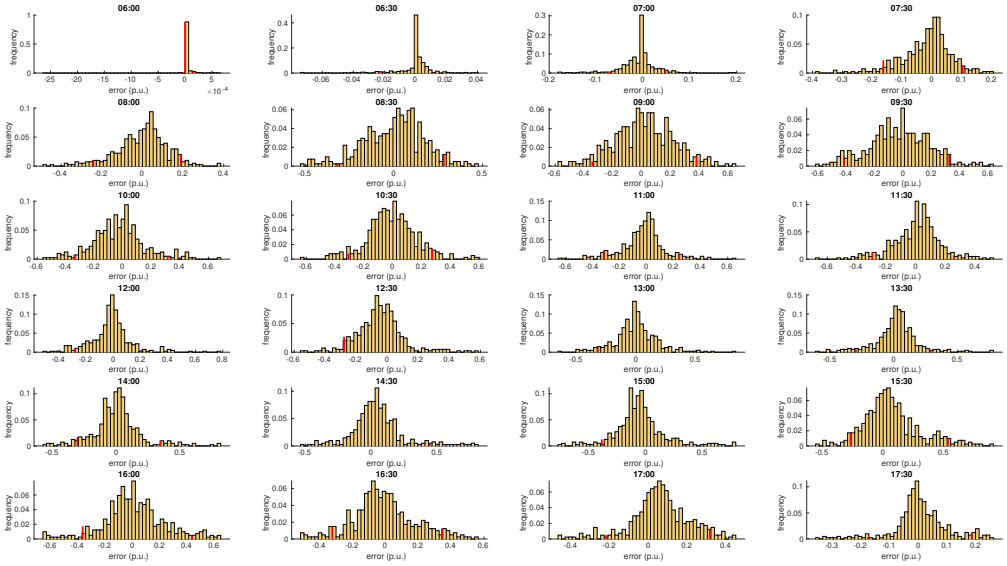


รูป 5.5: ค่าทางสถิติของความคลาดเคลื่อนของการพยากรณ์จากแบบจำลอง day-ahead ที่ใช้กับโรงไฟฟ้า A

เมื่อนำความคลาดเคลื่อนการพยากรณ์มาวิเคราะห์การกระจายตัวนั้น ในรูป 5.6 พบว่า กล่องสีน้ำเงินใน box plot ซึ่งแสดงถึง interquartile range (Q1-Q3) จะเป็นช่วงที่กว้างมากในเวลา 8:30, 9:00, 9:30, 16:00, 16:30 (ช่วงนี้ครอบคลุมความน่าจะเป็น 0.5) และสอดคล้องกับการแสดงผลด้วย histogram จุดสีดำใน box plot แสดงค่า extreme ที่คำนวณจาก median (ขีดสีแดง) นำมาบวกลบกับสัดส่วนของความกว้าง interquartile range โดยจะเห็นว่าค่า extreme ดังกล่าวจะกระโดดไปสูงในเวลา 8:30-9:30 ส่วนจุดสีแดงใน box plot แสดงถึง outliers ที่พบได้มากในช่วงเวลา 14:30-16:00 เมื่อพิจารณา histogram เราพบว่า เวลา 6:00-7:00 ความคลาดเคลื่อนจะมีการกระจุกตัวมากที่ค่าใกล้ๆ ศูนย์ และความคลาดเคลื่อนที่เวลา 9:30 จะมีการกระจายตัวที่ widespread มากที่สุด ลักษณะ shape ของ histogram นี้จะมีพารามิเตอร์ทางสถิติที่ต่างกันไปในแต่ละเวลา (เช่น ค่าเฉลี่ย ค่า median ความหนาของ tail ความสมมาตร เป็นต้น) ซึ่งสอดคล้องกับผลในรูป 5.5



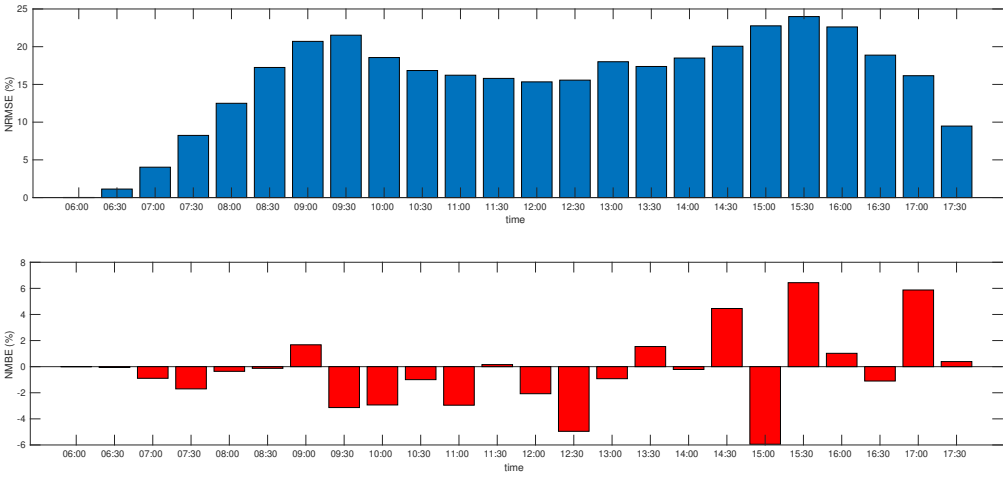
(a) box plot ของความคลาดเคลื่อน



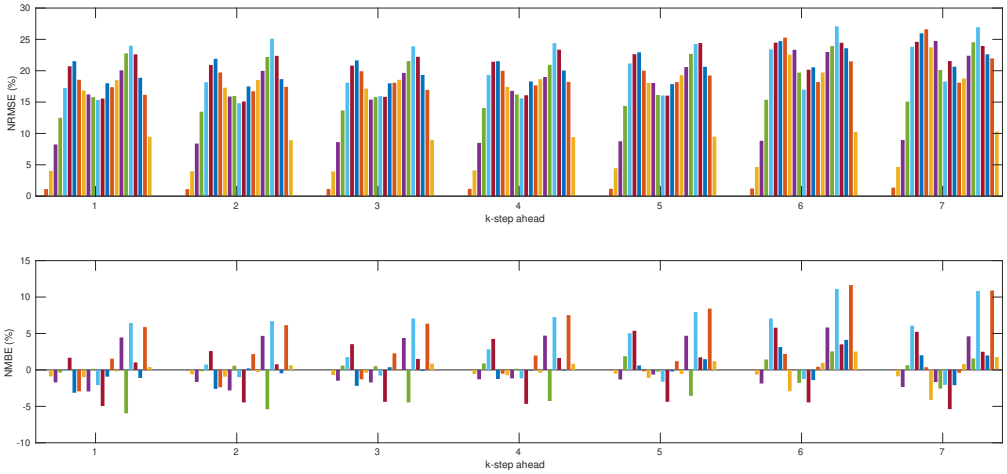
(b) histogram ของความคลาดเคลื่อน ช่วงสีแดงแสดงถึงช่วงความเชื่อมั่นด้วยความน่าจะเป็น 0.9

รูป 5.6: สมบัติการกระจายตัวของความคลาดเคลื่อนการพยากรณ์ด้วยแบบจำลอง day-ahead ของโรงไฟฟ้า A

สมรรถนะการพยากรณ์แสดงในรูป 5.7a เป็นการคำนวณจากพยากรณ์แบบ 1 วันล่วงหน้า พบว่าค่า NRMSE สูงประมาณ 22-24 % ณ เวลา 9:30, 15:00, 16:00 อันเป็นผลที่สอดคล้องกับความแปรปรวนของความคลาดเคลื่อนสูงในช่วงเวลานั้น แบบจำลองย่อยของเวลาเช้าพบว่า ส่วนใหญ่จะ underestimate เมื่อพิจารณาสมรรถนะการพยากรณ์ของ 7 วันล่วงหน้าในรูป 5.7b พบว่าแนวโน้มของค่าสมรรถนะในแต่ละเวลาพยากรณ์นั้นใกล้เคียงกัน กล่าวคือ เวลาที่ค่า NRMSE สูงจะเกิดขึ้นที่ช่วงเวลาเดียวกัน แบบจำลองมีแนวโน้มที่จะ overestimate มากขึ้นในการพยากรณ์ที่ 6 และ 7 วันล่วงหน้า (ซึ่งควรเป็นประเด็นที่ต้องตรวจสอบกับ GMC inputs ว่าเป็นสาเหตุหรือไม่) นอกจากนี้ การพยากรณ์ระยะ 5 วันล่วงหน้าเป็นต้นไป จะมีสมรรถนะที่แย่ลงอย่างเห็นได้ชัด การพยากรณ์ 7 วันล่วงหน้ามีค่า RMSE ที่สูงสุดที่ 28 % ณ 15:00 น.



(a) สมรรถนะการพยากรณ์ 1 วันล่วงหน้า

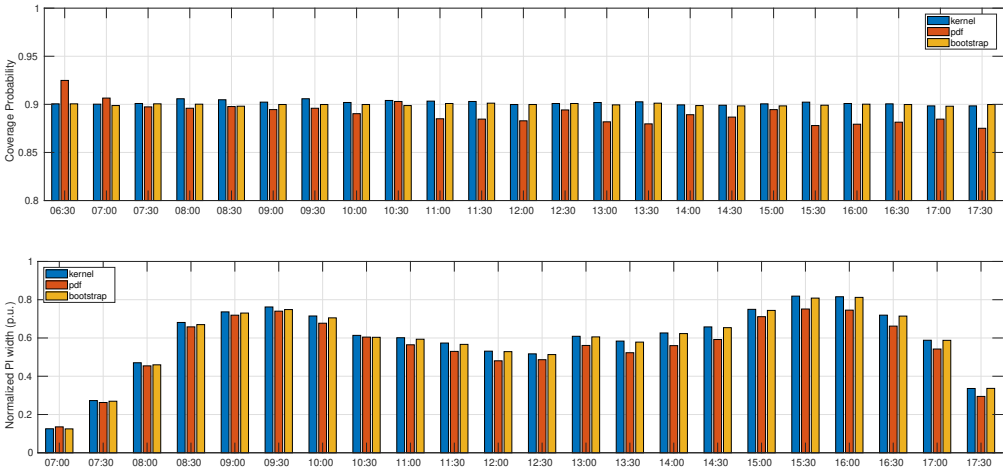


(b) สมรรถนะการพยากรณ์ 7 วันล่วงหน้า

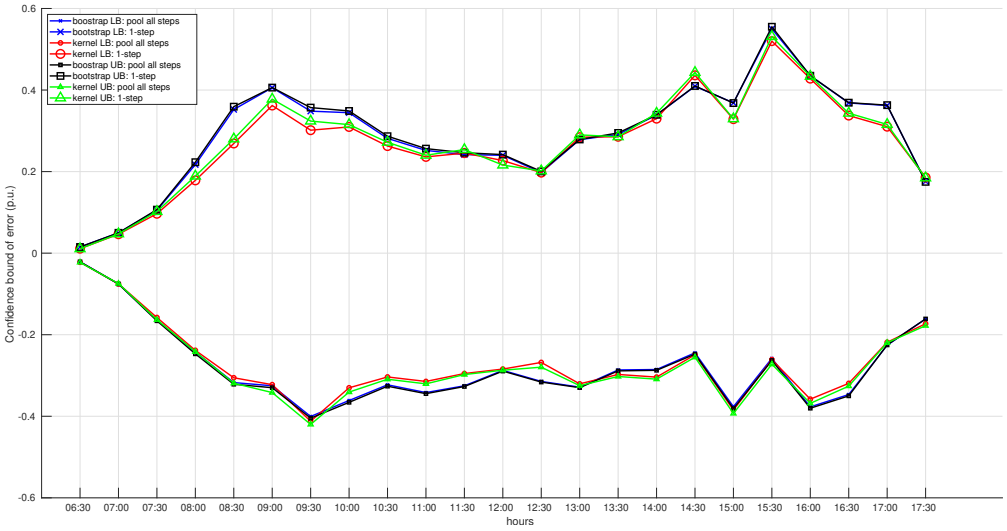
รูป 5.7: สมรรถนะการพยากรณ์ด้วยแบบจำลอง day-ahead ที่ใช้กับโรงไฟฟ้า A

จากการประมาณช่วงการทำนายค่าพยากรณ์ (PI) ด้วยสามวิธี อันได้แก่ kernel distribution, fitted distribution และ bootstrap ซึ่งใช้สัญลักษณ์ **`kernel, pdf, bootstrap`** ตามลำดับ ในรูป 5.8a แสดงความกว้างของช่วง PI ที่ normalized ให้ไม่เกิน 1 เราพบว่า PI ที่คำนวณจาก kernel และ bootstrap นั้นจะมีช่วงกว้างกว่า PI ที่คำนวณจาก fitted distribution เล็กน้อย และพบว่าช่วง PI มีความกว้างสูง ที่เวลา 8:00, 9:30, 15:30, 16:00,16:30 การประมาณช่วง PI ในรายงานนี้ ได้ตั้งค่าพารามิเตอร์ probability of coverage เป็น 0.9 (ซึ่งปรับได้) ในรูป 5.8a แสดงให้เห็นว่าเมื่อนำช่วง PI ดังกล่าวไปทดสอบกับข้อมูลใน test data set นั้น ค่า coverage probability ก็ใกล้เคียงกับ 0.9 ในเกือบทุกเวลาของค่าพยากรณ์ ยกเว้นวิธี fitted distribution ที่ให้ค่า coverage probability ที่ต่ำกว่า 0.9 ในหลายเวลา

รูป 5.8b เปรียบเทียบช่วงการทำนายที่คำนวณมาจากสองวิธี สำหรับเวลาหนึ่งๆ เช่น ณ เวลา 10:00 น. ค่าความคลาดเคลื่อนจะมีมาจากการพยากรณ์ 1-step,2-step,...,7-step ล่วงหน้า การคำนวณสองวิธีดังกล่าวคือ i) การรวมความคลาดเคลื่อนจากทุก step มาด้วยกันแล้วคำนวณ PI และ ii) การคำนวณ PI ของความคลาดเคลื่อนที่แยกตาม k-step เมื่อพิจารณาสองวิธีโดยที่วิธีหลังนั้นพิจารณา ที่ 1-step เราจะพบว่า PI ที่คำนวณจากวิธีแรกจะมีช่วงกว้างกว่าเล็กน้อย เนื่องจากความคลาดเคลื่อนย่อมมีการกระจายตัวที่สูงกว่า



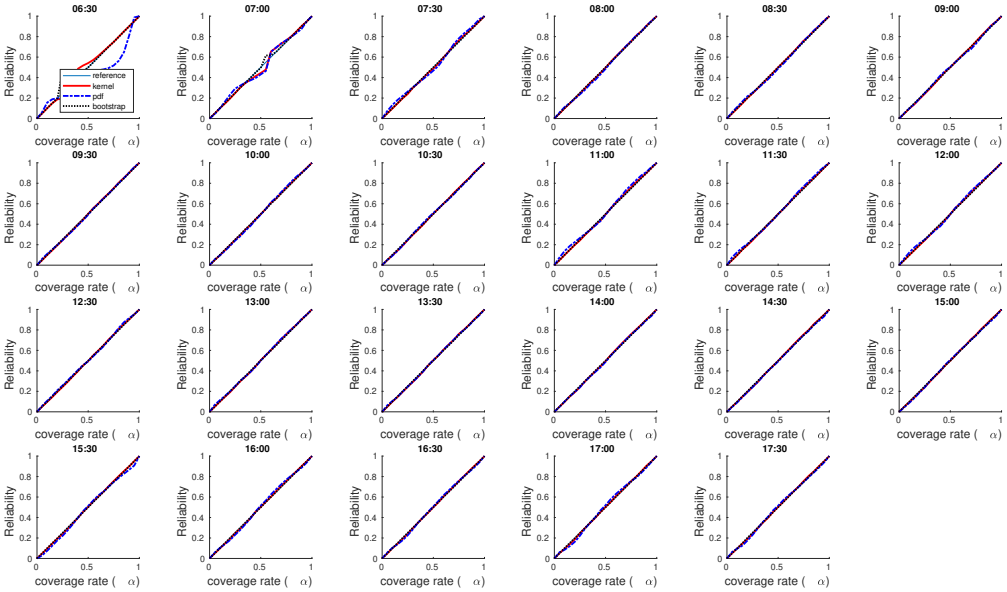
(a) Prediction interval normalized average width



(b) ช่วง PI จากการรวมความคลาดเคลื่อนทุก k-step และการวิเคราะห์จาก 1-step

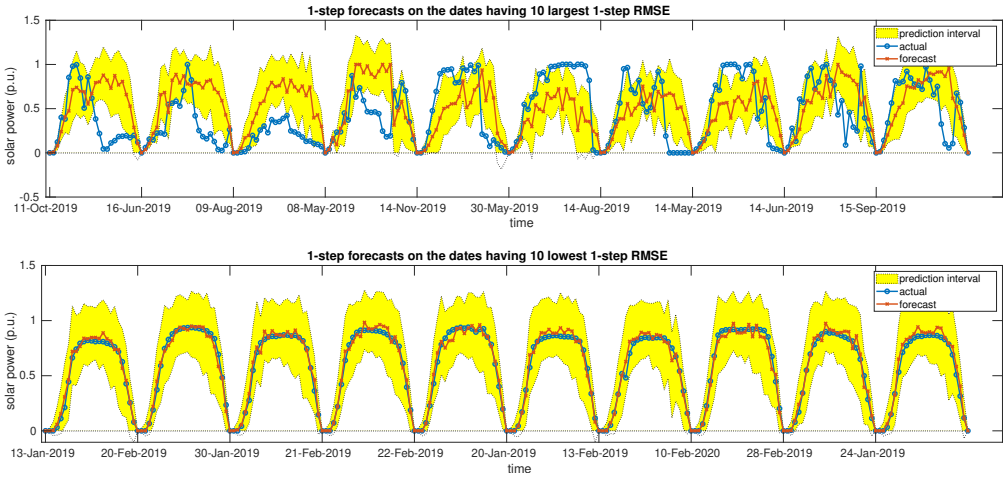
รูป 5.8: สมบัติช่วงการทำนายที่ประมาณได้ของแบบจำลอง day-ahead ของโรงไฟฟ้า A

จากรูป 5.9 พบว่า ในแต่ละเวลาพยากรณ์นั้น ช่วง PI ที่คำนวณมาจากสามวิธีนั้น มีค่า reliability ที่สอดคล้องกับค่า coverage rate (กราฟ reliability diagram ค่อนข้างเป็นเส้นตรงความชัน $45°$) ยกเว้นที่ข้อมูล ณ เวลา 6:00-6:30 น. พบว่า การคำนวณ coverage probability ของข้อมูลช่วงนั้น เกิดปัญหาเชิงเลข (numerical problem) และอีกทั้งเนื่องจากการกระจายตัวของข้อมูลเวลาดังกล่าวต่ำมาก จนทำให้ PI ที่ประมาณได้จากข้อมูลฝึกสอน ไม่สามารถครอบคลุมข้อมูลจากชุด test ได้



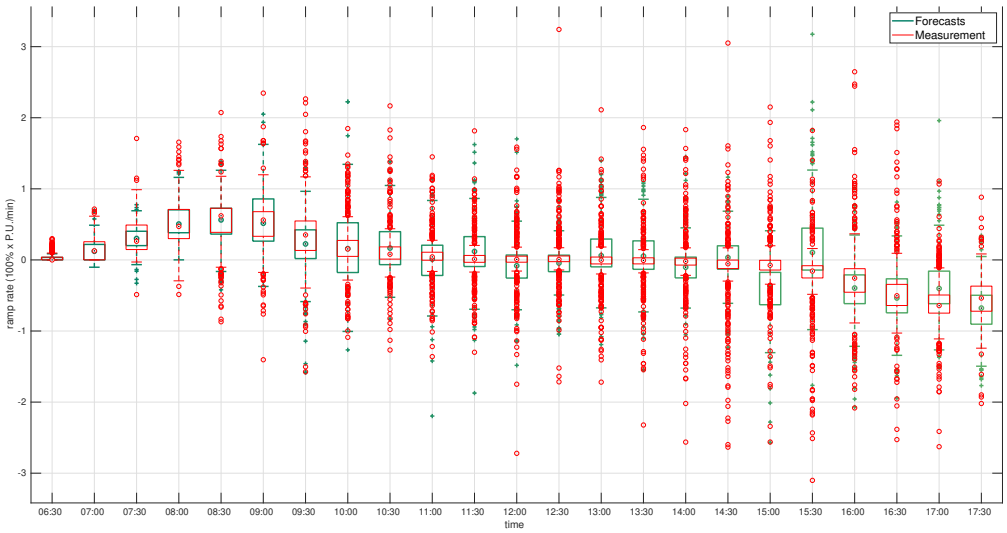
รูป 5.9: Reliability diagram (กราฟแต่ละเส้นคือผลที่คำนวณจากแต่ละ fold ใน cross validation) ที่ได้จากการตรวจสอบสมรรถนะของช่วง PI จากแบบจำลอง day-ahead ของโรงไฟฟ้า A

ในรูป 5.10 แสดงช่วงการทำนายที่คำนวณจากความคลาดเคลื่อนของการพยากรณ์ 1 step ล่วงหน้า ช่วงกล่าวแสดงใน shade สีเหลือง ของ 10 วันที่มีค่า NRMSE สูงสุดและต่ำสุด 10 ค่า และวันที่ที่แสดงของวันที่สมรรถนะแย่ก็เรียงตามค่า NRMSE จากสูงสุดไปยังค่าสูงสุดที่ 10 ส่วนวันที่ที่มีสมรรถนะดี ก็เรียงวันที่ตามค่า NRMSE ที่ต่ำสุดไปยังค่าที่ต่ำสุดที่ 10 เราสังเกตว่าวันที่สมรรถนะดีจะอยู่ในช่วง 2 เดือนแรกของต้นปี ส่วนวันที่สมรรถนะแย่จะกระจายในช่วงกลางปี ฤดูฝน ช่วงการทำนายนั้นตามหลักการแล้ว จะมีความน่าจะเป็น 0.9 (เป็นพารามิเตอร์ที่ตั้งไว้) ที่จะครอบคลุมค่ากำลังไฟฟ้าผลิตจริง (นั่นคือ ช่วงการทำนายก็เป็นค่าสุ่มค่าหนึ่ง เมื่อเรามีช่วงการทำนายหลายๆ samples นั้น จะมีสัดส่วน 0.9 ที่ช่วงการทำนายนั้นจะครอบคลุมค่ากำลังไฟฟ้าผลิตจริง) ในการประยุกต์ใช้จริง เราจะเห็นจากตัวอย่างกราฟว่า ในบางวันค่ากำลังผลิตจริง ก็อยู่นอกช่วงการทำนาย

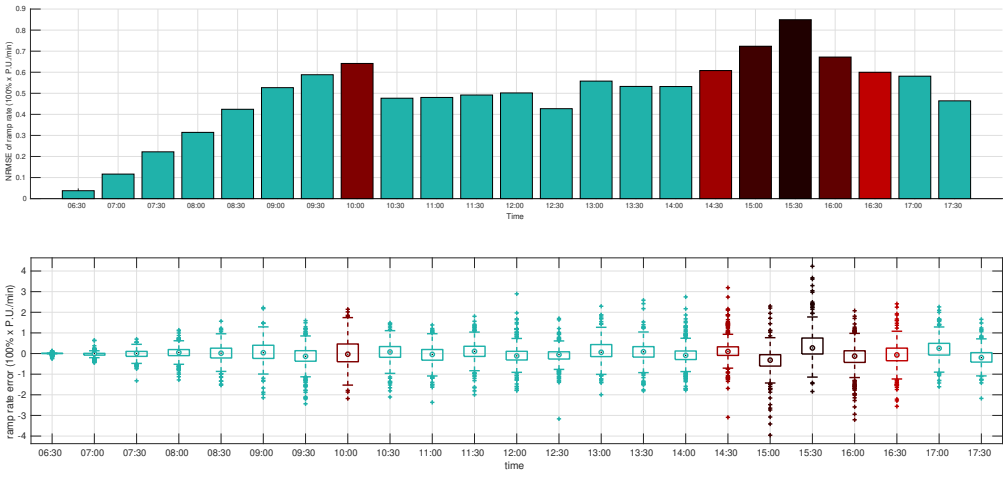


รูป 5.10: ตัวอย่างค่าพยากรณ์ และช่วงการทำนายจากวิธี bootstrap ของแบบจำลอง day-ahead ของโรงไฟฟ้า A

รูป 5.11a แสดง box plot ของ ramp rate ของค่ากำลังไฟฟ้าวัดจริงและค่าพยากรณ์ ที่เห็นได้ว่าค่า median ของ ramp rate ในช่วงเช้าถึง 10:00 นั้น มีค่าเป็นบวก และจะแกว่งไปมาค่าในช่วง 10:00-15:30 น. ในช่วงกลางวันดังกล่าวจะพบว่ามี outliers ของ ramp rate เป็นปริมาณที่สูงมาก ramp rate จะมีค่าเป็นลบอีกครั้งในช่วงหลัง 15:30 น. ค่า median ของ ramp rate ของค่าพยากรณ์ในช่วงบ่ายนั้นมักจะไม่ค่อยมีค่าใกล้เคียงกับ ramp rate ของค่ากำลังไฟฟ้าวัดจริง รูป 5.11b แสดงให้เห็นค่า NRMSE ของ ramp rate และเน้นจุดเวลาที่มีค่า NRMSE สูงสุด 6 ค่า ที่พบว่าในช่วงเวลา 14:30-16:30 น. นั้นมีความคลาดเคลื่อนของ ramp rate สูงมาก ซึ่งเมื่อดู box plot ประกอบก็จะเห็นว่าช่วง inter-quartile ณ เวลาดังกล่าวมีช่วงที่กว้างนั่นเอง



(a) Ramp rate of measured and forecasted solar power.

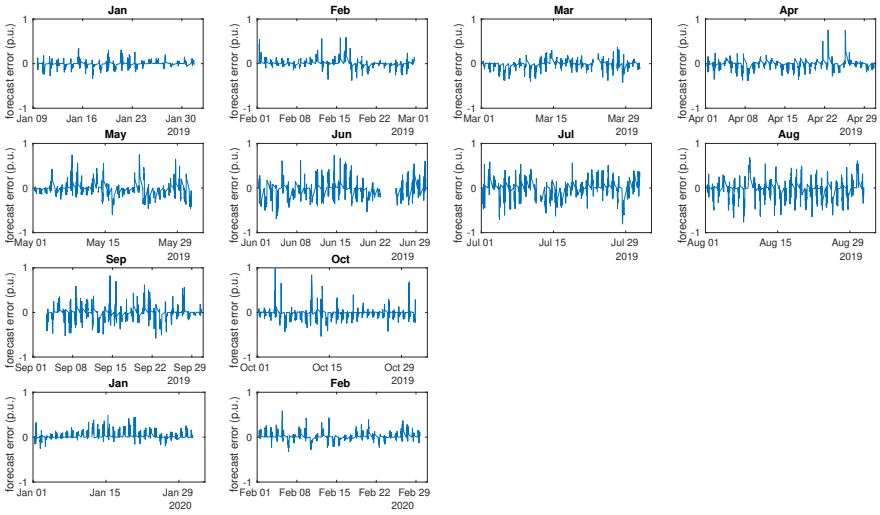


(b) Ramp rate error (the darker tone the higher NRMSE).

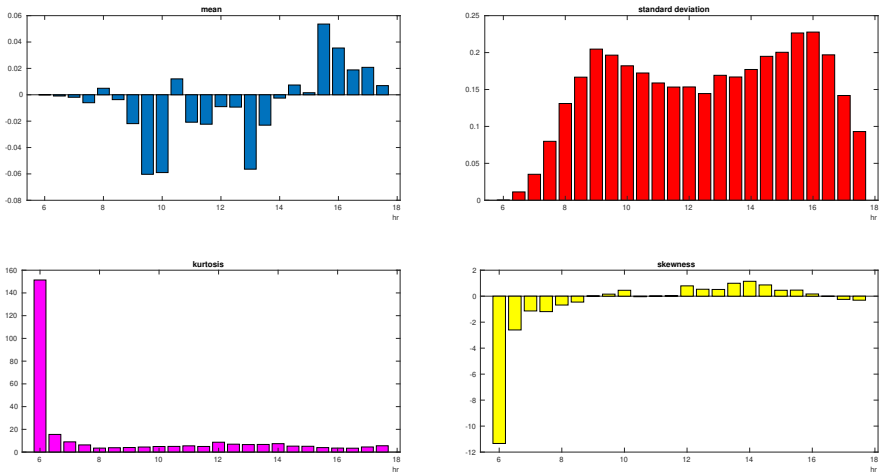
รูป 5.11: สมรรถนะของ ramp rate ที่ได้จากแบบจำลอง day-ahead ของโรงไฟฟ้า A

5.3 แบบจำลอง hour-ahead ที่ใช้กับโรงไฟฟ้า A

เมื่อนำค่าความคลาดเคลื่อนของการพยากรณ์มาแสดงกราฟในแต่ละเดือนดังรูป 5.12 พบว่าในช่วงเดือนในเดือนพฤษภาคม-สิงหาคม ซึ่งเริ่มเข้าหน้าฝน สัดส่วนของความคลาดเคลื่อนค่าสูงจะมีมากขึ้น กราฟในรูป 5.13 แสดงให้เห็นว่า สำหรับค่า mean ซึ่งบ่งชี้ว่าแบบจำลองย่อยของแต่ละเวลามี bias ต่างกันในแต่ละเวลา โดยจะ underestimate ในช่วงเช้าและเที่ยง และจะ over-estimate ในตอนเย็น ค่าความแปรปรวนนั้น จะมีค่าสูงในช่วง 8:30-10:00 และตอน 15:30-16:30 ค่า kurtosis มีค่าสูงมาก ที่ 6:00 และโดยรวมก็ยังมีค่าสูงกว่า 3 (ซึ่งเป็นค่า kurtosis ของ Gaussian distribution) ค่า skewness ที่ติดลบในช่วงเวลาเช้าบ่งชี้ว่า การกระจายตัวนั้นเป็นแบบ left-tailed

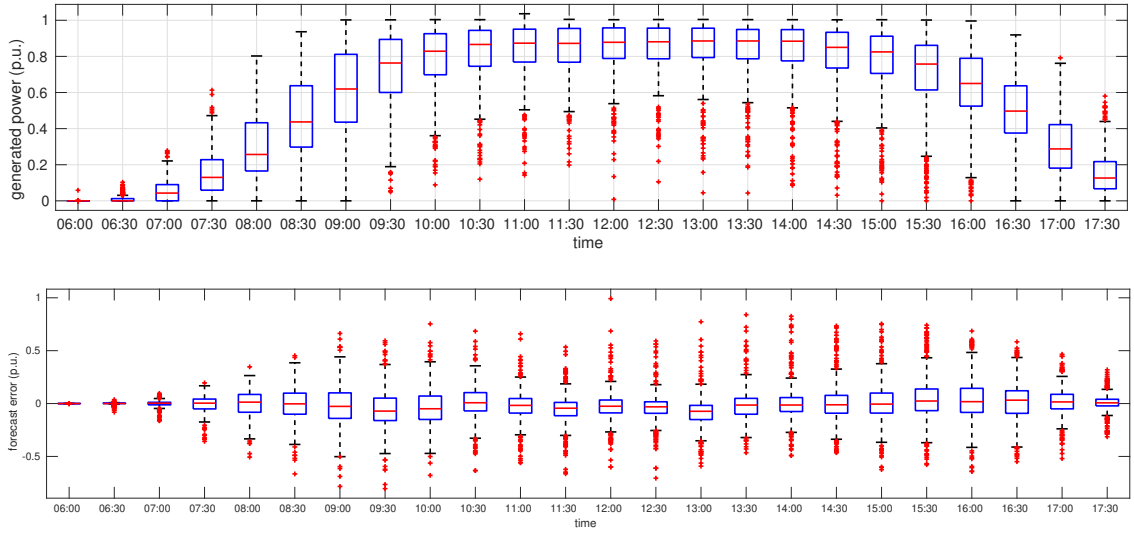


รูป 5.12: ความคลาดเคลื่อนในแต่ละเดือนของการพยากรณ์จากแบบจำลอง hour-ahead ที่ใช้กับโรงไฟฟ้า A

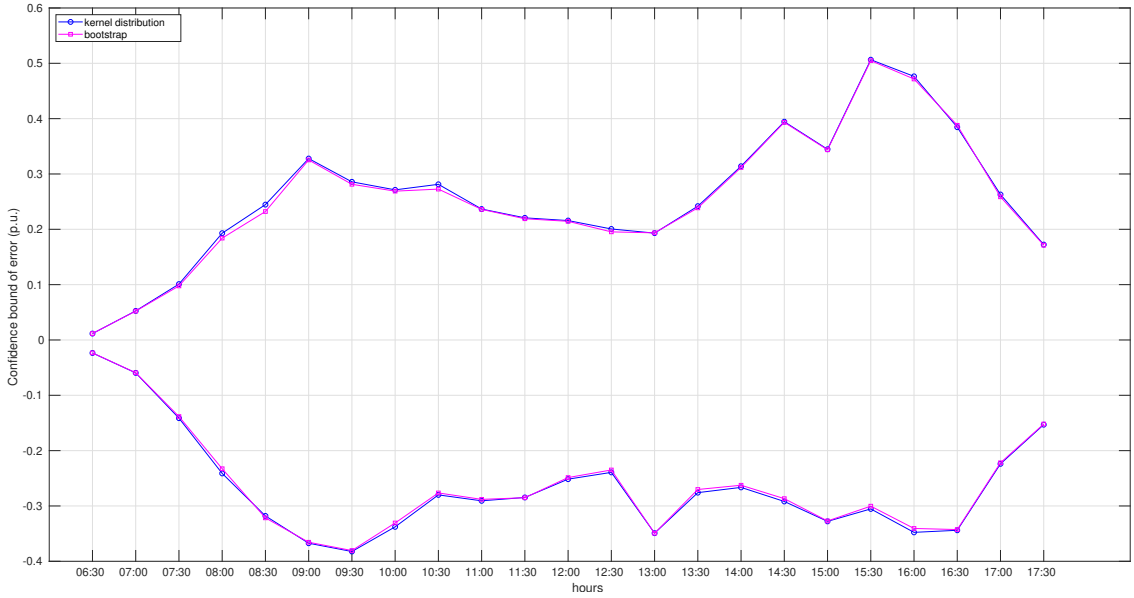


รูป 5.13: ค่าทางสถิติของความคลาดเคลื่อนของการพยากรณ์จากแบบจำลอง hour-ahead ที่ใช้กับโรงไฟฟ้า A

เมื่อนำความคลาดเคลื่อนการพยากรณ์มาวิเคราะห์การกระจายตัวนั้น ในรูป 5.14 พบว่า กล่องสีน้ำเงินใน box plot ซึ่งแสดงถึง interquartile range (Q1-Q3) จะเป็นช่วงที่กว้างมากในเวลา 8:30, 9:00, 9:30, 15:30-16:30 (ช่วงนี้ครอบคลุมความน่าจะเป็น 0.5) และสอดคล้องกับการแสดงผลด้วย histogram จุดสีดำใน box plot แสดงค่า extreme ที่คำนวณจาก median (ขีดสีแดง) นำมาบวกลบกับสัดส่วนของความกว้าง interquartile range โดยจะเห็นว่าค่า extreme ดังกล่าวจะกระโดดไปสูงในเวลา 8:30, 15:30-16:30 ส่วนจุดสีแดงใน box plot แสดงถึง outliers ที่พบได้มากในช่วงเวลา 14:00-16:00 เมื่อพิจารณา histogram เราพบว่า เวลา 6:00-7:00 ความคลาดเคลื่อนจะมีการกระจุกตัวมากที่ค่าใกล้ๆ ศูนย์ และความคลาดเคลื่อนที่เวลา 8:00-10:00 จะมีการกระจายตัวที่ widespread มากที่สุด ลักษณะ shape ของ histogram นี้จะมีพารามิเตอร์ทางสถิติที่ต่างกันไปในแต่ละเวลา (เช่น ค่าเฉลี่ย ค่า median ความหนาของ tail ความสมมาตร เป็นต้น) ซึ่งสอดคล้องกับผลในรูป 5.13



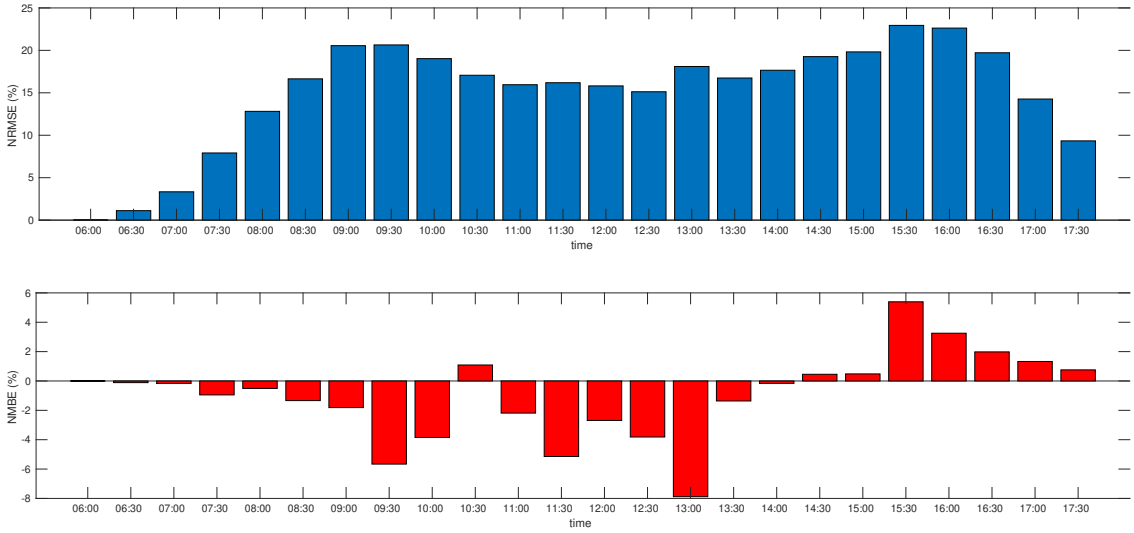
(a) box plot ของความคลาดเคลื่อน



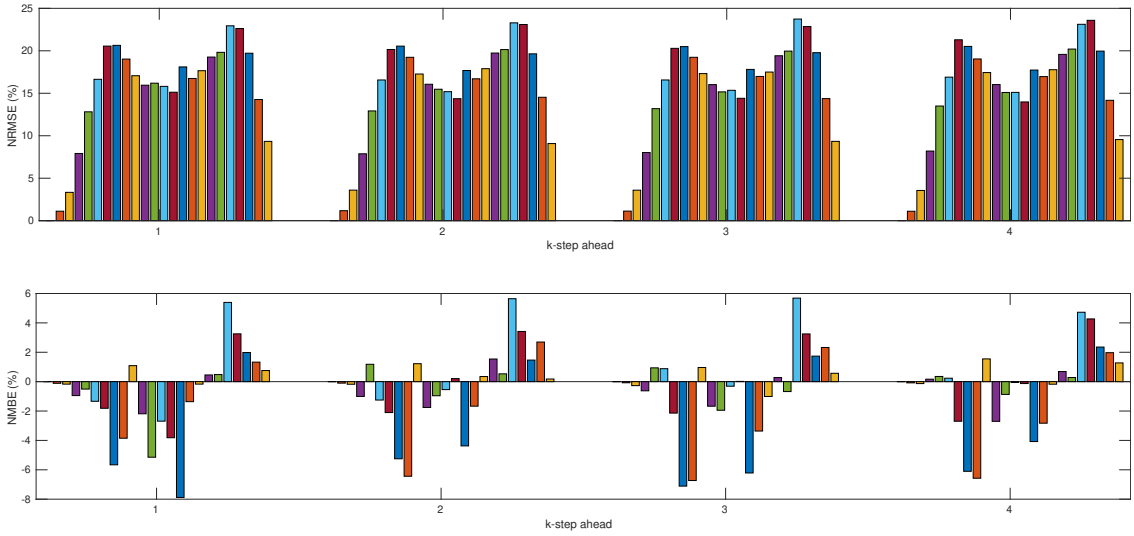
(b) histogram ของความคลาดเคลื่อน ช่วงสีแดงแสดงถึงช่วงความเชื่อมั่นด้วยความน่าจะเป็น 0.9

รูป 5.14: สมบัติการกระจายตัวของความคลาดเคลื่อนการพยากรณ์ด้วยแบบจำลอง hour-ahead ของโรงไฟฟ้า A

สมรรถนะการพยากรณ์แสดงในรูป 5.15a เป็นการคำนวณจากพยากรณ์แบบ 1-step ล่วงหน้า พบว่าค่า NRMSE สูงประมาณ 20-23 % ณ เวลา 9:30, 15:30, 16:00 อันเป็นผลที่สอดคล้องกับความแปรปรวนของความคลาดเคลื่อนสูงในช่วงเวลานั้น แบบจำลองย่อยของเวลาเช้าจนถึง 13:30 พบว่า ส่วนใหญ่จะ underestimate เมื่อพิจารณาสมรรถนะการพยากรณ์ของ 4-step ล่วงหน้าในรูป 5.15b พบว่าแนวโน้มของค่าสมรรถนะในแต่ละเวลาพยากรณ์นั้นใกล้เคียงกัน กล่าวคือ เวลาที่ค่า NRMSE สูงจะเกิดขึ้นที่ช่วงเวลาเดียวกัน การพยากรณ์ 1-step, 2-step, 3-step, 4-step ผลโดยรวมไม่ได้ต่างกันมากนัก



(a) สมรรถนะการพยากรณ์ 1 step ล่วงหน้า

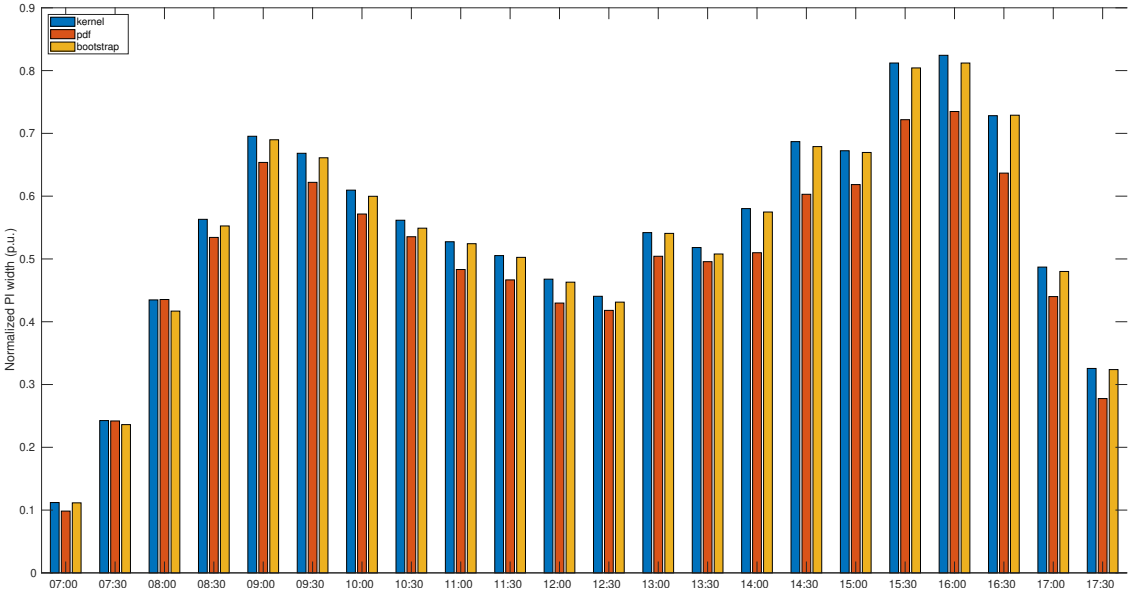


(b) สมรรถนะการพยากรณ์ 4 step ล่วงหน้า

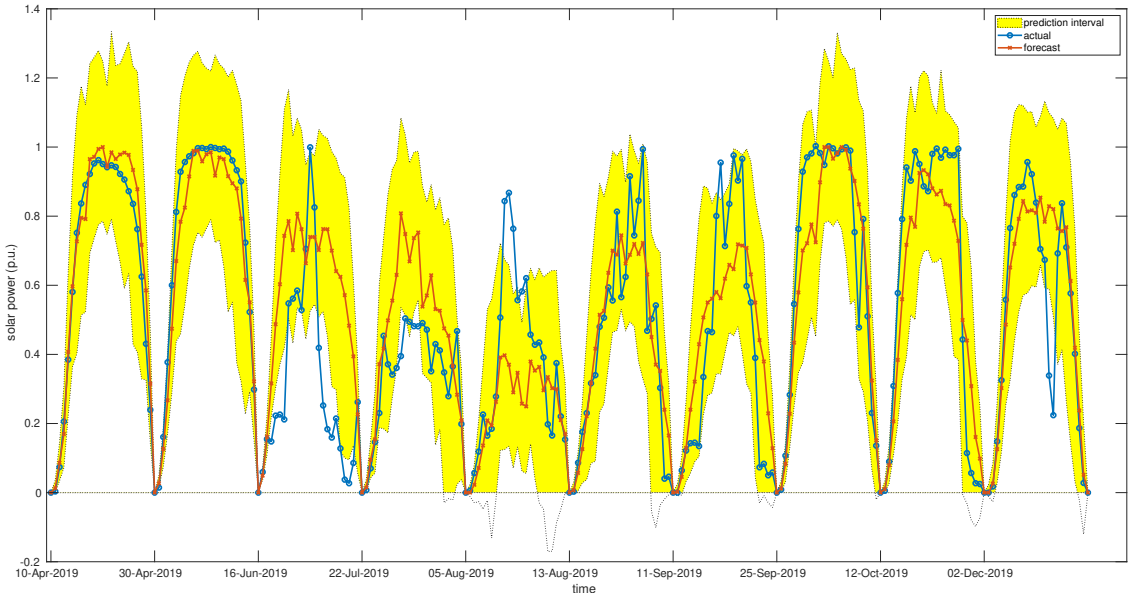
รูป 5.15: สมรรถนะการพยากรณ์ด้วยแบบจำลอง hour-ahead ที่ใช้กับโรงไฟฟ้า A

จากการประมาณช่วงการทำนายค่าพยากรณ์ (PI) ด้วยสามวิธี อันได้แก่ kernel distribution, fitted distribution และ bootstrap ซึ่งใช้สัญลักษณ์ `kernel, pdf, bootstrap` ตามลำดับ รูป 5.16b แสดงช่วงการทำนายที่คำนวณจากความคลาดเคลื่อนของการพยากรณ์ 1 step ล่วงหน้า ช่วงกล่าวแสดงใน shade สีเหลือง ของ 10 วันที่เลือกมาแบบสุ่ม ช่วงการทำนายนั้นตามหลักการแล้ว จะมีความน่าจะเป็น 0.9 (เป็นพารามิเตอร์ที่ตั้งไว้) ที่จะครอบคลุมค่ากำลังไฟฟ้าผลิตจริง (นั่นคือ ช่วงการทำนายก็เป็นค่าสุ่มค่าหนึ่ง เมื่อเรามีช่วงการทำนายหลายๆ samples นั้น จะมีสัดส่วน 0.9 ที่ช่วงการทำนายนั้นจะครอบคลุมค่ากำลังไฟฟ้าผลิตจริง) จะพบว่าค่ากำลังไฟฟ้าผลิตได้จริง บางครั้งก็อยู่ในช่วงการทำนาย บางครั้งก็เกือบจะออกนอกขอบบนของช่วงการทำนาย ส่วนกราฟรูป 5.16a แสดงความกว้างของช่วง PI ที่ normalized ให้ไม่เกิน 1 เราพบว่า PI ที่คำนวณจาก kernel และ fitted distribution นั้นจะมีช่วงกว้างกว่า PI ที่คำนวณจาก bootstrap เล็กน้อย ในแค่บางช่วงเวลา และพบว่าช่วง PI มี

ความกว้างสูง ที่เวลา 9:00-10:00 และ 15:00-16:30 น. ความกว้างของช่วง PI ที่สูงนั้น บ่งชี้ว่า ค่า point forecast ที่คำนวณได้ (กราฟเส้นสีส้ม) มีช่วงความเชื่อมั่นที่กว้างเกินไป แบบจำลองย่อยที่พยากรณ์ ณ เวลาดังกล่าว จึงควรมีการพิจารณาถึงความเป็นไปได้ในการปรับปรุงโครงสร้างแบบจำลอง หรือ GMC inputs ที่ใช้



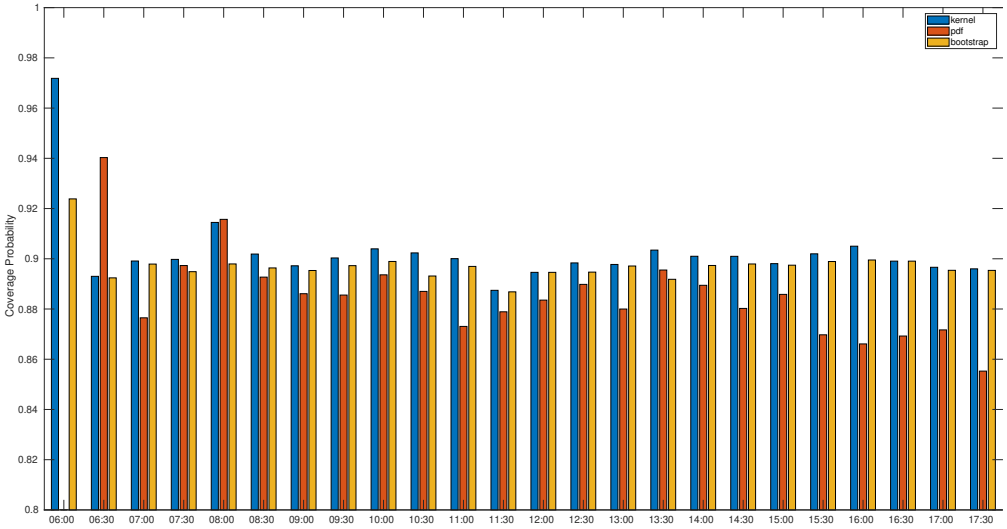
(a) Prediction interval normalized average width



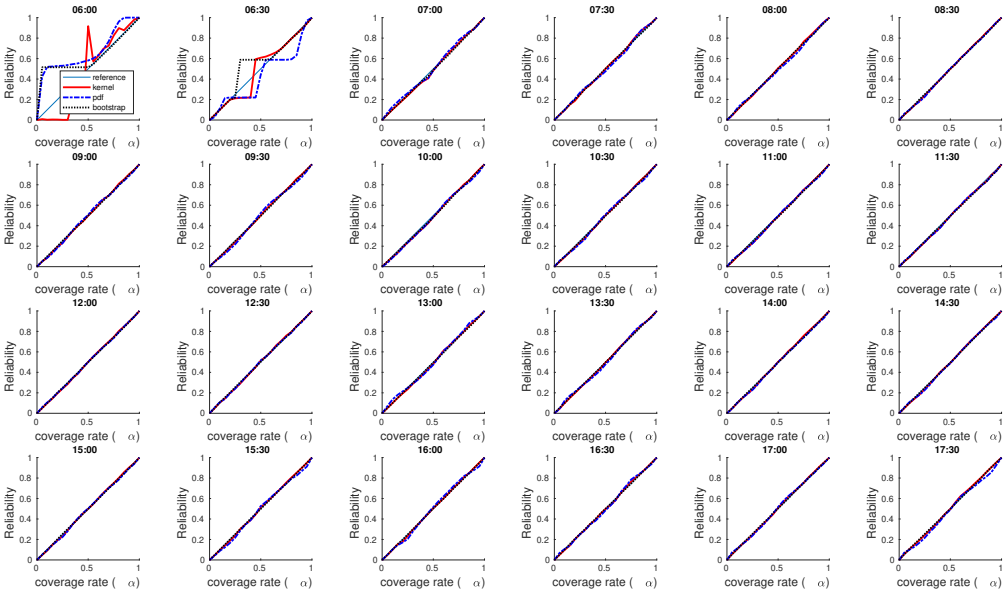
(b) ตัวอย่างค่าพยากรณ์ ค่าวัด และช่วงการทำนายจากวิธี bootstrap

รูป 5.16: สมบัติช่วงการทำนายที่ประมาณได้ของแบบจำลอง hour-ahead ของโรงไฟฟ้า A

การประมาณช่วง PI ในรายงานนี้ ได้ตั้งค่าพารามิเตอร์ probability of coverage เป็น 0.9 (ซึ่งปรับได้) ในรูป 5.17a แสดงให้เห็นว่าเมื่อนำช่วง PI ดังกล่าวไปทดสอบกับข้อมูลใน test data set นั้น ค่า coverage probability จากวิธี kernel และ bootstrap ก็ใกล้เคียงกับ 0.9 ในเกือบทุกเวลาของค่าพยากรณ์ ยกเว้นวิธี fitted distribution ที่มีค่า coverage probability ต่ำกว่า 0.9 เล็กน้อย จากรูป 5.17b พบว่า ในแต่ละเวลาพยากรณ์นั้น ช่วง PI ที่คำนวณมาจากสามวิธีนั้น มีค่า reliability ที่สอดคล้องกับค่า coverage rate (กราฟ reliability diagram ค่อนข้างเป็นเส้นตรงความชัน $45°$) ยกเว้นที่ข้อมูล ณ เวลา 6:00-6:30 น. พบว่า การคำนวณ coverage probability ของข้อมูลช่วงนั้น เกิดปัญหาเชิงเลข (numerical problem) จากวิธี kernel estimation และ fitted distribution และอีกทั้งเนื่องจากการกระจายตัวของข้อมูลเวลาดังกล่าวต่ำมาก จนทำให้ PI ที่ประมาณได้จากข้อมูลฝึกสอน ไม่สามารถครอบคลุมข้อมูลจากชุด test ได้



(a) Prediction interval coverage probability

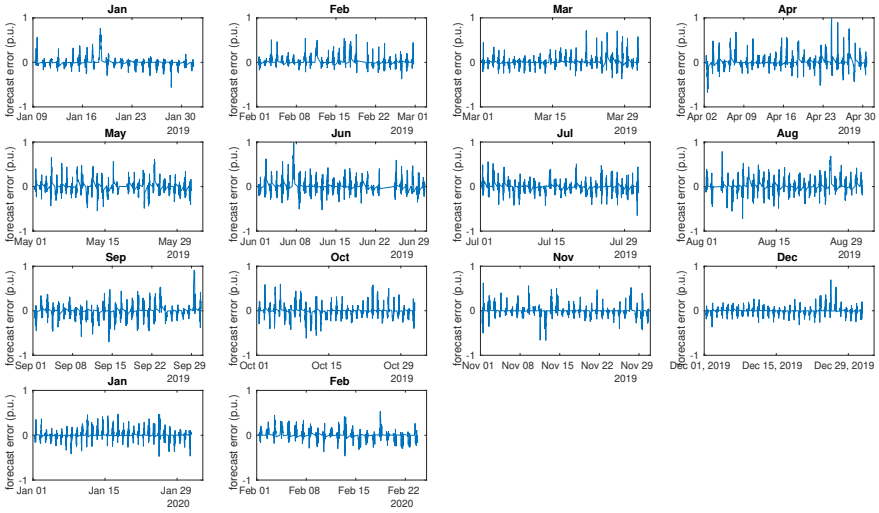


(b) Reliability diagram (กราฟแต่ละเส้นคือผลที่คำนวณจากแต่ละ fold ใน cross validation)

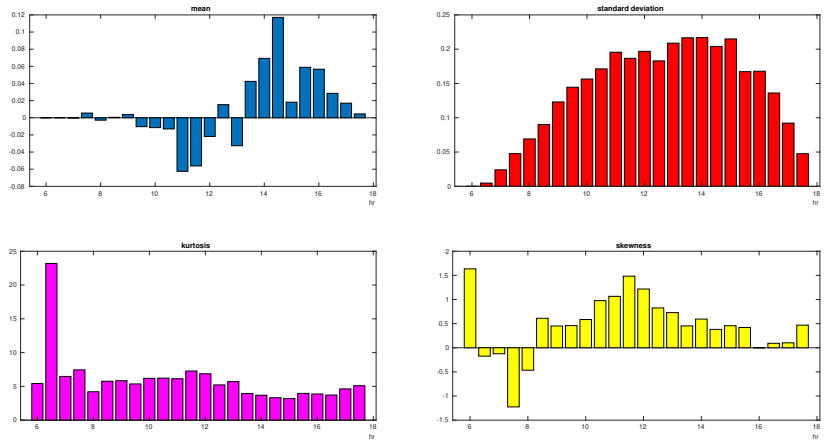
รูป 5.17: ผลการตรวจสอบสมรรถนะของช่วงการทำนายที่ประมาณได้ จากแบบจำลอง hour-ahead ของโรงไฟฟ้า A

5.4 แบบจำลอง day-ahead ที่ใช้กับโรงไฟฟ้า B

เมื่อนำค่าความคลาดเคลื่อนของการพยากรณ์มาแสดงกราฟในแต่ละเดือนดังรูป 5.18 พบว่าในช่วงเดือนมกราคม-กุมภาพันธ์นั้น ค่าความคลาดเคลื่อนที่มีค่าเกิน ± 0.5 p.u. อยู่ในสัดส่วนที่น้อย ค่าความคลาดเคลื่อนที่สูงเกิน 0.5 p.u. มีสัดส่วนมากขึ้นในเดือนเมษายน-มิถุนายน กราฟในรูป 5.19 แสดงให้เห็นว่า สำหรับค่า mean ซึ่งบ่งชี้ว่าแบบจำลองย่อยของแต่ละเวลามี bias ต่างกันในแต่ละเวลา โดยช่วงเช้าจะมีค่า bias ที่ต่ำไปจนถึง 14:00 ซึ่งแบบจำลองเริ่ม overestimate หลังจากตอนบ่ายเป็นต้นไป ค่าความแปรปรวนนั้น จะมีค่าสูงในช่วง 14:00-16:00 ค่า kurtosis มีค่าสูงมาก ที่ 6:00 และโดยรวม ค่า kurtosis ที่สูงกว่า 3 (ซึ่งเป็นค่า kurtosis ของ Gaussian distribution) จะเป็นของแบบจำลองช่วงเวลา 9:30-13:30 ค่า skewness ส่วนใหญ่จะเป็นบวก บ่งชี้ว่า การกระจายตัวนั้นเป็นแบบ right-tailed (histogram เบ้ซ้าย)

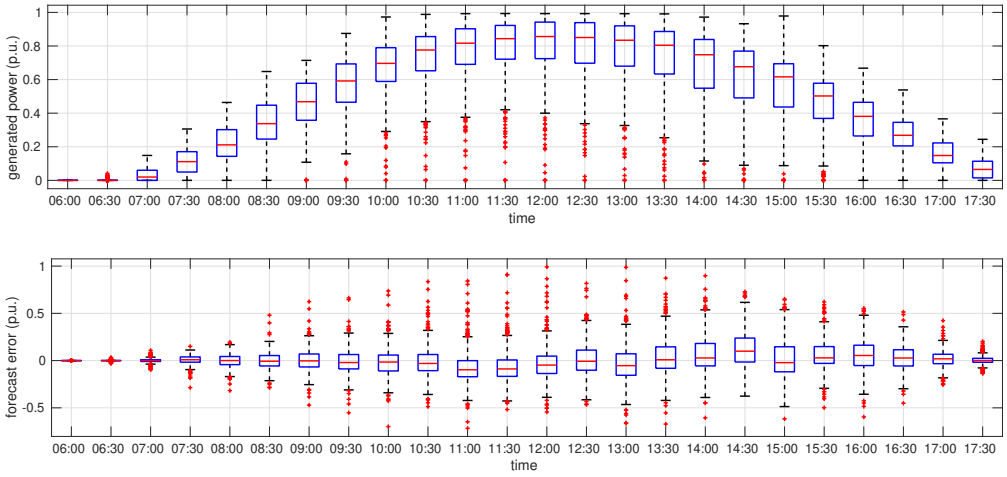


รูป 5.18: ความคลาดเคลื่อนในแต่ละเดือนของการพยากรณ์จากแบบจำลอง day-ahead ที่ใช้กับโรงไฟฟ้า B

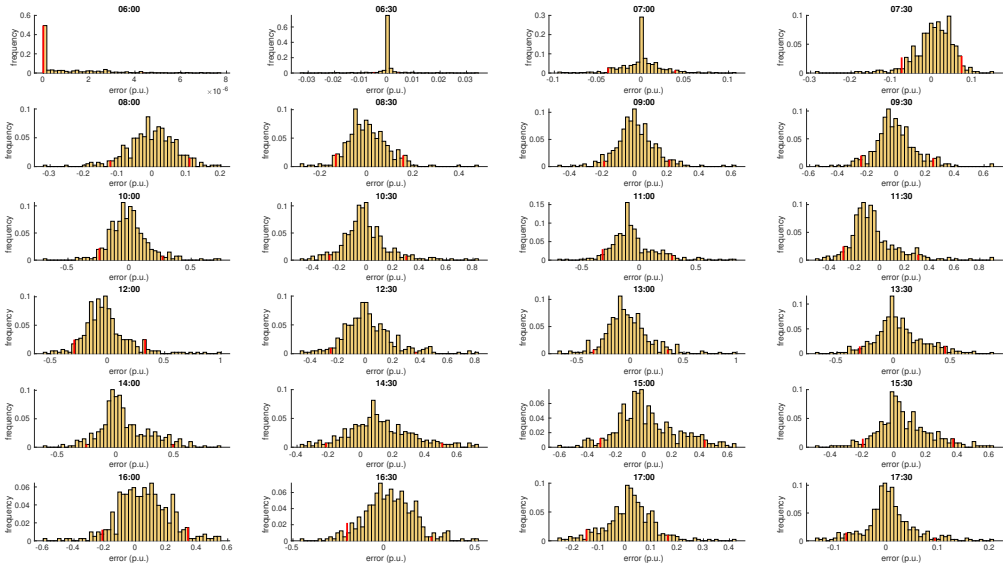


รูป 5.19: ค่าทางสถิติของความคลาดเคลื่อนของการพยากรณ์จากแบบจำลอง day-ahead ที่ใช้กับโรงไฟฟ้า B

เมื่อนำความคลาดเคลื่อนพยากรณ์มาวิเคราะห์การกระจายตัวนั้น ในรูป 5.20 พบว่า กล่องสีน้ำเงินใน box plot ซึ่งแสดงถึง interquartile range (Q1-Q3) จะเป็นช่วงที่กว้างมากในเวลา 14:00-15:00 (ช่วงนี้ครอบคลุมความน่าจะเป็น 0.5) และสอดคล้องกับการแสดงผลด้วย histogram จุดสีดำใน box plot แสดงค่า extreme ที่คำนวณจาก median (ขีดสีแดง) นำมาบวกลบกับสัดส่วนของความกว้าง interquartile range โดยจะเห็นว่าค่า extreme ดังกล่าวจะกระโดดไปสูงในเวลา 14:00-15:00 ส่วนจุดสีแดงใน box plot แสดงถึง outliers ที่พบได้มากในช่วงเวลา 11:00-12:00 เมื่อพิจารณา histogram เราพบว่า เวลา 6:00-7:00 ความคลาดเคลื่อนจะมีการกระจุกตัวมากที่ค่าใกล้ๆ ศูนย์ ลักษณะ shape ของ histogram นี้จะมีพารามิเตอร์ทางสถิติที่ต่างกันไปค่อนข้างเล็กน้อยในแต่ละเวลา (เช่น ค่าเฉลี่ย ค่า median ความหนาของ tail ความสมมาตร เป็นต้น) ซึ่งสอดคล้องกับผลในรูป 5.19



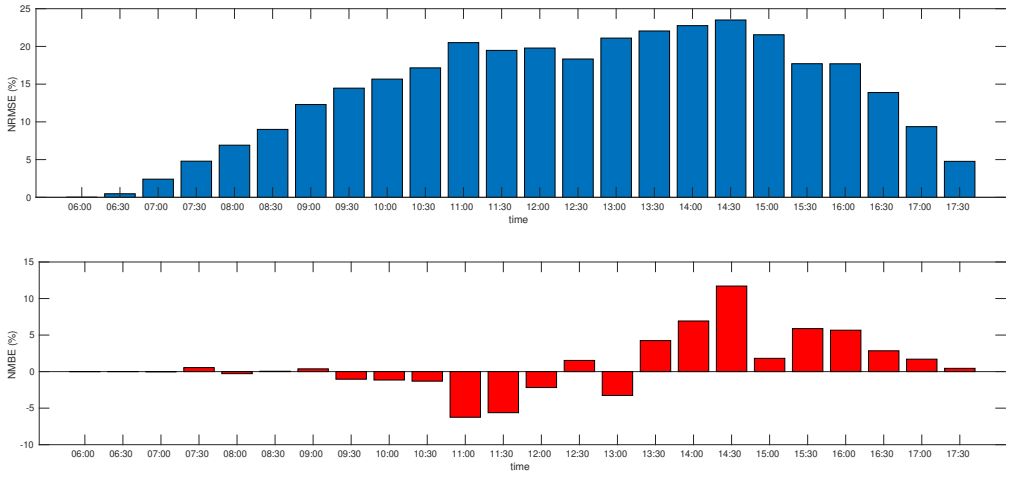
(a) box plot ของความคลาดเคลื่อน



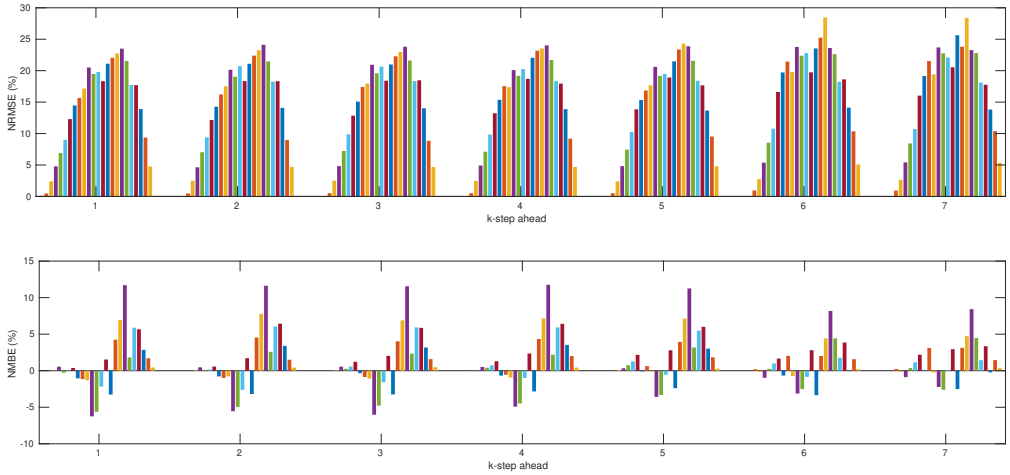
(b) histogram ของความคลาดเคลื่อน ช่วงสีแดงแสดงถึงช่วงความเชื่อมั่นด้วยความน่าจะเป็น 0.9

รูป 5.20: สมบัติการกระจายตัวของความคลาดเคลื่อนการพยากรณ์ด้วยแบบจำลอง day-ahead ของโรงไฟฟ้า B

สมรรถนะการพยากรณ์แสดงในรูป 5.21a เป็นการคำนวณจากพยากรณ์แบบ 1 วันล่วงหน้า พบว่าค่า NRMSE สูงประมาณ 22-24 % ณ เวลา 13:30-15:00 อันเป็นผลที่สอดคล้องกับความแปรปรวนของความคลาดเคลื่อนสูงในช่วงเวลานั้น แบบจำลองย่อยของเวลาเช้าพบว่า ส่วนใหญ่จะให้ bias ที่ต่ำมาก มีช่วงเวลา 11:00-11:30 ที่แบบจำลอง underestimate ไปเล็กน้อย แต่ตอนช่วงบ่าย แบบจำลองค่อนข้างจะ overestimate เมื่อพิจารณาสมรรถนะการพยากรณ์ของ 7 วันล่วงหน้าในรูป 5.21b พบว่าแนวโน้มของค่าสมรรถนะในแต่ละเวลาพยากรณ์นั้นใกล้เคียงกัน กล่าวคือ เวลาที่ค่า NRMSE สูงจะเกิดขึ้นที่ช่วงเวลาเดียวกัน นอกจากนี้ สมรรถนะโดยรวมที่วัดจากค่า NRMSE พบว่าไม่ได้มีต่างกันมากนัก เมื่อเป็นการพยากรณ์หลายวันล่วงหน้ามากขึ้น แต่ค่า NRMSE สูงสุดที่ 27-28% จะพบที่การพยากรณ์ 6-7 วันล่วงหน้าที่เวลา 13:30



(a) สมรรถนะการพยากรณ์ 1 วันล่วงหน้า

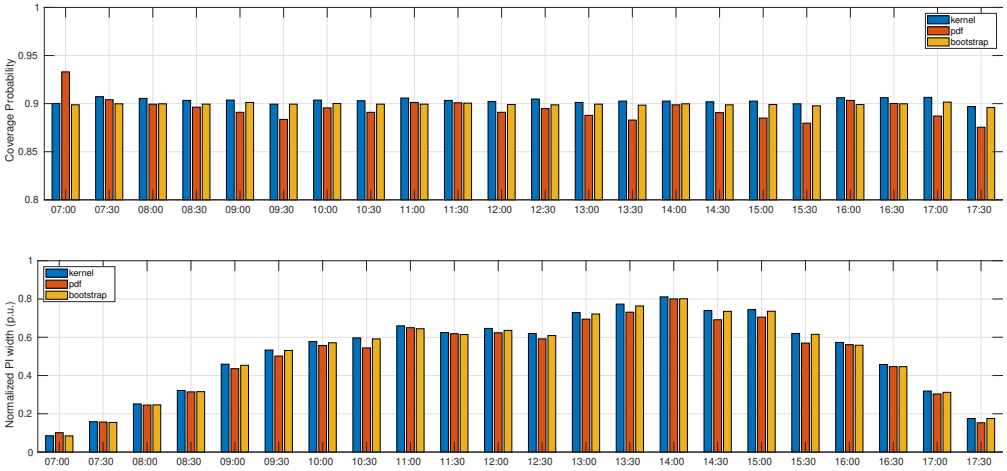


(b) สมรรถนะการพยากรณ์ 7 วันล่วงหน้า

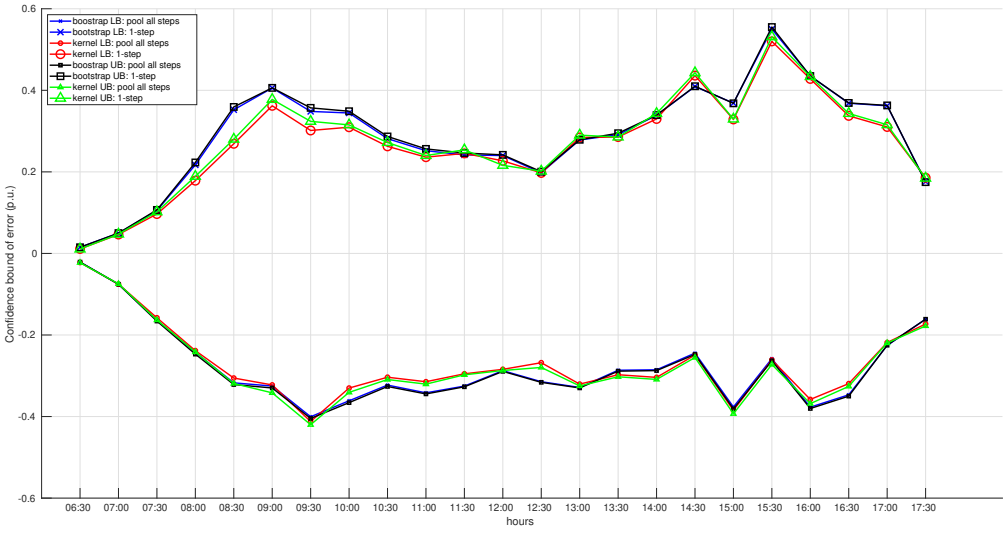
รูป 5.21: สมรรถนะการพยากรณ์ด้วยแบบจำลอง day-ahead ที่ใช้กับโรงไฟฟ้า B

จากการประมาณช่วงการทำนายค่าพยากรณ์ (PI) ด้วยสามวิธี อันได้แก่ kernel distribution, fitted distribution และ bootstrap ซึ่งใช้สัญลักษณ์ `kernel, pdf, bootstrap` ตามลำดับ ในรูป 5.22a แสดงความกว้างของช่วง PI ที่ normalized ให้ไม่เกิน 1 เราพบว่า PI ที่คำนวณจาก kernel density นั้นจะมีช่วงกว้างกว่าวิธีอื่นที่เวลา 6:30-7:00 PI ที่คำนวณจาก bootstrap เล็กน้อย และพบว่าช่วง PI มีความกว้างสูง ที่เวลา 13:00-15:00 การประมาณช่วง PI ในรายงานนี้ ได้ตั้งค่าพารามิเตอร์ probability of coverage เป็น 0.9 (ซึ่งปรับได้) ในรูป 5.22a แสดงให้เห็นว่าเมื่อนำช่วง PI ดังกล่าวไปทดสอบกับข้อมูลใน test data set นั้น ค่า coverage probability ก็ใกล้เคียงกับ 0.9 ในเกือบทุกเวลาของค่าพยากรณ์ ยกเว้นวิธี fitted distribution ที่มี coverage probability ที่ต่ำกว่า 0.9 ในหลายช่วงเวลา

รูป 5.22b เปรียบเทียบช่วงการทำนายที่คำนวณมาจากสองวิธี สำหรับเวลาหนึ่งๆ เช่น ณ เวลา 10:00 น. ค่าความคลาดเคลื่อนจะมีมาจากการพยากรณ์ 1-step,2-step,...,7-step ล่วงหน้า การคำนวณสองวิธีดังกล่าวคือ i) การรวมความคลาดเคลื่อนจากทุก step มาด้วยกันแล้วคำนวณ PI และ ii) การคำนวณ PI ของความคลาดเคลื่อนที่แยกตาม k-step เมื่อพิจารณาสองวิธีโดยที่วิธีหลังนั้นพิจารณา ที่ 1-step เราจะพบว่า PI ที่คำนวณจากวิธีแรกจะมีช่วงกว้างกว่าเล็กน้อย เนื่องจากความคลาดเคลื่อนย่อมมีการกระจายตัวที่สูงกว่า



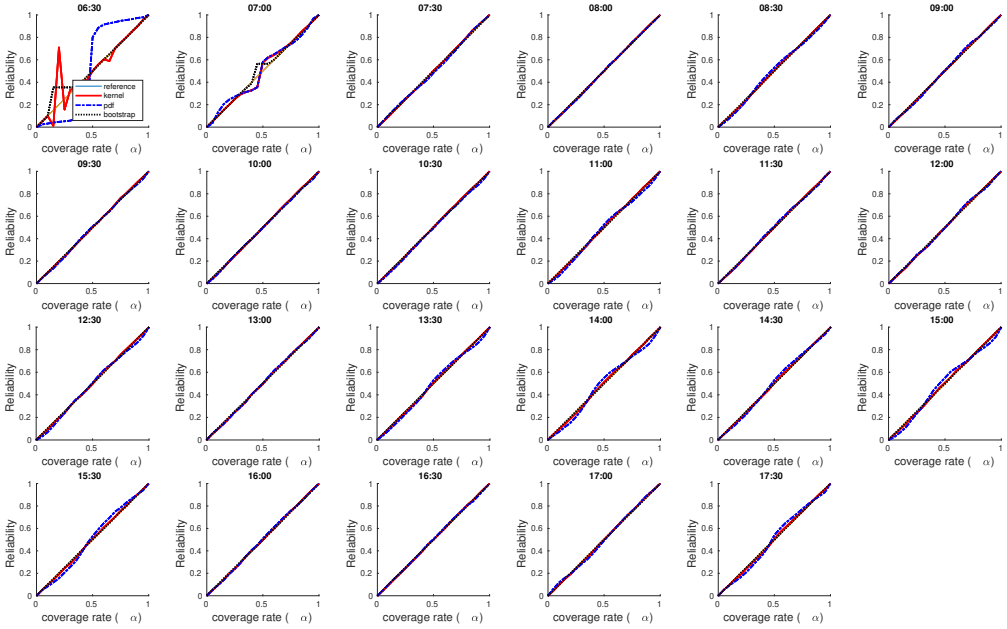
(a) Prediction interval normalized average width



(b) ช่วง PI จากการรวมความคลาดเคลื่อนทุก k-step และการวิเคราะห์จาก 1-step

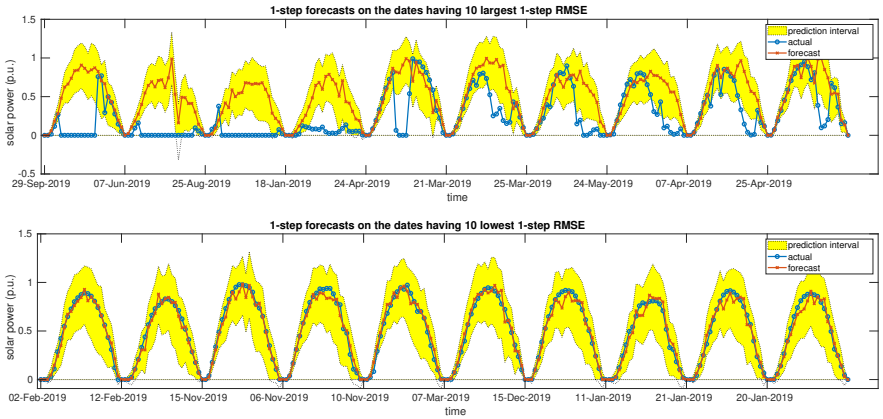
รูป 5.22: สมบัติช่วงการทำนายที่ประมาณได้ของแบบจำลอง day-ahead ของโรงไฟฟ้า B

จากรูป 5.23 พบว่า ในแต่ละเวลาพยากรณ์นั้น ช่วง PI ที่คำนวณมาจากสามวิธีนั้น มีค่า reliability ที่สอดคล้องกับค่า coverage rate (กราฟ reliability diagram ค่อนข้างเป็นเส้นตรงความชัน $45°$) ยกเว้นที่ข้อมูล ณ เวลา 6:00-6:30 น. พบว่า การคำนวณ coverage probability ของข้อมูลช่วงนั้น เกิดปัญหาเชิงเลข (numerical problem) และอีกทั้งเนื่องจากการกระจายตัวของข้อมูลเวลาดังกล่าวต่ำมาก จนทำให้ PI ที่ประมาณได้จากข้อมูลฝึกสอน ไม่สามารถครอบคลุมข้อมูลจากชุด test ได้



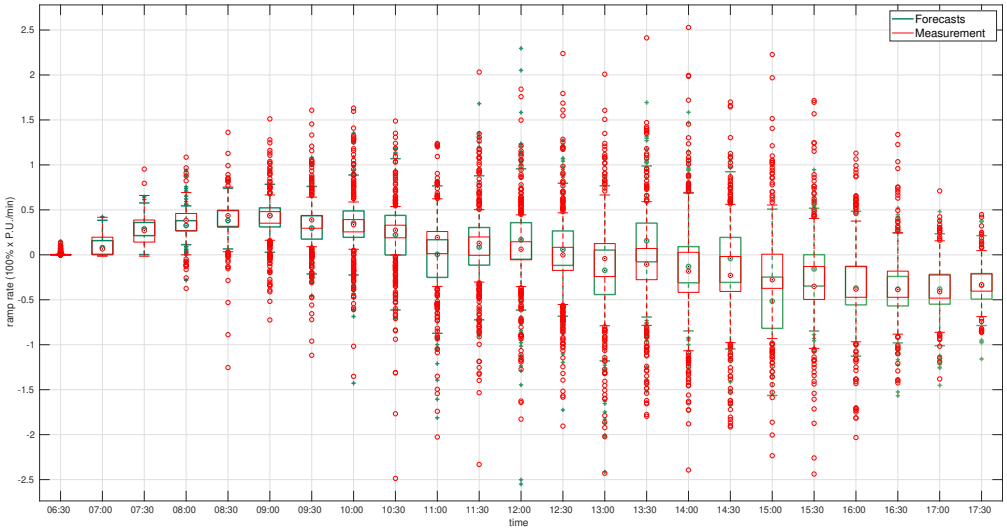
รูป 5.23: Reliability diagram (กราฟแต่ละเส้นคือผลที่คำนวณจากแต่ละ fold ใน cross validation) ที่ได้จากการตรวจสอบสมรรถนะของช่วง PI จากแบบจำลอง day-ahead ของโรงไฟฟ้า B

ในรูป 5.24 แสดงช่วงการทำนายที่คำนวณจากความคลาดเคลื่อนของการพยากรณ์ 1 step ล่วงหน้า ช่วงกล่าวแสดงใน shade สีเหลือง ของ 10 วันที่มีค่า NRMSE สูงสุดและต่ำสุด 10 ค่า และวันที่ที่แสดงของวันทีสมรรถนะแย่ก็เรียงตามค่า NRMSE จากสูงสุดไปยังค่าสูงสุดที่ 10 ส่วนวันที่ที่มีสมรรถนะดี ก็เรียงวันที่ตามค่า NRMSE ที่ต่ำสุดไปยังค่าที่ต่ำสุดที่ 10 เราสังเกตว่าวันที่สมรรถนะดีจะอยู่ในช่วง 2 เดือนแรกของต้นปีและปลายปี ส่วนวันที่สมรรถนะแย่ดูเสมือนจะมีความผิดพลาดของค่าวัด ช่วงการทำนายนั้นตามหลักการแล้ว จะมีความน่าจะเป็น 0.9 (เป็นพารามิเตอร์ที่ตั้งไว้) ที่จะครอบคลุมค่ากำลังไฟฟ้าผลิตจริง (นั่นคือ ช่วงการทำนายก็เป็นค่าสุ่มค่าหนึ่ง เมื่อเรามีช่วงการทำนายหลายๆ samples นั้น จะมีสัดส่วน 0.9 ที่ช่วงการทำนายนั้นจะครอบคลุมค่ากำลังไฟฟ้าผลิตจริง) ในการประยุกต์ใช้จริง เราจะเห็นจากตัวอย่างกราฟว่า ในบางวันค่ากำลังผลิตจริง ก็อยู่นอกช่วงการทำนาย

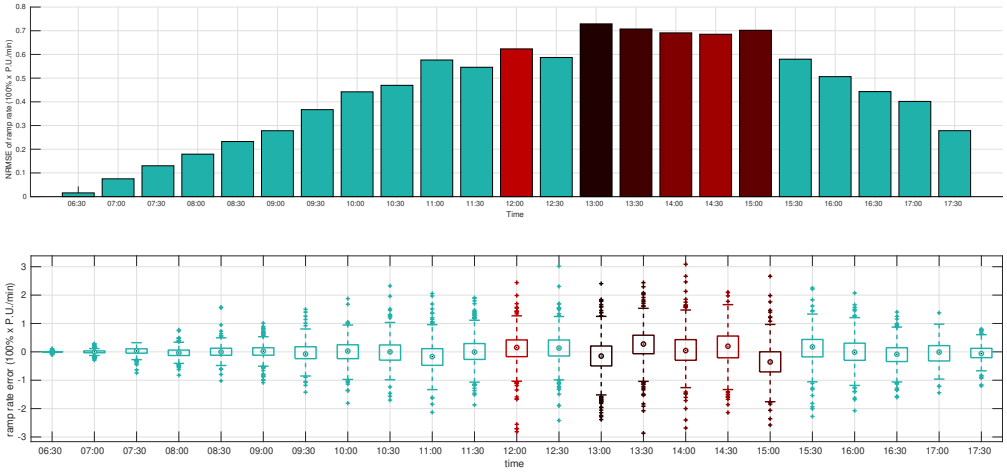


รูป 5.24: ตัวอย่างค่าพยากรณ์ และช่วงการทำนายจากวิธี bootstrap ของแบบจำลอง day-ahead ของโรงไฟฟ้า B

รูป 5.25a แสดง box plot ของ ramp rate ของค่ากำลังไฟฟ้าวัดจริงและค่าพยากรณ์ ที่เห็นได้ว่าค่า median ของ ramp rate ในช่วงเช้าถึง 10:00 นั้น มีค่าเป็นบวก และจะแกว่งไปมาค่าในช่วง 10:00-14:00 น. ในช่วงกลางวันดังกล่าวจะพบว่ามี outliers ของ ramp rate เป็นปริมาณที่สูงมาก ramp rate จะมีค่าเป็นลบอีกครั้งในช่วงหลัง 15:30 น. ค่า median ของ ramp rate ของค่าพยากรณ์ในช่วงบ่ายนั้นมักจะไม่ค่อยมีค่าใกล้เคียงกับ ramp rate ของค่ากำลังไฟฟ้าวัดจริง รูป 5.25b แสดงให้เห็นค่า NRMSE ของ ramp rate และเน้นจุดเวลาที่มีค่า NRMSE สูงสุด 6 ค่า ที่พบว่าในช่วงเวลา 13:00-15:00 น. นั้นมีความคลาดเคลื่อนของ ramp rate สูงมาก ซึ่งเมื่อดู box plot ประกอบก็จะเห็นว่าช่วง inter-quartile ณ เวลาดังกล่าวมีช่วงที่กว้างนั่นเอง



(a) Ramp rate of measured and forecasted solar power.

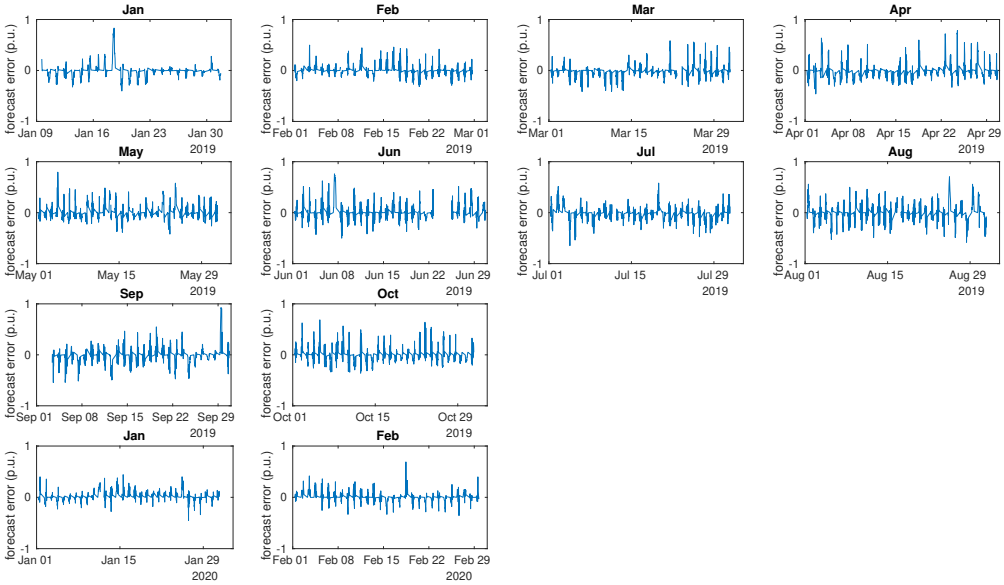


(b) Ramp rate error (the darker tone the higher NRMSE).

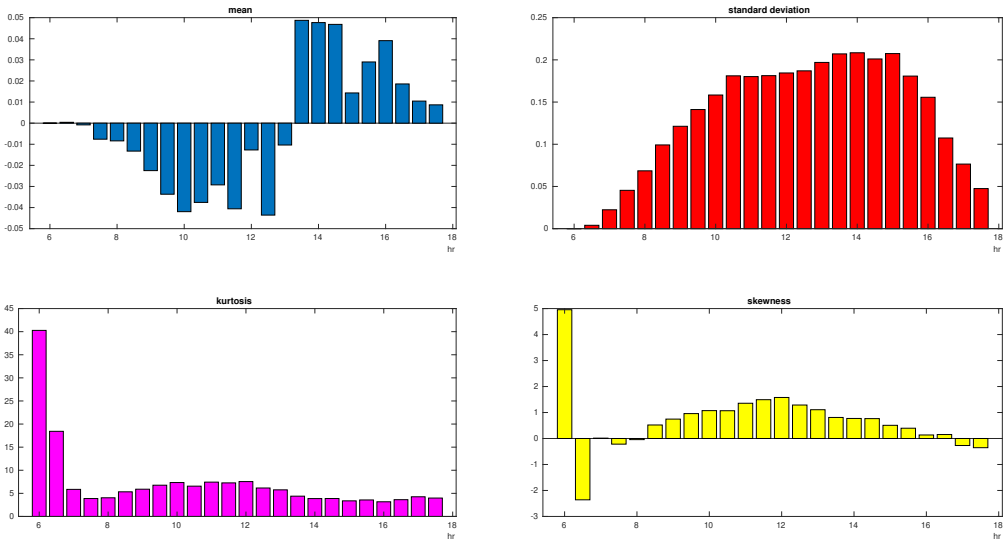
รูป 5.25: สมรรถนะของ ramp rate ที่ได้จากแบบจำลอง day-ahead ของโรงไฟฟ้า B

5.5 แบบจำลอง hour-ahead ที่ใช้กับโรงไฟฟ้า B

เมื่อนำค่าความคลาดเคลื่อนของการพยากรณ์มาแสดงกราฟในแต่ละเดือนดังรูป 5.26 พบว่าค่าความคลาดเคลื่อนที่สูงเกิน 0.4 p.u. มีสัดส่วนมากในเดือนเมษายน-มิถุนายน กราฟในรูป 5.27 แสดงให้เห็นว่า สำหรับค่า mean ซึ่งบ่งชี้ว่าแบบจำลองย่อยของแต่ละเวลามี bias ต่างกันในแต่ละเวลา โดยช่วงเช้าจะมีค่า bias ที่ติดลบจนถึง 13:00 จนมี bias เป็นค่าบวก (overestimate) ในช่วงบ่ายเป็นต้นไป ค่าความแปรปรวนนั้น จะมีค่าสูงในช่วง 14:30-15:00 ค่า kurtosis มีค่าสูงมาก ที่ 6:00 และโดยรวม ค่า kurtosis ที่สูงกว่า 3 (ซึ่งเป็นค่า kurtosis ของ Gaussian distribution) จะเป็นของแบบจำลองช่วงเวลา 9:30-13:30 ค่า skewness ส่วนใหญ่จะเป็นบวก บ่งชี้ว่า การกระจายตัวนั้นเป็นแบบ right-tailed (histogram เบ้ซ้าย)

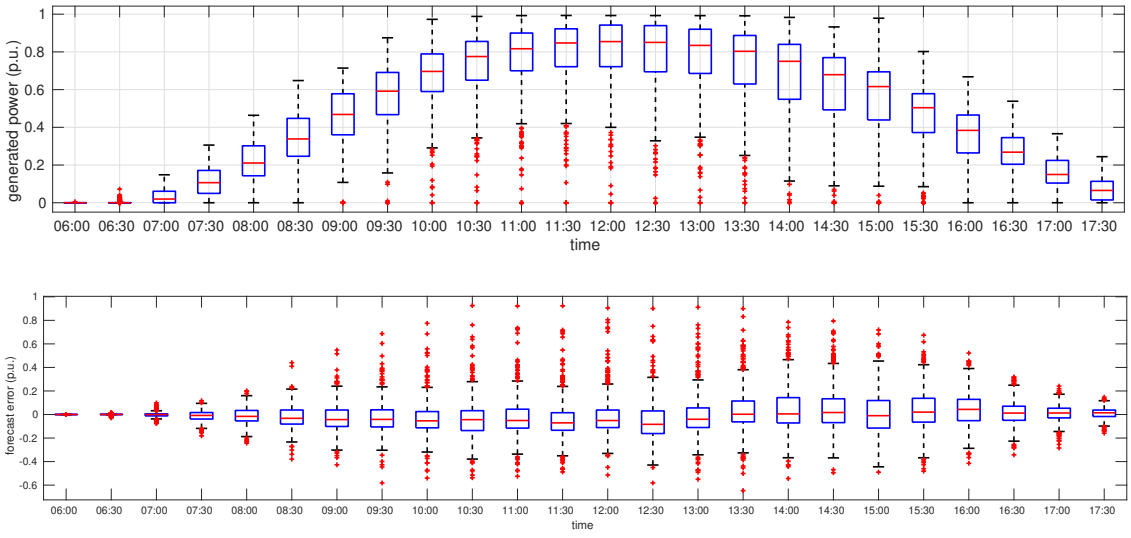


รูป 5.26: ความคลาดเคลื่อนในแต่ละเดือนของการพยากรณ์จากแบบจำลอง hour-ahead ที่ใช้กับโรงไฟฟ้า B

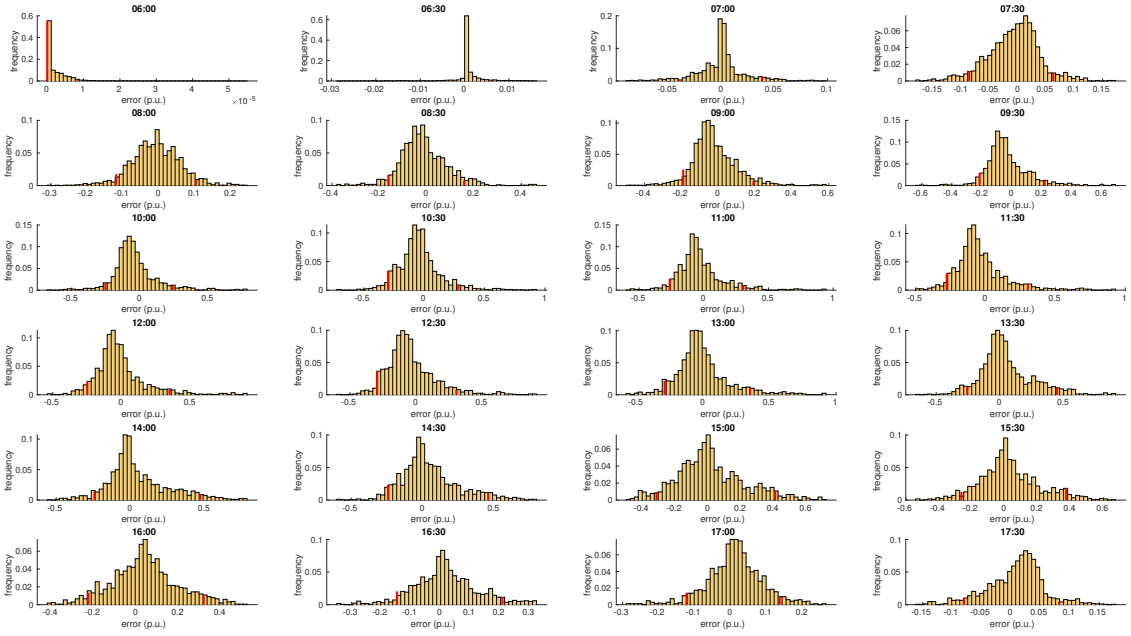


รูป 5.27: ค่าทางสถิติของความคลาดเคลื่อนของการพยากรณ์จากแบบจำลอง hour-ahead ที่ใช้กับโรงไฟฟ้า B

เมื่อนำความคลาดเคลื่อนการพยากรณ์มาวิเคราะห์การกระจายตัวนั้น ในรูป 5.28 พบว่า กล่องสีน้ำเงินใน box plot ซึ่งแสดงถึง interquartile range (Q1-Q3) จะเป็นช่วงที่กว้างมากในเวลา 14:00-16:00 (ช่วงนี้ครอบคลุมความน่าจะเป็น 0.5) และสอดคล้องกับการแสดงผลด้วย histogram จุดสีดำใน box plot แสดงค่า extreme ที่คำนวณจาก median (ขีดสีแดง) นำมาบวกลบกับสัดส่วนของความกว้าง interquartile range โดยจะเห็นว่าค่า extreme ดังกล่าวจะกระโดดไปสูงในเวลา 14:00-15:00 ส่วนจุดสีแดงใน box plot แสดงถึง outliers ที่พบได้มากในช่วงเวลา 12:30-13:30 เมื่อพิจารณา histogram เราพบว่า เวลา 6:00-7:00 ความคลาดเคลื่อนจะมีการกระจุกตัวมากที่ค่าใกล้ๆ ศูนย์ ลักษณะ shape ของ histogram นี้ พบกว่ามีการกระจายตัวกว้างในช่วงเวลา 14:30-17:30 และจะมีพารามิเตอร์ทางสถิติที่ต่างกันไปค่อนข้างเล็กน้อยในแต่ละเวลา (เช่น ค่าเฉลี่ย ค่า median ความหนาของ tail ความสมมาตร เป็นต้น) ซึ่งสอดคล้องกับผลในรูป 5.27



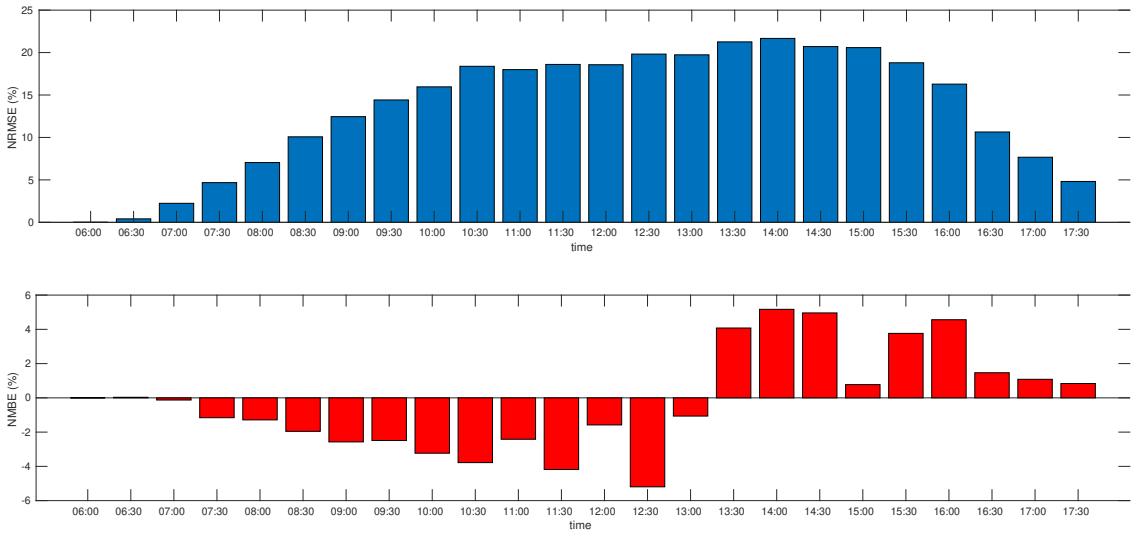
(a) box plot ของความคลาดเคลื่อน



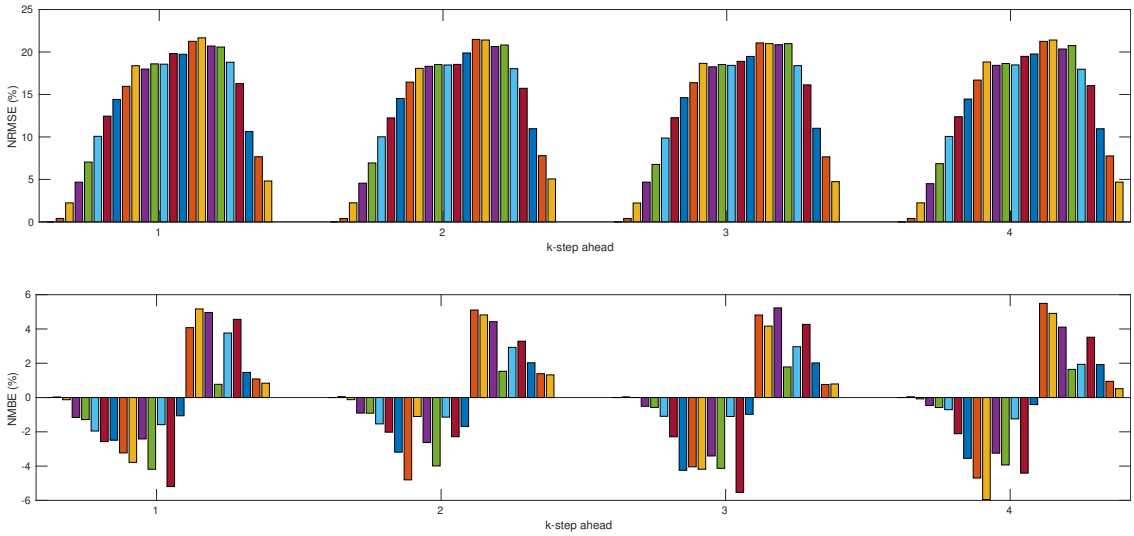
(b) histogram ของความคลาดเคลื่อน ช่วงสีแดงแสดงถึงช่วงความเชื่อมั่นด้วยความน่าจะเป็น 0.9

รูป 5.28: สมบัติการกระจายตัวของความคลาดเคลื่อนการพยากรณ์ด้วยแบบจำลอง hour-ahead ของโรงไฟฟ้า B

สมรรถนะการพยากรณ์แสดงในรูป 5.29a เป็นการคำนวณจากพยากรณ์แบบ 1-step ล่วงหน้า พบว่าค่า NRMSE สูงประมาณ 20-21% ณ เวลา 13:30-15:00 อันเป็นผลที่สอดคล้องกับความแปรปรวนของความคลาดเคลื่อนสูงในช่วงเวลานั้น แบบจำลองย่อยของเวลาเช้าพบว่า underestimate (จากค่า MBE ที่ติดลบ) แต่ตอนช่วงบ่าย แบบจำลองค่อนข้างจะ overestimate เมื่อพิจารณาสมรรถนะการพยากรณ์ของ 4-step ล่วงหน้าในรูป 5.29b พบว่าแนวโน้มของค่าสมรรถนะในแต่ละเวลาพยากรณ์นั้นใกล้เคียงกัน กล่าวคือ เวลาที่ค่า NRMSE สูงจะเกิดขึ้นที่ช่วงเวลาเดียวกัน นอกจากนี้ สมรรถนะโดยรวมที่วัดจากค่า NRMSE พบว่าไม่ได้มีต่างกันมากนัก เมื่อเป็นการพยากรณ์หลาย step ล่วงหน้ามากขึ้น



(a) สมรรถนะการพยากรณ์ 1-step ล่วงหน้า

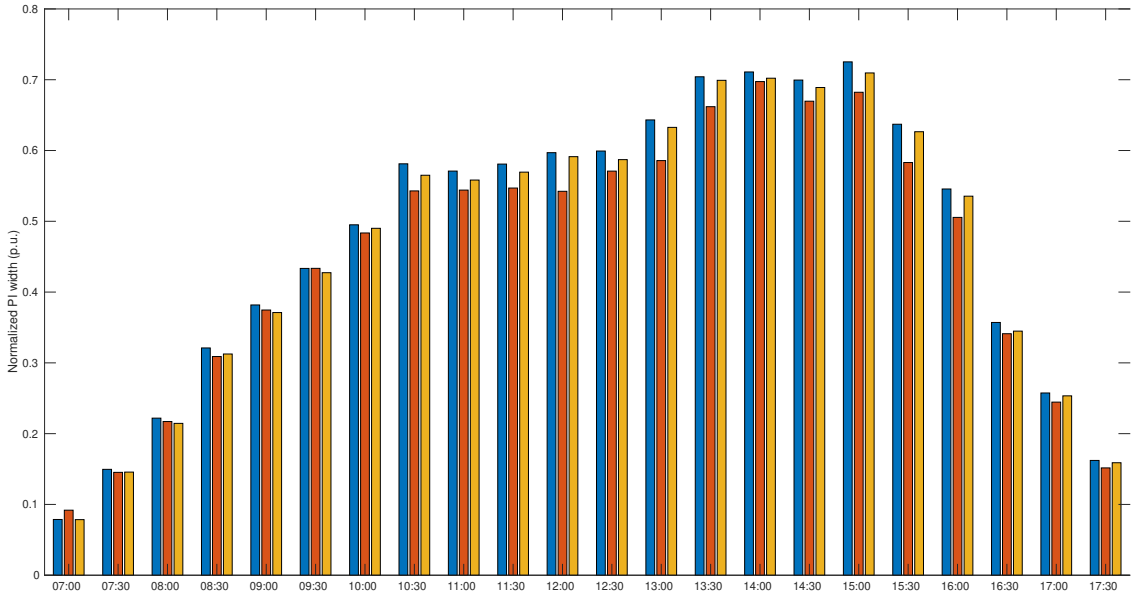


(b) สมรรถนะการพยากรณ์ 4-step ล่วงหน้า

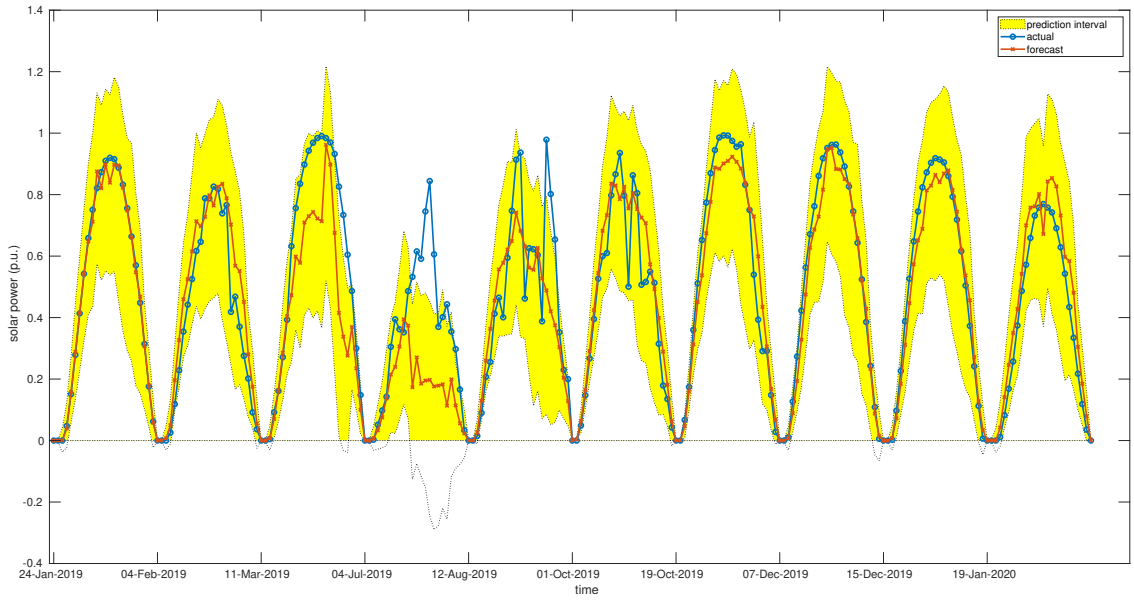
รูป 5.29: สมรรถนะการพยากรณ์ด้วยแบบจำลอง hour-ahead ที่ใช้กับโรงไฟฟ้า B

จากการประมาณช่วงการทำนายค่าพยากรณ์ (PI) ด้วยสามวิธี อันได้แก่ kernel distribution, fitted distribution และ bootstrap ซึ่งใช้สัญลักษณ์ `kernel, pdf, bootstrap` ตามลำดับ รูป 5.30b แสดงช่วงการทำนายที่คำนวณจากความคลาดเคลื่อนของการพยากรณ์ 1 step ล่วงหน้า ช่วงกล่าวแสดงใน shade สีเหลือง ของ 10 วันที่เลือกมาแบบสุ่ม ช่วงการทำนายนั้นตามหลักการแล้ว จะมีความน่าจะเป็น 0.9 (เป็นพารามิเตอร์ที่ตั้งไว้) ที่จะครอบคลุมค่ากำลังไฟฟ้าผลิตจริง (นั่นคือ ช่วงการทำนายก็เป็นค่าสุ่มค่าหนึ่ง เมื่อเรามีช่วงการทำนายหลายๆ samples นั้น จะมีสัดส่วน 0.9 ที่ช่วงการทำนายนั้นจะครอบคลุมค่ากำลังไฟฟ้าผลิตจริง) ในการประยุกต์ใช้จริง เราจะเห็นจากตัวอย่างกราฟว่า ในบางวันค่ากำลังผลิตจริง ก็อยู่นอกช่วงการทำนาย ส่วนกราฟรูป 5.30a แสดงความกว้างของช่วง PI ที่ normalized ให้ไม่เกิน 1 เราพบว่า PI ที่คำนวณจาก kernel และ fit-

ted distribution นั้นจะมีช่วงกว้างกว่า PI ที่คำนวณจาก bootstrap เล็กน้อย ในแค่บางช่วงเวลา และพบว่าช่วง PI มีความกว้างสูง ที่เวลา 14:00-15:00 ความกว้างของช่วง PI ที่สูงถึง 0.7 p.u. บ่งชี้ว่า ค่า point forecast ที่คำนวณได้ (กราฟเส้นสีส้ม) มีช่วงความเชื่อมั่นที่กว้างเกินไป แบบจำลองย่อยที่พยากรณ์ ณ เวลาดังกล่าว จึงควรมีการพิจารณาถึงความเป็นไปได้ในการปรับปรุงโครงสร้างแบบจำลอง หรือ GMC inputs ที่ใช้



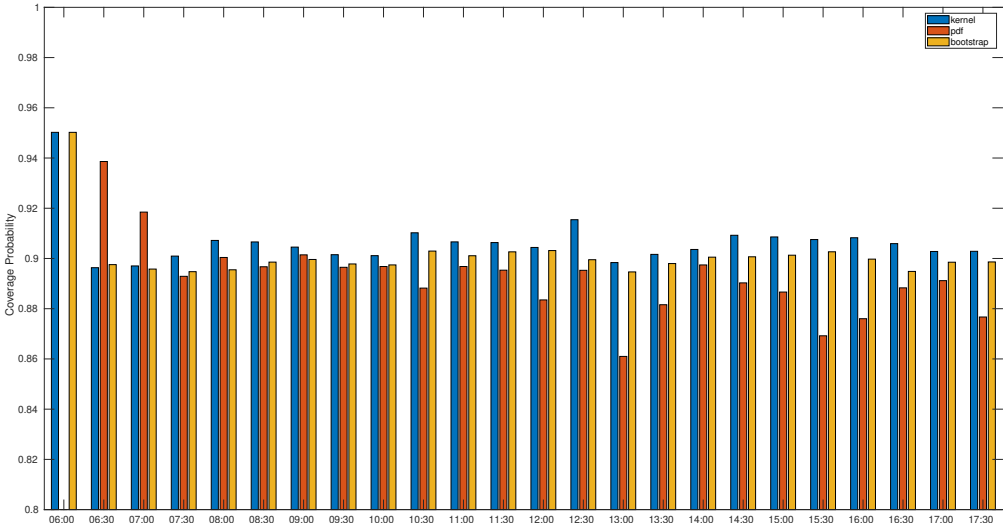
(a) Prediction interval normalized average width



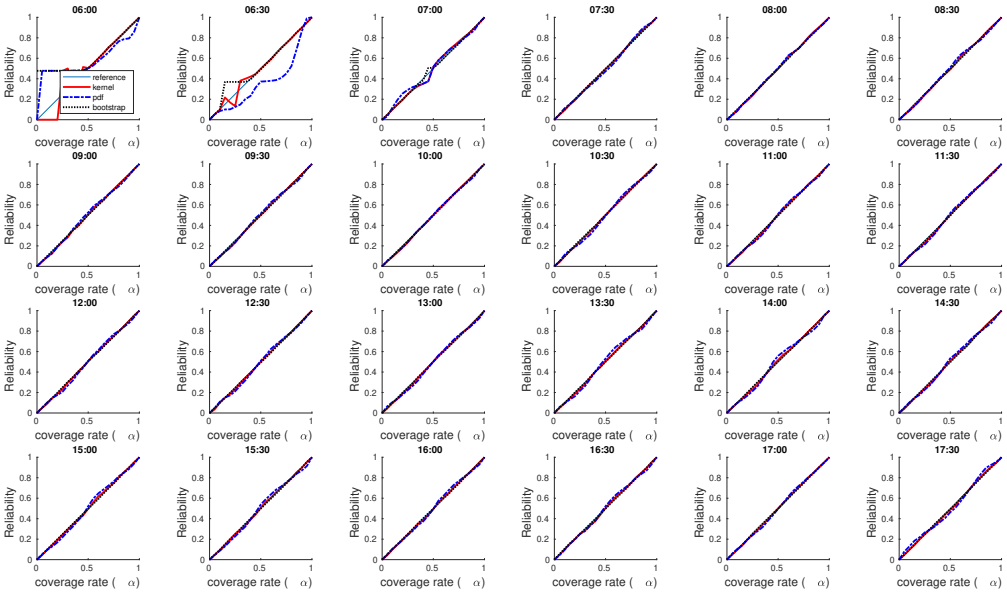
(b) ตัวอย่างค่าพยากรณ์ ค่าวัด และช่วงการทำนาย

รูป 5.30: สมบัติช่วงการทำนายที่ประมาณได้จากวิธี bootstrap ของแบบจำลอง hour-ahead ของโรงไฟฟ้า B

การประมาณช่วง PI ในรายงานนี้ ได้ตั้งค่าพารามิเตอร์ probability of coverage เป็น 0.9 (ซึ่งปรับได้) ในรูป 5.31a แสดงให้เห็นว่าเมื่อนำช่วง PI ดังกล่าวไปทดสอบกับข้อมูลใน test data set นั้น ค่า coverage probability ที่ได้จากวิธี kernel และ bootstrap ก็ใกล้เคียงกับ 0.9 ในเกือบทุกเวลาของค่าพยากรณ์ ยกเว้นวิธีจาก fitted distribution จากรูป 5.31b พบว่า ในแต่ละเวลาพยากรณ์นั้น ช่วง PI ที่คำนวณมาจากสามวิธีนั้น มีค่า reliability ที่สอดคล้องกับค่า coverage rate (กราฟ reliability diagram ค่อนข้างเป็นเส้นตรงความชัน $45°$) ยกเว้นที่ข้อมูล ณ เวลา 6:00-6:30 น. พบว่า การคำนวณ coverage probability ของข้อมูลช่วงนั้น เกิดปัญหาเชิงเลข (numerical problem) และอีกทั้งเนื่องจากการกระจายตัวของข้อมูลเวลาดังกล่าวต่ำมาก จนทำให้ PI ที่ประมาณได้จากข้อมูลฝึกสอน ไม่สามารถครอบคลุมข้อมูลจากชุด test ได้



(a) Prediction interval coverage probability



(b) Reliability diagram (กราฟแต่ละเส้นคือผลที่คำนวณจากแต่ละ fold ใน cross validation)

รูป 5.31: ผลการตรวจสอบสมรรถนะของช่วงการทำนายที่ประมาณได้ จากแบบจำลอง hour-ahead ของโรงไฟฟ้า B

บทที่ 6

ผลสรุปการวิเคราะห์ความคลาดเคลื่อนการพยากรณ์

เราจะแสดงผลเปรียบเทียบของสองโรงไฟฟ้าตามประเด็นดังนี้

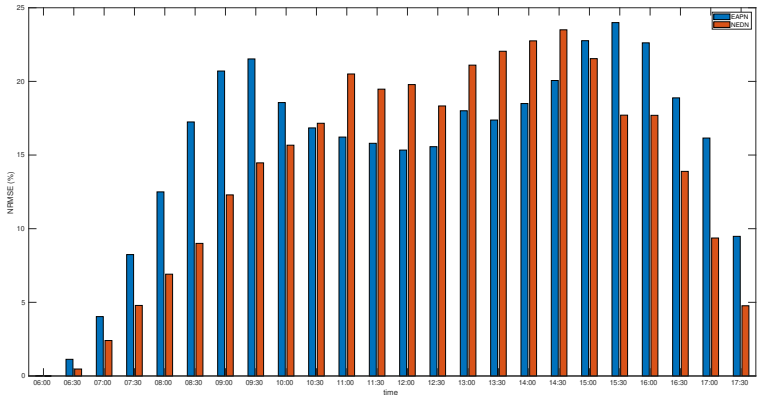
1. สมรรถนะการพยากรณ์ในแต่ละเวลาพยากรณ์ (NRMSE และ NMBE)
2. ความกว้างของช่วงการทำนายในแต่ละเวลาพยากรณ์
3. ช่วงความเชื่อมั่นของความคลาดเคลื่อนในแต่ละเวลาพยากรณ์

การเปรียบเทียบดังข้างต้น จึงแยกกันระหว่างแบบจำลอง day-ahead และ hour-ahead

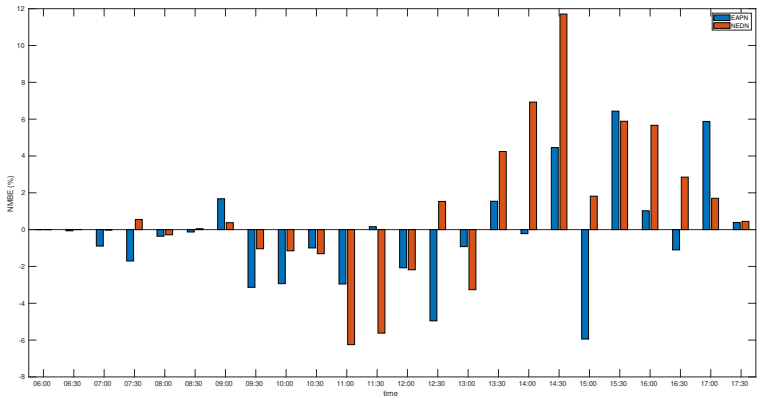
ในส่วนสุดท้ายจะสรุปถึงสมรรถนะโดยรวม และข้อคำนึงถึงในการประยุกต์ใช้ผลลัพธ์ของแบบจำลองเชิงสถิติความคลาดเคลื่อนพยากรณ์นี้

6.1 การเปรียบเทียบระหว่างสองโรงไฟฟ้าของแบบจำลอง day-ahead

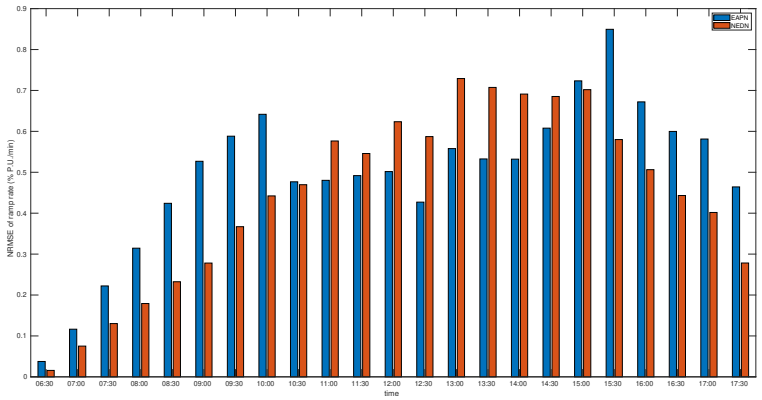
จากผลการเปรียบเทียบ เราจะเห็นว่า ช่วงเวลาการพยากรณ์ที่ค่า NRMSE มีค่าสูงนั้นจะต่างกันไปในแต่ละโรงไฟฟ้า แบบจำลองของโรงไฟฟ้า A ให้ค่า NRMSE ที่สูงในช่วง 8:30-10:00 และ 15:00-16:00 และมีแนวโน้มจะ underestimate มากกว่าโรง B ในช่วงเช้า ส่วนแบบจำลองของโรงไฟฟ้า B ให้ค่า NRMSE ที่สูงในช่วงเวลา 11-12:00 และ 13:30-15:00 และ overestimate มากกว่าในช่วงเวลา 13:30-15:30 นอกจากนี้ เราจะเห็นว่าความคลาดเคลื่อนของ ramp rate ที่มีค่าสูงนั้น เกิดในช่วงเวลาที่ต่างกัน สำหรับสองโรงไฟฟ้า กล่าวคือ โรง A จะมีค่า NRMSE ของ ramp rate ที่สูง ณ เวลา 10:00,15:00-16:00 น. ส่วนโรง B จะมีค่าสูง เริ่มตั้งแต่เวลาบ่ายต้นๆ คือ 13:30-15:00 น.



(a) NRMSE



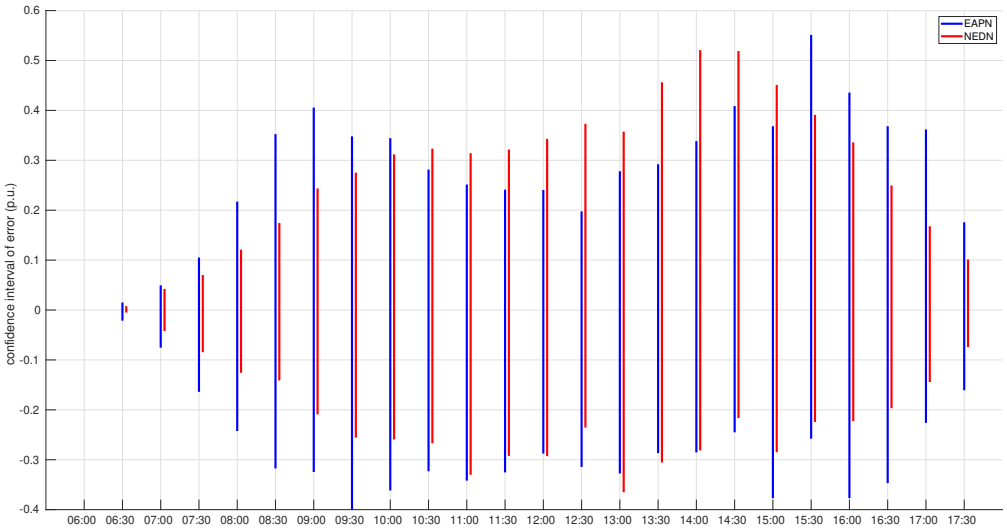
(b) NMBE



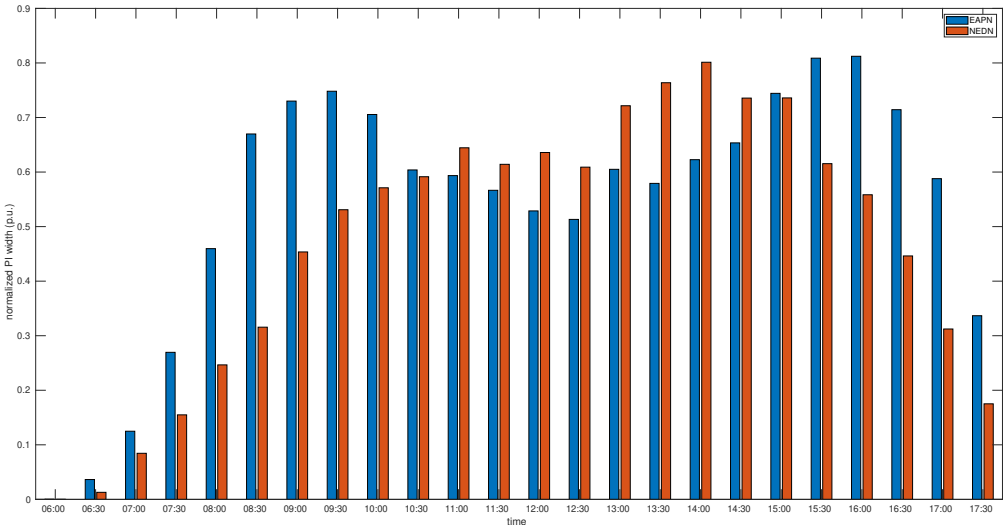
(c) NRMSE of ramp rate

รูป 6.1: การเปรียบเทียบสมรรถนะการพยากรณ์ด้วยแบบจำลอง day-ahead ของสองโรงไฟฟ้า

จุดกลางของช่วงความเชื่อมั่นของความคลาดเคลื่อนบ่งชี้ว่า 90% ของค่าพยากรณ์นั้นมี bias เราจะเห็นว่าโรงไฟฟ้า A มีช่วง CI ของความคลาดเคลื่อนที่สูงในช่วงเช้าและ 15:30-16:30 ส่วนการพยากรณ์ของโรงไฟฟ้า B มีช่วงความเชื่อมั่นของความคลาดเคลื่อนที่กว้างในช่วง 13:00-15:00 ทั้งสองโรงไฟฟ้ามีผลที่ร่วมกันคือ ช่วงเวลาที่แบบจำลองมีสมรรถนะที่แย่ (NRMSE,NMBE มีค่าสูง) จะค่อนข้างเป็นช่วงเวลาเดียวกันที่ช่วงการทำนายมีความกว้างที่สูง



(a) ช่วงความเชื่อมั่นของความคลาดเคลื่อนการพยากรณ์

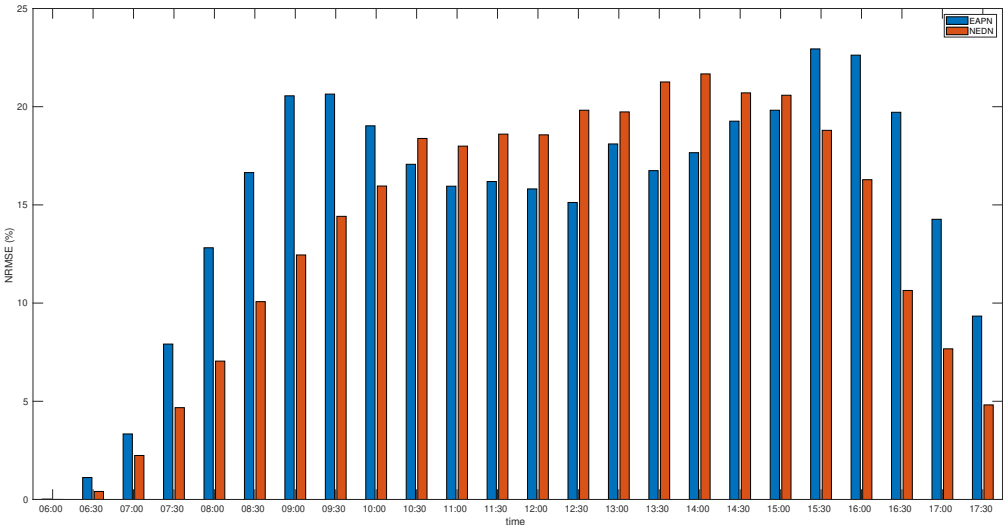


(b) Prediction interval normalized average width

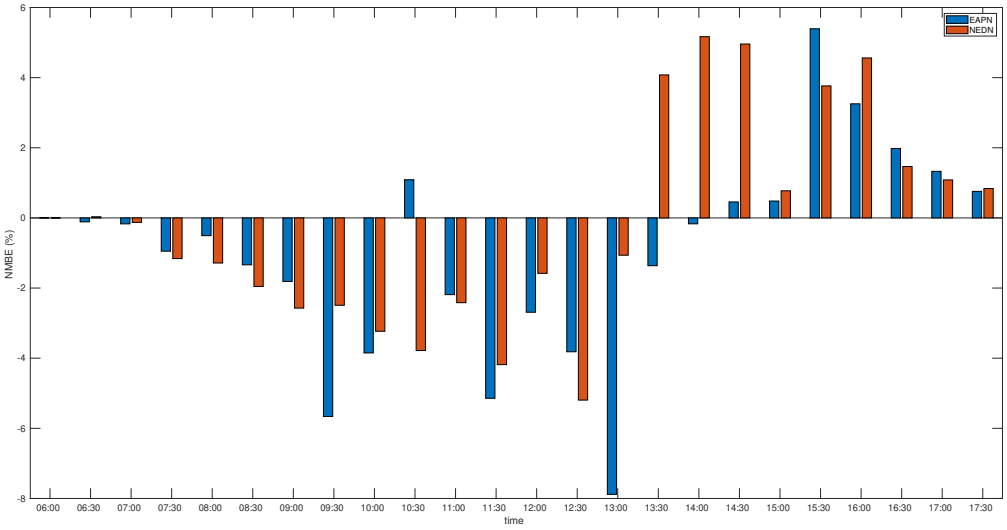
รูป 6.2: การเปรียบเทียบสมบัติของช่วงการทำนายของแบบจำลอง day-ahead ของสองโรงไฟฟ้า

6.2 การเปรียบเทียบระหว่างสองโรงไฟฟ้าของแบบจำลอง hour-ahead

จากผลการเปรียบเทียบ เราจะเห็นว่า ช่วงเวลาการพยากรณ์ที่ค่า NRMSE มีค่าสูงนั้นจะต่างกันไปในแต่ละโรงไฟฟ้า แบบจำลองของโรงไฟฟ้า A ให้ค่า NRMSE ที่สูงในช่วง 9:00-10:00 และ 15:30-16:30 ส่วนแบบจำลองของโรงไฟฟ้า B ให้ค่า NRMSE ที่สูงในช่วงเวลา 13:30-15:30 เมื่อพิจารณา NMBE นั้นแบบจำลองของทั้งสองโรงไฟฟ้ามีแนวโน้มจะ underestimate ในช่วงเช้าจนถึง 13:30 โดยที่โรงไฟฟ้า A มีค่า bias ที่ติดลบด้วยขนาดที่มากกว่า ส่วนช่วงเวลา 13:30-14:30 แบบจำลองของโรง B กลับมา overestimate ด้วยค่า NMBE ที่มากกว่า



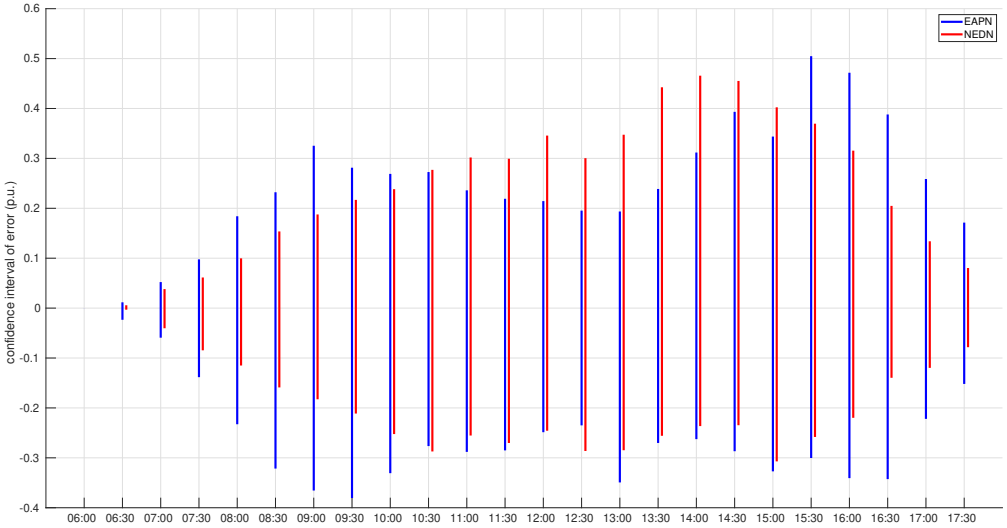
(a) NRMSE



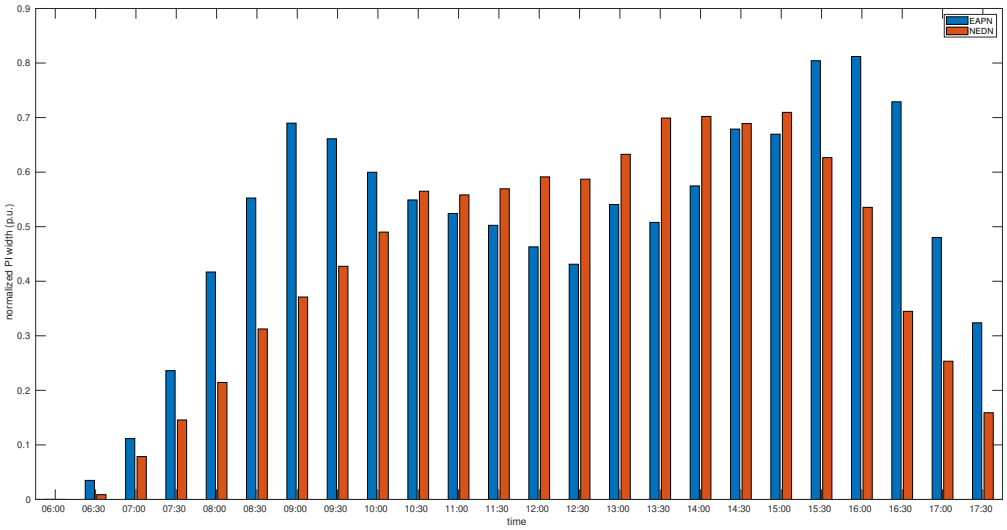
(b) NMBE

รูป 6.3: การเปรียบเทียบสมรรถนะการพยากรณ์ด้วยแบบจำลอง hour-ahead ของสองโรงไฟฟ้า

เราจะเห็นว่าโรงไฟฟ้า A มีช่วง CI ของความคลาดเคลื่อนที่สูงในช่วงเช้า 9:00 และ 14:30-15:00 ส่วนการพยากรณ์ของโรงไฟฟ้า B มีช่วงความเชื่อมั่นของความคลาดเคลื่อนที่กว้างในช่วง 13:30-15:00 ทั้งสองโรงไฟฟ้ามีผลที่ร่วมกันคือ ช่วงเวลาที่แบบจำลองมีสมรรถนะที่แย่ (NRMSE,NMBE มีค่าสูง) จะค่อนข้างเป็นช่วงเวลาเดียวกันที่ช่วงการทำนายมีความกว้างที่สูง



(a) ช่วงความเชื่อมั่นของความคลาดเคลื่อนการพยากรณ์



(b) Prediction interval normalized average width

รูป 6.4: การเปรียบเทียบสมบัติของช่วงการทำนายของแบบจำลอง hour-ahead ของสองโรงไฟฟ้า

6.3 ข้อสรุปสมรรถนะการพยากรณ์

จากการวิเคราะห์ความคลาดเคลื่อนการพยากรณ์กำลังไฟฟ้าแสงอาทิตย์ ด้วยแบบจำลอง Nostradamus ซึ่งเป็น neural-network based ของ Provider เพื่อพยากรณ์ในระยะ hour-ahead (4 ชั่วโมงล่วงหน้า) และระยะ day-ahead (7 วันล่วงหน้า) ที่โรงไฟฟ้าสองแห่ง อันได้แก่ A และ B นั้น ผลการทดลองได้ข้อสรุปเชิงกว้างดังนี้

1. ลักษณะของความคลาดเคลื่อนจากแบบจำลอง day-ahead และ hour-ahead ของโรงไฟฟ้าหนึ่งๆ นั้นมีจุดร่วมกันคือ ค่าสมรรถนะการพยากรณ์ (point forecasts) เช่น NRMSE จะมีค่าสูงที่เวลาใกล้เคียงกัน

2. เมื่อเราพิจารณาระยะการพยากรณ์หนึ่งๆ แต่เปรียบเทียบผลระหว่าง 2 โรงไฟฟ้า พบว่าลักษณะสมบัติของความคลาดเคลื่อนนั้นต่างกัน

3. ปัจจัยที่ทำให้เกิดความต่างกันในสองโรงไฟฟ้านั้น อาจมีหลายประเด็น เช่น GMC inputs ลักษณะการเทรนแบบจำลอง หรือโครงสร้างของแบบจำลอง (เป็นสิ่งที่ Provider ควรพิจารณา) แต่ปัจจัยหนึ่งที่สามารถเห็นได้ชัดจากข้อมูลนั่นคือ การกระจายตัวของค่ากำลังไฟฟ้าที่ผลิตได้ ที่มักจะมีความแปรปรวนสูงในช่วงเวลาที่แบบจำลองที่สมรรถนะการพยากรณ์ที่แย่

4. การประมาณฟังก์ชันการกระจายตัวจากวิธี kernel density estimation นั้น ให้ผลการประมาณที่ดีกว่า การมีสมมติฐานของฟังก์ชันการกระจายตัว (โดยดูจากผล KS test) ซึ่งวิธีหลังนั้น ค่า PI ที่คำนวณได้ เมื่อไปทดสอบกับข้อมูล test set พบว่ามี coverage probability ที่ต่ำกว่าความน่าจะเป็น (ที่กำหนดไว้) ในบางเวลา อันเกิดจากปัญหาการคำนวณเชิงเลขที่ไม่สามารถคำนวณ inverse cumulative function ได้อย่างแม่นยำ สำหรับวิธี kernel และวิธี bootstrap นั้น เมื่อนำมาคำนวณหา PI จะให้ผลไม่ต่างกันมากนัก ทั้งในแง่ความกว้างของช่วง และ coverage probability แต่ว่าช่วง PI ที่ประมาณได้จะมีความต่างเล็กน้อยจากวิธี bootstrap และผลความต่างเห็นชัดขึ้นของข้อมูล hour-ahead ในโรงไฟฟ้า B

5. การประมาณช่วงการทำนายนั้นได้คำนวณจาก training data set และตรวจสอบสมรรถนะของช่วงการทำนายดังกล่าวบน test data set พบว่า ผลการตรวจสอบให้ค่าตัวชี้วัดเช่น coverage probability และ reliability ที่สอดคล้องกับพารามิเตอร์ที่ตั้งไว้ ยกเว้นที่ข้อมูลในเวลา 6:00 และ 6:30 น. ที่ข้อมูลมีการกระจายตัวต่ำมาก ส่งผลให้เกิดความคลาดเคลื่อนเชิงเลข และการประมาณฟังก์ชันการกระจายตัวที่ได้นั้นมี bias กับ training data set มากเกินไป แนวทางที่สามารถทำได้คือ อาจจะละเว้นการวิเคราะห์ของข้อมูลที่เวลาดังกล่าว เนื่องจากความคลาดเคลื่อนการพยากรณ์ของเวลา 6:00-6:30 น. นั้นน้อยมากอยู่แล้ว

ในตาราง 6.1 ได้สรุปช่วงเวลาของแบบจำลองพยากรณ์ย่อย ที่มีสมรรถนะแย่ (ค่า NRMSE สูง) มีช่วงการทำนายที่กว้าง บ่งชี้ถึงการพยากรณ์ขาด และการพยากรณ์เกิน

ตาราง 6.1: ช่วงเวลาที่แบบจำลองการพยากรณ์ย่อยมีสมบัติต่างๆ

Solar plant	Horizon	Large NRMSE	Large PI	Underestimate	Overestimate
A	DA	9:00-10:00, 15:00, 16:00	8:30-10:00, 15:30-16:30	12:30, 15:00	14:30, 15:30, 17:00
A	HA	9:00-10:00, 13:00, 15:30-16:30	9:00-10:00, 14:30-16:30	9:30, 13:00	15:30
B	DA	11:00-12:00, 13:00-15:00	13:00-15:00	11:00, 11:00, 11:30	14:00-14:30
B	HA	13:30-15:00	13:00-15:30	11:30, 12:30	13:30-14:30, 15:30-16:00

6.4 สิ่งที่ต้องคำนึงถึงในการประยุกต์

การศึกษาลักษณะสมบัติเชิงสถิติของความคลาดเคลื่อนการพยากรณ์นั้น มีประโยชน์ในงานประยุกต์ดังตัวอย่างต่อไปนี้

1. เมื่อมีข้อมูลจำนวนมากพอ การกระจายตัวของความคลาดเคลื่อนการพยากรณ์ที่ไม่เป็นไปตาม normal assumption นั้น (รวมถึงการมี non-zero mean) บ่งชี้ว่า แบบจำลองที่ใช้ในการพยากรณ์อาจจะยังมี bias หรือยังอธิบายพลวัตของตัวแปรกำลังผลิตไฟฟ้าได้ไม่ดีพอ จึงสามารถนำข้อสังเกตดังกล่าวไปปรับปรุงแบบจำลองพยากรณ์ต่อไป

2. ฟังก์ชันการกระจายตัวของความคลาดเคลื่อนที่ประมาณได้ (estimated distribution function) สามารถนำไปอธิบายปริมาณทางสถิติอื่นๆ ของความคลาดเคลื่อนได้ เช่น เราอาจจะสนใจ บริเวณข้อมูลที่เป็น outlier ที่ defined ด้วยบริเวณข้อมูลที่มีความน่าจะเป็นที่เกิดนั้นต่ำมาก การ detect outlier ก็เพื่อประกอบการตัดสินใจว่าจะเชื่อถือข้อมูลนั้นขนาดไหน ปริมาณทางสถิติที่นำเสนอในรายงานก็คือช่วง prediction โดยในงานประยุกต์ เราอาจจะให้ความสำคัญกับค่าขอบล่างของ PI มากกว่าขอบบน เพราะจะบ่งชี้ถึงการที่ต้องเตรียมแหล่งจ่ายพลังงานอื่นเสริม หากพลังงานไฟฟ้าแสงอาทิตย์ที่ผลิตได้จริงต่ำกว่าค่าที่พยากรณ์ และอาจจะไม่เพียงพอกับความต้องการของโหลด การวิเคราะห์ดังกล่าว สามารถทำเชื่อมโยงเป็นระบบหลายตัวแปรสุ่ม (multivariate random data) เนื่องจากความต้องการของโหลดก็มีความไม่แน่นอนเช่นกัน มีผลการวิเคราะห์เชิงสถิติของช่วง PI ของโหลดในงานวิจัยที่ผ่านมาเช่นกัน

3. หากจุดประสงค์ของการนำแบบจำลองเชิงสถิติของความคลาดเคลื่อนการพยากรณ์ไปใช้ คือการหา PI เท่านั้น เราสามารถใช้วิธีที่ไม่อิงแบบจำลอง นั่นคือวิธี bootstrap ได้ แต่วิธีนี้ต้องมี resampling data และคำนวณวนซ้ำหลายรอบ (เช่นมากกว่า 1000) ดังนั้น การคำนวณ PI จะไม่ได้ทำแบบ real-time การประมาณการกระจายตัวของความคลาดเคลื่อนก็เช่นเดียวกัน ในแง่ของ implementation นั้น เราต้องมีข้อมูลในอดีตเก็บไว้จำนวนหนึ่ง (เช่นแนะนำว่าอย่างน้อย 1 ปี เพื่อให้ครอบคลุมลักษณะความคลาดเคลื่อนอันเกิดมาจากทุกฤดูอากาศ) การมีข้อมูลที่มาก เพื่อให้ผลการประมาณการกระจายตัวมีความเชื่อถือได้ดี การคำนวณจึงจะเป็นไปในลักษณะ off-line โดยมี input จากผู้ใช้คือ i) ข้อมูลการพยากรณ์และข้อมูลค่าวัดจริงที่เก็บไว้ตั้งแต่อดีต ii) ค่าพารามิเตอร์ coverage probability ที่จะบ่งชี้ถึงช่วง PI และมี output เป็น i) การกระจายตัวที่ประมาณได้ (ไม่ว่าจะเป็น kernel หรือ fitted distribution) และ ii) ช่วง PI ที่ประมาณได้

4. สมบัติเชิงสถิติของความคลาดเคลื่อนนั้น ขึ้นกับ 3 ปัจจัยหลัก อันได้แก่ 1) แบบจำลองพยากรณ์ 2) เวลาของค่าพยากรณ์ (6:00,12:30,etc.) และ 3) k-step prediction ดังที่กล่าวไว้ในหลักการ ในทางปฏิบัตินั้น หากมีการเปลี่ยนแบบจำลองพยากรณ์บน computing server จึงควรมี log เก็บไว้เพื่อทราบว่าระยะเวลาช่วงใด ในแบบจำลองใด มิเช่นนั้น ผลสมรรถนะของความคลาดเคลื่อนที่วิเคราะห์ได้จะคลุมเคลือว่า มาจากแบบจำลองพยากรณ์หรือไม่

5. ในทางปฏิบัติ หากคำนวณ PI ของ ณ เวลาหนึ่งๆ จะใช้ความคลาดเคลื่อนที่รวม (pool) จากมาจากหลายๆ k-step นั้น (อาจจะเนื่องด้วยข้อจำกัดด้านจำนวนข้อมูลในอดีตที่มาวิเคราะห์) จะให้ผลค่า PI ที่กว้างกว่า การใช้ความคลาดเคลื่อนที่มาจาก 1-step prediction แต่กว้างกว่าเพียงเล็กน้อย

Bibliography

[ABCP16] A. Antoniadis, X. Brossat, J. Cugliari, and J.M. Poggi. A prediction interval for a function-valued forecast model: Application to load forecasting. *International Journal of Forecasting*, 32(3):939–947, 2016.

[BDNL08] H. Bludszuweit, J. A. Domínguez-Navarro, and A. Llombart. Statistical analysis of wind power forecast error. *IEEE Transactions on Power Systems*, 23(3):983–991, 2008.

[DPP16] V. Dordonnat, A. Pichavant, and A. Pierrot. Gefcom2014 probabilistic electric load forecasting using time series and semi-parametric regression models. *International Journal of Forecasting*, 32(3): 1005–1011, 2016.

[ET93] B. Efron and R.J. Tibshirani. *An introduction to the bootstrap*. Springer-Science+Business Media, B.V., 1993.

[FH11] S. Fan and R.J. Hyndman. Short-term load forecasting based on a semi-parametric additive model. *IEEE Transactions on Power Systems*, 27(1):134–141, 2011.

[GPG16] F. Golestaneh, P. Pinson, and H.B. Gooi. Very short-term nonparametric probabilistic forecasting of energy generation: With application to solar energy. *IEEE Transactions on Power Systems*, 31(5): 3850–3863, 2016.

[HF16] T. Hong and S. Fan. Probabilistic electric load forecasting: A tutorial review. *International Journal of Forecasting*, 32(3):914–938, 2016.

[HM11] B. Hodge and M. Milligan. Wind power forecasting error distributions over multiple timescales. In *Power and Energy Society General Meeting, 2011 IEEE*, pages 1–8. IEEE, 2011.

[HTF09] T. Hastie, R. Tibshirani, and J. Friedman. *The Elements of Statistical Learning: Data Mining, Inference and Prediction*. Springer, 2nd edition, 2009.

[JKB70] N. L. Johnson, S. Kotz, and N. Balakrishnan. *Continuous univariate distributions: volume 2*, volume 2. Houghton Mifflin Boston, 2 edition, 1970.

[JWHT13] G. James, D. Witten, T. Hastie, and R. Tibshirani. *An Introduction to statistical learning*, volume 112. Springer, 2013.

[KNC13] A. Khosravi, S. Nahavandi, and D. Creighton. Prediction intervals for short-term wind farm power generation forecasts. *IEEE Transactions on sustainable energy*, 4(3):602–610, 2013.

[KNCA11] A. Khosravi, S. Nahavandi, D. Creighton, and A.F. Atiya. Comprehensive review of neural network-based prediction intervals and new advances. *IEEE Transactions on neural networks*, 22(9):1341–1356, 2011.

[LHHB09] E. Lorenz, J. Hurka, D. Heinemann, and H.G. Beyer. Irradiance forecasting for the power prediction of grid-connected photovoltaic systems. *IEEE Journal of selected topics in applied earth observations and remote sensing*, 2(1):2–10, 2009.

[Lov11] M. Lovric. *International Encyclopedia of Statistical Science*. Springer, 2011.

[Nol18] J.P. Nolan. Stable distributions: Models for heavy tailed data. Technical report, American University, 2018.

[NW18] J. Nowotarski and R. Weron. Recent advances in electricity price forecasting: A review of probabilistic forecasting. *Renewable and Sustainable Energy Reviews*, 81:1548–1568, 2018.

[PNM+07] P. Pinson, H.A. Nielsen, J.K. Møller, H. Madsen, and G.N. Kariniotakis. Non-parametric probabilistic forecasts of wind power: required properties and evaluation. *Wind Energy: An International Journal for Progress and Applications in Wind Power Conversion Technology*, 10(6):497–516, 2007.

[Sil98] B.W. Silverman. *Density estimation for statistics and data analysis*. Chapman & Hall/CRC, 1998.

[SM00] A. P. Da Silva and L. S. Moulin. Confidence intervals for neural network based short-term load forecasting. *IEEE Transactions on Power Systems*, 15(4):1191–1196, 2000.

[WXP+14] C. Wan, Z. Xu, P. Pinson, Z. Dong, and K.P. Wong. Optimal prediction intervals of wind power generation. *IEEE Transactions on Power Systems*, 29(3):1166–1174, 2014.

[ZMAP14] M. Zamo, O. Mestre, P. Arbogast, and O. Pannekoucke. A benchmark of statistical regression methods for short-term forecasting of photovoltaic electricity production. part ii: Probabilistic forecast of daily production. *Solar Energy*, 105:804–816, 2014.