การเปรียบเทียบประสิทธิภาพของตัวแบบสำหรับข้อมูลจำนวนนับที่มีปัญหาการกระจายและค่าศูนย์เฟ้อ

โดย

นายธนสิทธิ์ พูลสมบัติ

วิทยานิพนธ์นี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตร
วิทยาศาสตรมหาบัณฑิต (สถิติประยุกต์)
สาขาวิชาคณิตศาสตร์และสถิติ
คณะวิทยาศาสตร์และเทคโนโลยี มหาวิทยาลัยธรรมศาสตร์
ปีการศึกษา 2563

การเปรียบเทียบประสิทธิภาพของตัวแบบสำหรับข้อมูลจำนวนนับที่มีปัญหาการกระจายและค่าศูนย์เฟ้อ

โดย

นายธนสิทธิ์ พูลสมบัติ

วิทยานิพนธ์นี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตร
วิทยาศาสตรมหาบัณฑิต (สถิติประยุกต์)
สาขาวิชาคณิตศาสตร์และสถิติ
คณะวิทยาศาสตร์และเทคโนโลยี มหาวิทยาลัยธรรมศาสตร์
ปีการศึกษา 2563

A COMPARISON STUDY OF MODEL EFFICIENCY FOR DISPERSION PROBLEMS AND ZERO-INFLATED COUNT DATA

BY

MR.TANASIT POOLSOMBUT

A THESIS SUBMITTED IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR THE DEGREE OF MASTER OF SCIENCE (APPLIED STATISTICS)
DEPARTMENT OF MATHEMATICS AND STATISTICS
FACULTY OF SCIENCE AND TECHNOLOGY
THAMMASAT UNIVERSITY
ACADEMIC YEAR 2020

มหาวิทยาลัยธรรมศาสตร์
คณะวิทยาศาสตร์และเทคโนโลยี

วิทยานิพนธ์

ของ

นายธนสิทธิ์ พูลสมบัติ

เรื่อง

การเปรียบเทียบประสิทธิภาพของตัวแบบสำหรับข้อมูลจำนวนนับที่มีปัญหาการกระจาย
และค่าศูนย์เฟ้อ

ได้รับการตรวจสอบและอนุมัติ ให้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตร
วิทยาศาสตรมหาบัณฑิต (สถิติประยุกต์)

เมื่อ วันที่ 29 มิถุนายน พ.ศ. 2564

ประธานกรรมการสอบวิทยานิพนธ์ _____
(รองศาสตราจารย์ ดร.อัชฌา อระวีพร)

กรรมการและอาจารย์ที่ปรึกษาวิทยานิพนธ์หลัก _____
(ผู้ช่วยศาสตราจารย์ ดร.ธีระวัฒน์ สิมมาจันทร์)

กรรมการและอาจารย์ที่ปรึกษาวิทยานิพนธ์ร่วม _____
(ผู้ช่วยศาสตราจารย์ ดร.แสงดาว วงค์สาย)

กรรมการสอบวิทยานิพนธ์ _____
(ผู้ช่วยศาสตราจารย์ ดร.พัทธ์ชนก ศรีสุรเดชชัย)

กรรมการสอบวิทยานิพนธ์ _____
(ผู้ช่วยศาสตราจารย์ ดร.ภทรวรรณ แสงนวกิจ)

คณบดี _____
(รองศาสตราจารย์ ดร.ณัฐธนนท์ หงส์วริทธิ์ธร)

หัวข้อวิทยานิพนธ์	การเปรียบเทียบประสิทธิภาพของตัวแบบสำหรับข้อมูลจำนวนนับที่มีปัญหาการกระจายและค่าศูนย์เฟ้อ
ชื่อผู้เขียน	นายธนสิทธิ์ พูลสมบัติ
ชื่อปริญญา	วิทยาศาสตรมหาบัณฑิต (สถิติประยุกต์)
สาขาวิชา/คณะ/มหาวิทยาลัย	สาขาวิชาคณิตศาสตร์และสถิติ คณะวิทยาศาสตร์และเทคโนโลยี มหาวิทยาลัยธรรมศาสตร์
อาจารย์ที่ปรึกษาวิทยานิพนธ์	ผู้ช่วยศาสตราจารย์ ดร.ธีระวัฒน์ สิมมาจันทร์
อาจารย์ที่ปรึกษาวิทยานิพนธ์ร่วม	ผู้ช่วยศาสตราจารย์ ดร.แสงดาว วงค์สาย
ปีการศึกษา	2563

บทคัดย่อ

งานวิจัยนี้มีวัตถุประสงค์เพื่อเปรียบเทียบประสิทธิภาพการพยากรณ์ของตัวแบบสำหรับจำนวนนับจากข้อมูลจริงและข้อมูลจำลองที่มีปัญหาการกระจายและค่าศูนย์เฟ้อ โดยแบ่งเป็นสองกรณี คือ การกระจายต่ำกว่าเกณฑ์ ได้แก่ ข้อมูลอุบัติเหตุทางถนนที่ส่งผลให้มีผู้เสียชีวิตในประเทศไทย ปี พ.ศ. 2558 และการกระจายเกินเกณฑ์ ได้แก่ ข้อมูลการเรียกร้องค่าสินไหมทดแทนจากบริษัทประกันภัยรถยนต์แห่งหนึ่งในประเทศไทย ปี พ.ศ. 2560 ตัวแบบพยากรณ์ที่ใช้มี 5 ตัวแบบ ได้แก่ ควอไซปัวซง (QP), คอนเวย์แม็กซ์เวลล์ปัวซง (CMP), ปัวซงค่าศูนย์เฟ้อ (ZIP), ทวินามเชิงลบค่าศูนย์เฟ้อ (ZINB) และเทคนิคป่าสุ่ม (RF) การสร้างข้อมูลจำลองใช้ค่าสัมประสิทธิ์การถดถอยของตัวแบบคอนเวย์แม็กซ์เวลล์ปัวซงค่าศูนย์เฟ้อ (ZICMP) ที่ได้จากการวิเคราะห์ข้อมูลจริงเป็นตัวตั้งต้น ขนาดตัวอย่าง ได้แก่ 250, 500, 1,000, 3,000 และ 5,000 และเพื่อศึกษาอิทธิพลของการเพิ่มขึ้นของค่าสัมประสิทธิ์การถดถอย คือ สัมประสิทธิ์การถดถอยจากข้อมูลจริงและสัมประสิทธิ์การถดถอยที่เพิ่มขึ้น 50 เปอร์เซ็นต์ การเปรียบเทียบประสิทธิภาพของตัวแบบใช้วิธีเค-โฟลด์ตรวจสอบไขว้ เกณฑ์การพิจารณา คือ ค่าเฉลี่ยของรากของค่าคลาดเคลื่อนกำลังสองเฉลี่ย (mRMSE) ผลการจำลอง พบว่าเมื่อค่าสัมประสิทธิ์การถดถอยเพิ่มขึ้นส่งผลให้ประสิทธิภาพของทุกตัวแบบดีขึ้น กรณีการกระจายต่ำกว่าเกณฑ์ พบว่า CMP มีประสิทธิภาพดีกว่า RF โดยเฉพาะที่ขนาดตัวอย่างน้อย (250) กรณีการกระจายเกินเกณฑ์ พบว่า เมื่อขนาดตัวอย่างน้อยตัวแบบมีประสิทธิภาพต่างกัน แต่เมื่อขนาดตัวอย่างเพิ่มทุกตัวแบบมีประสิทธิภาพใกล้เคียงกัน ยกเว้น RF ซึ่งให้ผลสรุปสอดคล้องกับข้อมูลจริง

คำสำคัญ: ตัวแบบสำหรับจำนวนนับ, เทคนิคป่าสุ่ม, การกระจายต่ำกว่าเกณฑ์, การกระจายเกินเกณฑ์, จำนวนนับ, การมีค่าศูนย์เฟ้อ

Thesis Title	A COMPARISON STUDY OF MODEL EFFICIENCY FOR DISPERSION PROBLEMS AND ZERO-INFLATED COUNT DATA
Author	Mr.Tanasit Poolsombut
Degree	Master of Science (Applies Statistics)
Department/Faculty/University	Mathematics and Statistics Faculty of science and Technology Thammasat University
Thesis Advisor	Assistant Professor Teerawat Simmachan, Ph.D.
Thesis Co-Advisor	Assistant Professor Sangdao Wongsai, Ph.D.
Academic Year	2020

ABSTRACT

The objective of this research was to compare the efficiency of prediction models for count data from simulations and real data with dispersion and zero inflated problems. Data with under-dispersion is the number of deaths from road traffic accidents in Thailand in 2015 and data with over-dispersion is the claim counts from a car insurance company in Thailand in 2017. Five prediction models considered were Quasi-Poisson (QP), Conway Maxwell Poisson (CMP), Zero-inflated Poisson (ZIP), Zero-inflated Negative binomial (ZINB) and Random Forest (RF). The simulations were based on the Zero-inflated Conway Maxwell Poisson regression model (ZICMP) regression coefficients from the model fitting to the real data. To study the impact of a change in the regression coefficient, a set of regression coefficients with an increase of 50 percent was simulated. The size of the sample ranged from 250, 500, 1,000, 3,000 to 5,000. The K-fold cross-validation approach was used to compare the model efficiency by a criterion of a mean of root mean square errors. The results of the simulations show that the efficiency of all models was improved when the regression coefficient was increased. The CMP outperformed RF for under-dispersion, especially when the sample size was small (250). For over-dispersion with a small sample size, there was a difference in model performance.

However, when the sample size increases, the models performed in the same way, except for RF, providing the least efficiency. These results were consistent with those obtained from the real data.

Keywords: Count model, Random Forest, Under-dispersion, Over-dispersion, Count data, Zero-inflation.

กิตติกรรมประกาศ

วิทยานิพนธ์ฉบับนี้สำเร็จลุล่วงได้ อันเนื่องมาจากความกรุณาจากอาจารย์ที่ปรึกษาวิทยานิพนธ์ ผู้ช่วยศาสตราจารย์ ดร.ธีระวัฒน์ สิมมาจันทร์ และ ผู้ช่วยศาสตราจารย์ ดร.แสงดาว วงค์สาย ที่กรุณาให้คำแนะนำ ให้คำปรึกษาที่เป็นประโยชน์ต่องานวิจัยในครั้งนี้ตลอดจนช่วยเหลือและแก้ไขข้อบกพร่องต่าง ๆ จนกระทั่งวิทยานิพนธ์ฉบับนี้สำเร็จลุล่วงไปได้ด้วยดี ทางผู้วิจัยจึงขอ กราบขอบพระคุณ ไว้ ณ ที่นี้

ผู้วิจัยขอกราบขอบพระคุณ รองศาสตราจารย์ ดร.อัชฌา อระวีพร ในฐานะประธานกรรมการและกรรมการสอบวิทยานิพนธ์ ผู้ช่วยศาสตราจารย์ ดร.พัทธ์ชนก ศรีสุรเดชชัย และผู้ช่วยศาสตราจารย์ ดร.ภทรวรรณ แสงนวกิจ ที่กรุณาสละเวลา ให้ข้อเสนอแนะตลอดจนตรวจสอบข้อบกพร่องต่าง ๆ เพื่อการปรับแก้วิทยานิพนธ์ฉบับนี้ให้สมบูรณ์ยิ่งขึ้น

สุดท้ายนี้ผู้วิจัยขอกราบขอบพระคุณกรมป้องกันและบรรเทาสาธารณภัยที่ให้ความอนุเคราะห์ข้อมูลอันเป็นประโยชน์ในงานวิจัยฉบับนี้ รวมถึงบิดา มารดา ตลอดจนญาติพี่น้องและผู้ที่เกี่ยวข้องทั้งหลายที่ส่งเสริมและสนับสนุนด้านการศึกษาแก่ผู้วิจัยเสมอมา

นายธนสิทธิ์ พูลสมบัติ

สารบัญ

	หน้า
บทคัดย่อภาษาไทย	(1)
บทคัดย่อภาษาอังกฤษ	(2)
กิตติกรรมประกาศ	(4)
สารบัญ	(5)
สารบัญตาราง	(9)
สารบัญภาพ	(10)
รายการสัญลักษณ์และคำย่อ	(12)
บทที่ 1 บทนำ	1
1.1 ที่มาและความสำคัญ	1
1.2 วัตถุประสงค์ของการศึกษา	5
1.3 ขอบเขตการศึกษา	5
1.3.1 ขอบเขตของตัวแบบและโปรแกรมที่ใช้	5
1.3.2 ขอบเขตของข้อมูลจริง	5
1.3.3 ขอบเขตของข้อมูลจำลอง	7
1.4 เกณฑ์ที่ใช้ในการพิจารณา	8
1.4.1 รากของค่าคลาดเคลื่อนกำลังสองเฉลี่ย	8
1.4.2 ส่วนเบี่ยงเบนมาตรฐาน	9
1.5 ประโยชน์ที่คาดว่าจะได้รับ	11

บทที่ 2 วรรณกรรมและงานวิจัยที่เกี่ยวข้อง 12

2.1 ตัวแบบที่เกี่ยวข้อง 12
2.1.1 ตัวแบบการถดถอยปัวซง 12
2.1.2 ตัวแบบการถดถอยควอไซปัวซง 13
2.1.3 ตัวแบบการถดถอยปัวซงค่าศูนย์เฟ้อ 14
2.1.4 ตัวแบบการถดถอยทวินามเชิงลบค่าศูนย์เฟ้อ 16
2.1.5 ตัวแบบการถดถอยคอนเวย์แม็กซ์เวลล์ปัวซง 17
2.1.6 ตัวแบบการถดถอยคอนเวย์แม็กซ์เวลล์ปัวซงค่าศูนย์เฟ้อ 20
2.1.7 การเรียนรู้ของเครื่อง 21
2.1.7.1 ต้นไม้ตัดสินใจ 22
2.1.7.2 เทคนิคป่าสุ่ม 24
2.1.7.3 กระบวนการของเทคนิคป่าสุ่ม 24
2.1.7.4 การระบุพารามิเตอร์ในเทคนิคป่าสุ่ม 25
2.1.7.5 ค่าคลาดเคลื่อนกำลังสองเฉลี่ยของ OOB 26
2.1.7.6 หลักการเลือกตัวแปรอิสระเพื่อแบ่งโหนดการตัดสินใจ 27
2.1.7.7 การหาจุดแบ่งที่ดีที่สุด 28
2.2 สถิติทดสอบสกอร์ 32
2.3 การทดสอบการกระจาย 33
2.4 เกณฑ์สารสนเทศของอะกะอิเกะ 33
2.5 วิธีวัดประสิทธิภาพ 34
2.6 งานวิจัยที่เกี่ยวข้อง 35

บทที่ 3 วิธีการวิจัย 41

3.1 การวิเคราะห์ข้อมูลกรณีข้อมูลมีการกระจายต่ำกว่าเกณฑ์และค่าศูนย์เฟ้อ 41
3.1.1 การคัดเลือกตัวแปรอิสระโดยใช้ตัวแบบ ZICMP 41
3.1.2 การวิเคราะห์ข้อมูลจริง 45
3.1.3 การวิเคราะห์ข้อมูลจำลอง 45
3.2 การวิเคราะห์ข้อมูลกรณีข้อมูลมีการกระจายเกินเกณฑ์และค่าศูนย์เฟ้อ 47

3.2.1 การจำลองตัวแปรเชิงปริมาณ 47

บทที่ 4 ผลการวิจัยและอภิปรายผล 53

4.1 ผลการวิจัยกรณีการกระจายต่ำกว่าเกณฑ์และค่าศูนย์เฟ้อ 53
4.1.1 ผลการเปรียบเทียบประสิทธิภาพของตัวแบบสำหรับข้อมูลจริง 53
4.1.2 ผลการเปรียบเทียบประสิทธิภาพของตัวแบบสำหรับข้อมูลจำลอง 55
4.2 ผลการวิจัยกรณีการกระจายเกินเกณฑ์และค่าศูนย์เฟ้อ 63
4.2.1 ผลการเปรียบเทียบของตัวแบบสำหรับข้อมูลจริง 63
4.2.2 ผลการเปรียบเทียบของตัวแบบสำหรับข้อมูลจำลอง 66

บทที่ 5 สรุปผลการวิจัยและข้อเสนอแนะ 75

5.1 สรุปผลการวิจัย 75
5.2 ข้อจำกัดและข้อเสนอแนะ 78

รายการอ้างอิง 79

ภาคผนวก

ภาคผนวก โปรแกรมสำหรับจำลองข้อมูลในงายวิจัย 86

ประวัติผู้เขียน 157

สารบัญตาราง

ตารางที่	หน้า
1.1 ตัวแปรของข้อมูลการเกิดอุบัติเหตุทางถนนทั่วประเทศไทย ปี พ.ศ. 2558 (n = 4,666)	6
1.2 ตัวแปรของข้อมูลเรียกร้องค้าสินไหมทดแทนของลูกค้า ปี พ.ศ. 2560 (n = 2,991)	7
3.1 ตารางวิเคราะห์ข้อมูลจำนวนผู้เสียชีวิตจากอุบัติเหตุทางถนนทั่วประเทศไทย เดือนเมษายน ปี พ.ศ. 2558 จากตัวแบบการถดถอย ZICMP (n=4,666)	44
3.2 สัดส่วนของลักษณะที่สนใจในแต่ละตัวแปรอิสระที่มีผลต่อการประมาณค่าพารามิเตอร์	46
3.3 ตารางวิเคราะห์ข้อมูลการเรียกร้องค่าสินไหมทดแทนจากข้อมูลกรมธรรม์ประกันภัยรถยนต์ ปี พ.ศ. 2560 จากการตัวแบบการถดถอย ZICMP (n = 2,991)	50
4.1 สถิติพรรณนาข้อมูลอุบัติเหตุทางถนนที่ส่งผลให้มีผู้เสียชีวิตทั่วประเทศไทย เดือนเมษายน ปี พ.ศ. 2558 (n = 4,666)	54
4.2 ผลการเปรียบเทียบประสิทธิภาพการพยากรณ์ของตัวแบบการถดถอย CMP และเทคนิค RF กรณีข้อมูลอุบัติเหตุทางถนนที่ส่งผลให้มีผู้เสียชีวิตทั่วประเทศไทย เดือนเมษายน ปี พ.ศ. 2558	55
4.3 ค่า mRMSE ของตัวแบบการถดถอย CMP และเทคนิค RF จากข้อมูลจำลองชุดที่ 14.7 ตารางวิเคราะห์ข้อมูลการเรียกร้องค่าสินไหมทดแทนจากข้อมูลกรมธรรม์ประกันภัยรถยนต์ ปี พ.ศ. 2560	56
4.4 ค่า mRMSE ของตัวแบบการถดถอย CMP และเทคนิค RF จากข้อมูลจำลองชุดที่ 2	56
4.5 สถิติพรรณนาการเรียกร้องค้าสินไหมทดแทนของลูกค้า ปี พ.ศ. 2560 (n = 2,991)	64
4.6 ผลการเปรียบเทียบประสิทธิภาพการพยากรณ์ของตัวแบบ กรณีศึกษาข้อมูลการเรียกร้องค่าสินไหมจากกรมธรรม์ประกันภัย (n=2,991)	65
4.7 ค่า mRMSE ของตัวแบบการถดถอย QP, CMP, ZIP, ZINB และเทคนิค RF จากข้อมูลจำลองชุดที่ 1	66

4.8 ค่า mRMSE ของตัวแบบการถดถอย QP, CMP, ZIP, ZINB และเทคนิค RF จากข้อมูลจำลองชุดที่ 2 67

1ก ค่าพยากรณ์จำนวนผู้เสียชีวิตจากอุบัติเหตุทั่วประเทศไทยจากข้อมูลชุดทดสอบ (n=933) ของตัวแบบ CMP และ RF กรณีข้อมูลจริงที่มีการกระจายตัวต่ำกว่าเกณฑ์และค่าศูนย์เฟ้อ 104

2ก ค่าพยากรณ์จำนวนครั้งการเรียกร้องค่าสินไหมทดแทนของลูกค้าของบริษัทประกันภัยแห่งหนึ่งจากข้อมูลชุดทดสอบ (n=597) กรณีข้อมูลจริงที่มีการกระจายตัวเกินเกณฑ์และค่าศูนย์เฟ้อ 136

สารบัญภาพ

ภาพที่	หน้า
2.1 โครงสร้างต้นไม้ตัดสินใจ	23
2.2 กระบวนการของเทคนิคป่าสุ่ม	24
2.3 ความสำคัญของตัวแปร	28
2.4 (a) ขั้นตอนการหาจุดแบ่งที่ดีที่สุด (Step 1)	30
2.4 (b) ขั้นตอนการหาจุดแบ่งที่ดีที่สุด (Step 1)	30
2.4 (c) ขั้นตอนการหาจุดแบ่งที่ดีที่สุด (Step 2)	31
2.4 (d) ขั้นตอนการหาจุดแบ่งที่ดีที่สุด (Step 3)	31
2.4 (e) ขั้นตอนการหาจุดแบ่งที่ดีที่สุด (Step 4)	32
2.5 กระบวนการเค-โฟลด์ตรวจสอบไขว้	34
3.1 ขั้นตอนการวิเคราะห์ข้อมูลจริงกรณีการกระจายต่ำกว่าเกณฑ์	42
3.2 ขั้นตอนการวิเคราะห์ข้อมูลจำลองกรณีการกระจายต่ำกว่าเกณฑ์	43
3.3 ขั้นตอนการวิเคราะห์ข้อมูลจริงกรณีการกระจายเกินเกณฑ์	48
3.4 ขั้นตอนการวิเคราะห์ข้อมูลจำลองกรณีการกระจายเกินเกณฑ์	49
3.5 การแจกแจงเบอร์และค่าพารามิเตอร์ของตัวแปรเบี้ยประกันภัยจากโปรแกรม Easy Fit	51
3.6 การแจกแจงเบิร์นบัม-แซนเดอร์และค่าพารามิเตอร์ของตัวแปรส่วนลดประวัติดีจากโปรแกรม Easy Fit	52
4.1 แผนภูมิแสดงความถี่ของการเกิดอุบัติเหตุทางถนนที่ส่งผลให้มีผู้เสียชีวิตทั่วประเทศไทย เดือนเมษายน ปี พ.ศ. 2558	54
4.2 ค่า mRMSE ของตัวแบบการถดถอย CMP และเทคนิค RF เมื่อขนาดตัวอย่าง n = 250 500 1,000 3,000 และ 5,000 จากข้อมูลจำลองชุดที่ 1	57
4.3 ค่า mRMSE ของตัวแบบการถดถอย CMP และเทคนิค RF เมื่อขนาดตัวอย่าง n = 250 500 1,000 3,000 และ 5,000 จากข้อมูลจำลองชุดที่ 2	58
4.4 ค่า mRMSE ของตัวแบบการถดถอย CMP และเทคนิค RF เมื่อขนาดตัวอย่าง n = 250 500 1,000 3,000 และ 5,000 ข้อมูลจำลองชุดที่ 1 เปรียบเทียบกับข้อมูลจำลองชุดที่ 2	59
4.5 ค่า mRMSE ของตัวแบบการถดถอย CMP และเทคนิค RF ที่ค่า $\overline{\lambda_i}$ ต่าง ๆ ของข้อมูล	60

4.6 ค่า mRMSE ของตัวแบบการถดถอย CMP และเทคนิค RF ที่ค่า $\overline{\pi}_i$ ต่าง ๆ ของข้อมูลจำลองชุดที่ 1 เปรียบเทียบกับข้อมูลจำลองชุดที่ 2 61

4.7 ค่า $\overline{\lambda}_i$ ที่ขนาดตัวอย่างต่าง ๆ ของข้อมูลจำลองชุดที่ 1 เปรียบเทียบกับข้อมูลจำลองชุดที่ 2 62

4.8 ค่า $\overline{\pi}_i$ ที่ขนาดตัวอย่างต่าง ๆ ของข้อมูลจำลองชุดที่ 1 เปรียบเทียบกับข้อมูลจำลองชุดที่ 2 63

4.9 แผนภูมิแสดงความถี่การเรียกร้องค่าสินไหมทดแทนจากข้อมูลกรมธรรม์ ปี พ.ศ. 2560 64

4.10 ค่า mRMSE ของตัวแบบทั้ง 5 เมื่อขนาดตัวอย่าง n = 250 500 1,000 3,000 และ 5,000 จากข้อมูลจำลองชุดที่ 1 68

4.11 ค่า mRMSE ของตัวแบบทั้ง 5 เมื่อขนาดตัวอย่าง n = 250 500 1,000 3,000 และ 5,000 จากข้อมูลจำลองที่ใช้ค่าสัมประสิทธิ์การถดถอยของข้อมูลจริงเพิ่มขึ้น 50 เปอร์เซ็นต์ 69

4.12 ค่า mRMSE ของตัวแบบทั้ง 5 เมื่อขนาดตัวอย่าง n = 250 500 1,000 3,000 และ 5,000 ของข้อมูลจำลองชุดที่ 1 เปรียบเทียบกับข้อมูลจำลองชุดที่ 2 70

4.13 ค่า mRMSE ของตัวแบบทั้ง 5 ที่ค่า $\overline{\lambda}_i$ ต่าง ๆของข้อมูลจำลองชุดที่ 1 เปรียบเทียบกับข้อมูลจำลองชุดที่ 2 71

4.14 ค่า mRMSE ของตัวแบบทั้ง 5 ที่ค่า $\overline{\pi}_i$ ต่าง ๆ ของข้อมูลจำลองชุดที่ 1 เปรียบเทียบกับข้อมูลจำลองชุดที่ 2 72

4.15 ค่า $\overline{\lambda}_i$ ที่ขนาดตัวอย่างต่าง ๆ ของข้อมูลจำลองชุดที่ 1 เปรียบเทียบกับข้อมูลจำลองชุดที่ 2 73

4.16 ค่า $\overline{\pi}_i$ ที่ขนาดตัวอย่างต่าง ๆ ของข้อมูลจำลองชุดที่ 1 เปรียบเทียบกับข้อมูลจำลองที่ชุดที่ 2 74

รายการสัญลักษณ์และคำย่อ

สัญลักษณ์/คำย่อ	คำเต็ม/คำจำกัดความ
QP	ตัวแบบการถดถอยควอไซปัวซง
CMP	ตัวแบบการถดถอยคอนเวย์แม็กซ์เวลล์ปัวซง
ZIP	ตัวแบบการถดถอยปัวซงค่าศูนย์เฟ้อ
$ZINB$	ตัวแบบการถดถอยทวินามเชิงลบค่าศูนย์เฟ้อ
$ZICMP$	ตัวแบบการถดถอยคอนเวย์แม็กซ์เวลล์ปัวซงค่าศูนย์เฟ้อ
RF	เทคนิคป่าสุ่ม
n	ขนาดตัวอย่าง
λ	พารามิเตอร์ค่าเฉลี่ยของการแจกแจงปัวซง โดย $\lambda > 0$
ν	ตัวประมาณพารามิเตอร์การกระจาย โดยมีค่า $\nu > 0$ หาก $\nu < 1$ กรณีที่เป็น Over-dispersion และ $\nu > 1$ กรณีที่เป็น Under-dispersion
θ	พารามิเตอร์การกระจายของการแจกแจงควอ-ไซปัวซง QP
α	พารามิเตอร์การกระจายของ NB โดย $\alpha > 0$
μ_{CMP}	พารามิเตอร์ปรับใหม่ของ CMP โดย $\mu_{CMP} = \lambda^{1/\nu}$
β_0	ค่าคงที่สัมประสิทธิ์สมการถดถอย
β_i	ค่าสัมประสิทธิ์สมการถดถอยที่ i
y	ค่าสังเกต
$\hat{y}$	ค่าพยากรณ์
$\bar{y}$	ค่าเฉลี่ยของค่าสังเกต
π	ความน่าจะเป็นที่จะเกิดศูนย์

p	จำนวนตัวแปรอิสระ
D	สเกลที่แสดงฟังก์ชันความสัมพันธ์ระหว่างความแปรปรวนและค่าเฉลี่ยของการแจกแจงทวินามเชิงลบ
m	จำนวนตัวแปรอิสระที่ถูกสุ่ม
C_i	จุดแบ่งจุดตัดที่ดีที่สุดที่ตำแหน่งที่ i
MSE_{OOB}	ค่าคลาดเคลื่อนกำลังสองเฉลี่ยของ OOB
K	จำนวนข้อมูลที่ถูกแบ่งกลุ่มเพื่อทำตรวจสอบไขว้
q	ความน่าจะเป็นที่จะเกิดลักษณะที่สนใจ
M	จำนวนรอบการทำซ้ำในการจำลอง
$\overline{\lambda}_i$	ค่าเฉลี่ยของพารามิเตอร์ค่าเฉลี่ยที่ค่าสังเกตที่ i
$\overline{\nu}_i$	ค่าเฉลี่ยของพารามิเตอร์การกระจายที่ค่าสังเกตที่ i
$\overline{\pi}_i$	ค่าเฉลี่ยของพารามิเตอร์ความน่าจะเป็นที่จะเกิดศูนย์ที่ค่าสังเกตที่ i

บทที่ 1
บทนำ

1.1 ที่มาและความสำคัญ

ในการพยากรณ์ตัวแปรตอบสนองที่เป็นตัวแปรต่อเนื่องโดยใช้ตัวแปรอิสระเป็นตัวแปรอธิบายโดยทั่วไป การวิเคราะห์ที่ใช้ คือ ตัวแบบการวิเคราะห์การถดถอยเชิงเส้น (Linear regression model) หากตัวแปรตอบสนองเป็นจำนวนนับ (Count data) ซึ่งเป็นตัวแปรสุ่มไม่ต่อเนื่อง (Discrete random variable) ตัวแบบดังกล่าวจะไม่สามารถใช้ได้ เมื่อตัวแปรที่สนใจเป็นข้อมูลจำนวนนับและถูกเก็บรวบรวมในขอบเขตหรือช่วงเวลาที่ศึกษา อาทิเช่น จำนวนอุบัติเหตุบนถนนในแต่ละเดือน, จำนวนการขอรับสินไหมทดแทนจากอุบัติเหตุทางรถยนต์ของบริษัทประกันภัยแห่งรายไตรมาส, จำนวนผู้ป่วยที่เป็นโรค COVID-19 ที่มารับการรักษาที่โรงพยาบาลเฉลิมพระเกียรติในปี 2562 และจำนวนสินค้าที่ผลิตไม่ได้มาตรฐานตามที่กำหนดในรอบของการผลิต เป็นที่รู้จักกันดีว่าตัวแบบการถดถอยปัวซง (Poisson: P regression model) มักถูกใช้เป็นตัวแบบพื้นฐานในการวิเคราะห์ข้อมูลจำนวนนับ อีกทั้งตัวแบบการถดถอยปัวซงนั้นมีคุณสมบัติ คือ ค่าเฉลี่ยเท่ากับค่าความแปรปรวน ซึ่งคุณสมบัติดังกล่าวนี้ถูกนำเสนอโดยนักคณิตศาสตร์ชาวฝรั่งเศส Siméon Denis Poisson ปี ค.ศ. 1781-1840 และมีนักวิจัยนำไปประยุกต์ใช้กับข้อมูลต่าง ๆ เช่น การเกิดปอดบวมของผู้ป่วยโรคหลอดเลือดสมองที่เข้ารับการรักษาในโรงพยาบาลน่าน วันที่ 1 มกราคม พ.ศ. 2556 ถึง 31 ธันวาคม พ.ศ. 2559 พบว่า ตัวแบบการถดถอย P มีความเหมาะสม (สุภัทรา โนนคล้อ และคณะ, 2561) เป็นต้น

อย่างไรก็ตามในทางปฏิบัติ ข้อมูลจำนวนนับไม่ได้มีความแปรปรวนเท่ากับค่าเฉลี่ยหรือเรียกว่า การกระจายเท่ากัน (Equi-dispersion) ซึ่งข้อมูลอาจเกิดการกระจายเกินเกณฑ์ (Over-dispersion) ในกรณีที่ความแปรปรวนมากกว่าค่าเฉลี่ย หรือการกระจายต่ำกว่าเกณฑ์ (Under-dispersion) ถ้าความแปรปรวนต่ำกว่าค่าเฉลี่ย มีนักสถิติเป็นจำนวนมากที่พยายามนำเสนอตัวแบบใหม่ ๆ เพื่อรองรับปัญหาการกระจายของข้อมูลดังกล่าว สำหรับการแก้ปัญหาการกระจายเกินเกณฑ์ Cameron และ Trivedi (1998) ได้เสนอตัวแบบการถดถอยทวินามเชิงลบ (Negative Binomial: NB regression model) ซึ่งเป็นการแจกแจงแบบผสมระหว่างการแจกแจงปัวซงและการแจกแจงแกมมาต่อมา Potts และ Elith (2006) ได้นำเสนอตัวแบบการถดถอยควอไซปัวซง (Quasi-Poisson: QP regression model) เป็นตัวแบบที่มีพารามิเตอร์การกระจายเกินเกณฑ์ซึ่งสามารถจัดการกับปัญหาค่าความแปรปรวนมากกว่าค่าเฉลี่ยได้ และสำหรับการแก้ปัญหาการกระจายเกินเกณฑ์และการกระจายต่ำกว่าเกณฑ์ ได้แก่ Conway และ Maxwell (1962) ได้มีการนำเสนอตัวแบบการถดถอยคอนเวย์แม็กซ์เวลล์ปัวซง (Conway Maxwell Poisson: CMP Regression model) ที่สามารถใช้ได้

กับปัญหาการกระจายทั้ง 2 รูปแบบ และ Annafari (2010) ได้นำเสนอตัวแบบการถดถอยปัวซงนัยทั่วไปแบบปรับใหม่ (Generalize Poisson: GP regression model) จึงเป็นอีกทางเลือกหนึ่งที่นำมาใช้ในการแก้ปัญหาดังกล่าว อีกทั้งมีนักวิจัยหลายท่านได้นำตัวแบบดังกล่าวไปประยุกต์ใช้กับข้อมูลจำนวนนับต่าง ๆ เพื่อหาความเหมาะสมของตัวแบบโดยใช้การเปรียบเทียบภาวะสารูปดี (Goodness of fit) สำหรับในกรณีการกระจายเกินเกณฑ์ เช่น ข้อมูลการเกิดอุบัติเหตุในประเทศไทย ปี พ.ศ. 2544 ถึง ปี พ.ศ. 2550 พบว่า ตัวแบบการถดถอย P เหมาะสมกับการคาดการณ์จำนวนอุบัติเหตุกับจำนวนผู้บาดเจ็บ ส่วนตัวแบบการถดถอย NB เหมาะสมกับการคาดการณ์จำนวนผู้เสียชีวิต (Thakali, 2008) ข้อมูลจำนวนสินค้าที่บกพร่องในโรงงานอุตสาหกรรมผลิตชิ้นส่วนรถยนต์แห่งหนึ่งในจังหวัดปทุมธานี ตั้งแต่เดือน มกราคม พ.ศ. 2555 ถึงเดือน กรกฎาคม พ.ศ. 2555 พบว่า ตัวแบบการถดถอย NB มีความเหมาะสมกว่า P (อดิเทพ ไชยวรรณ, วสันต์ บุญโฮ้ และ พิษณุ ทองขาว, 2555) ข้อมูลการเกิดอุบัติเหตุบนท้องถนนที่เป็นถนนทางหลวง โดยเก็บรวบรวมข้อมูลจากสำนักอำนวยความปลอดภัยกรมทางหลวง ตั้งแต่เดือนมกราคมถึงเดือนธันวาคม ปี พ.ศ. 2559 พบว่า ตัวแบบการถดถอย QP มีความเหมาะสมกว่า NB (นวพรรณ เชื้ออ่ำ, บุญอ้อม โฉมที และ อภิญญา หิรัญวงษ์, 2018) และข้อมูลจำนวนผู้ป่วยที่มีการเปลี่ยนแปลงจำนวนเซลล์ CD4 เริ่มต้นเนื่องจากการรักษาด้วยยาต้านไวรัสที่ให้กับผู้ใหญ่ที่ติดเชื้อเอชไอวีในประเทศเอธิโอเปียเหนือ - ตะวันตก (ภูมิภาค Amhara) พบว่า ตัวแบบการถดถอย QP มีความเหมาะสมกว่า NB (Seyoum, Ndlovu และ Zewotir, 2016) นอกจากนี้ (Lord, Guikema และ Geedipally, 2008) ได้เปรียบเทียบประสิทธิภาพการพยากรณ์โดยใช้ตัวแบบการถดถอย CMP และ NB โดยใช้กับข้อมูลการชนกันของรถยนต์ โดยข้อมูลชุดแรกบริเวณแยกสัญญาณสี่เลนที่ตั้งอยู่ในรัฐ Toronto Ontario ปี ค.ศ. 1995 และชุดที่สองข้อมูลย้อนหลัง 5 ปี บริเวณถนนสี่เลนในเขตชนบทของรัฐเท็กซัส (Texas) พบว่า ตัวแบบการถดถอย CMP และ NB มีประสิทธิภาพใกล้เคียงกัน เป็นต้น

เมื่อกล่าวถึงข้อมูลจำนวนนับจะสังเกตได้ว่าตัวแปรตอบสนองมักจะได้รับผลกระทบจากค่าศูนย์เป็นจำนวนมาก การมองข้ามปัญหาค่าศูนย์เฟ้อในการวิเคราะห์ข้อมูลโดยการใช้ตัวแบบการถดถอย P นั้น ส่งผลต่อการปฏิเสธสมมติฐานหลักที่ว่าค่าสัมประสิทธิ์การถดถอยเท่ากับศูนย์บ่อยกว่าความเป็นจริง เนื่องจากมีความคลาดเคลื่อนชนิดที่ 1 (Type I Error) สูงขึ้น ทั้งยังส่งผลให้ค่าเบี่ยงเบนมาตรฐานของตัวสถิติมีค่ามากกว่าหรือน้อยกว่าที่ควรจะเป็น (Payne และคณะ, 2017) จึงมีการนำเสนอตัวแบบการถดถอยใหม่สำหรับข้อมูลจำนวนนับที่มีค่าศูนย์เฟ้อ โดย Lambert (1992) ได้แก่ ตัวแบบการถดถอยปัวซงค่าศูนย์เฟ้อ (Zero-inflated Poisson: ZIP regression model) และ Ridout, Demetrio และ Hinde (1998) ได้ทำการขยายเป็นตัวแบบการถดถอยทวินามเชิงลบค่าศูนย์เฟ้อ (Zero-inflated Negative Binomial: ZINB regression model) เป็นอีกทางเลือกหนึ่งเพื่อจัดการกับปัญหาการกระจายเกินเกณฑ์และค่าศูนย์เฟ้อ นอกจากนี้ตัวแบบคอนเวย์แม็กซ์เวลล์ค่าศูนย์

เฟ้อ (Zero-inflated Conway Maxwell Poisson: ZICMP regression model) เป็นตัวแบบที่ถูกพัฒนาขึ้นเพื่อจัดการกับข้อมูลค่าศูนย์เฟ้อที่มีปัญหาการกระจายทั้ง 2 รูปแบบ การหาความเหมาะสมของตัวแบบโดยใช้การเปรียบเทียบภาวะสารรูปดี (Goodness of fit) ของตัวแบบสำหรับข้อมูลที่มีกระกระจายเกินเกณฑ์และค่าศูนย์เฟ้อ เช่น การจำลองข้อมูลด้วยวิธีมอนติคาร์โล โดยกำหนดขนาดตัวอย่าง 3 ระดับคือ 10 20 และ 50 และค่าเฉลี่ย 2 ระดับ คือ 5 และ 10 พบว่า ตัวแบบการถดถอย ZIP มีความเหมาะสมกว่า P (Xie, He และ Goh, 2001) ข้อมูลที่เป็นศูนย์และระดับความเบ้ของข้อมูลที่ไม่เป็นศูนย์หลายระดับและใช้วิธีการจำลองข้อมูลโดยเทคนิคมอนติคาร์โล พบว่า ตัวแบบการถดถอย ZINB มีความเหมาะสมกว่า ZIP และตัวแบบการถดถอยปัวซงนัยทั่วไปค่าศูนย์เฟ้อ (Zero-inflated Generalize Poisson: ZIGP regression model) (กษมะ นิจจันทร์พันศรี, 2554) นอกจากนี้ยังมีการใช้ข้อมูลจริงในการศึกษาได้แก่ ข้อมูลการยิงแสงไปยังรากของแอปเปิลภายใต้ระดับความเข้มข้นของไซโตยานินที่แตกต่างกัน พบว่า ตัวแบบการถดถอย ZINB มีความเหมาะสมกว่า P, NB และ ZIP (Ridout, Demétrio และ Hinde, 1998) Alqawba และ Diawara (2020) ใช้ข้อมูลจำนวนครั้งของการเกิดพายุทรายแรงตามรายเดือน บันทึกโดยสถานีสนามบิน AQI ในจังหวัดตะวันออกประเทศซาอุดิอาระเบีย ตั้งแต่เดือนมกราคม ปี ค.ศ. 1978 ถึงเดือนธันวาคม ปี ค.ศ. 2013 พบว่า ตัวแบบการถดถอย ZINB มีความเหมาะสมกว่า ZIP และ ZICMP ข้อมูลการชนของรถยนต์ในการขนส่งบนทางหลวง Mashhad Urban ประเทศอิหร่าน ปี ค.ศ. 2006 ถึง ปี ค.ศ. 2009 พบว่าตัวแบบการถดถอย ZINB มีประสิทธิภาพดีกว่า ZIP (Ayati และ Abbasi, 2014) และข้อมูลอุบัติเหตุทางถนนของเส้นทางเส้นทาง F0050 ประเทศซาอุดิอาระเบีย (Route F0050 Kluang-Air Hitam-Batu Pahat) ตั้งแต่กิโลเมตรที่ 0 ถึง กิโลเมตรที่ 58 ประเทศมาเลเซีย ปี ค.ศ. 2010 ถึง ปี ค.ศ. 2014 นักวิจัยมีวัตถุประสงค์ในการเปรียบเทียบประสิทธิภาพการพยากรณ์ของตัวแบบการถดถอย P, NB, ZIP และ ZINB พบว่า ตัวแบบการถดถอย ZINB และ NB มีประสิทธิภาพการพยากรณ์ดีกว่าตัวแบบ P และ ตัวแบบ ZIP (Prasetijo และคณะ, 2020) การจำลองข้อมูลแบบผสมระหว่างสัดส่วนของค่าศูนย์ (ร้อยละ 20 40 60 และ 80) และพารามิเตอร์การกระจายที่แตกต่างกัน (10 50 และ 100) จากการแจกแจง NB รวมถึงการใช้ข้อมูลจริงเกี่ยวกับการสำรวจด้านสุขภาพจาก Behavioral Risk Factor Surveillance System พบว่า ตัวแบบการถดถอย ZINB และตัวแบบการถดถอยทวินามเชิงลบค่าศูนย์ผันแปร (Zero-altered Negative Binomial: ZANB regression model) มีความเหมาะสมกว่าตัวแบบการถดถอย P, NB, ZIP และตัวแบบการถดถอยปัวซงค่าศูนย์ผันแปร (Zero-altered Poisson: ZAP regression model) (Yang และคณะ, 2017) ข้อมูลจำลองด้วยวิธีมอนติคาร์โลที่ขนาดตัวอย่างเท่ากับ 100 500 และ 1,000 โดยจำลองข้อมูลกรณี Over-dispersion และ Under-dispersion มีการทำซ้ำ 1,000 รอบ รวมถึงการใช้ข้อมูลจริงของจำนวนรากที่เกิดจาก

หน่อที่ขยายพันธุ์ขนาดเล็ก 270 ยอดของแอปเปิลที่ได้รับการปรับปรุงสายพันธุ์ พบว่า ตัวแบบการถดถอย CMP มีความเหมาะสมกว่า ZIP และ ZIGP (Sim และคณะ, 2018)

ในปัจจุบันมีการใช้การเรียนรู้ของเครื่อง (Machine leaning) เข้ามามีบทบาทในการพยากรณ์ข้อมูลจำนวนนับ เนื่องจากมีข้อจำกัดหรือเงื่อนไขในการใช้งานน้อยกว่าตัวแบบทางสถิติ เช่น Tin Kam ในปี ค.ศ. 1995 นำเสนอต้นไม้การจำแนกและต้นไม้การถดถอย (Classification and Regression tree) ต่อมาถูกพัฒนาต่อเป็นเทคนิคป่าสุ่ม (Random Forest: RF) โดย Leo Breiman ใน ปี ค.ศ. 2001 ตัวอย่างการประยุกต์ใช้เทคนิคป่าสุ่มเปรียบเทียบประสิทธิภาพในการพยากรณ์กับตัวแบบทางสถิติต่าง ๆ ได้แก่ ข้อมูลอุบัติเหตุบนทางหลวง National Freeway 1 ในไต้หวัน ปี ค.ศ. 2001 ถึง 2002 พบว่าตัวแบบการถดถอย NB และเทคนิค CART มีประสิทธิภาพใกล้เคียงกันและมีความเหมาะสมกว่าตัวแบบการถดถอย P (Chen และ Chang, 2005), งานวิจัยของ Ma และ Yuan (2018) ข้อมูลอุบัติเหตุบนท้องถนนแห่งหนึ่งในประเทศจีน พบว่าตัวแบบการถดถอย ZINB มีประสิทธิภาพดีที่สุด รองลงมาคือ NB, เทคนิค RF และ P (Ma และ Yuan, 2018) และข้อมูลจำนวนครั้งในการเกิดน้ำท่วมที่ได้รับแจ้งจากประชาชนในพื้นที่จากบันทึกของ Norfolk Virginia USA ตั้งแต่เดือนกันยายน ปี ค.ศ. 2010 ถึงตุลาคม ปี ค.ศ. 2016 พบว่า เทคนิค RF มีประสิทธิภาพมากกว่าตัวแบบการถดถอย P (Sadler และคณะ, 2018)

จากการศึกษาวรรณกรรมข้างต้น พบว่า งานวิจัยส่วนใหญ่จะเป็นการเปรียบเทียบภาวะสารูปดีในการวิเคราะห์ข้อมูลจริงหรือข้อมูลจำลอง โดยอาศัยการกำหนดค่าพารามิเตอร์ต่าง ๆ จากงานวิจัยที่เกี่ยวข้องก่อนหน้านี้ และ/หรือกำหนดเพิ่มเติม งานวิจัยฉบับนี้ มุ่งเน้นการเปรียบเทียบประสิทธิภาพการพยากรณ์ข้อมูลจำนวนนับที่มีปัญหาการกระจายและค่าศูนย์เฟ้อ ซึ่งมักเกิดขึ้นได้ในทางปฏิบัติ โดยใช้ตัวแบบทางสถิติและการเรียนรู้ของเครื่อง ได้แก่ ตัวแบบการถดถอย QP, NB, CMP, ZIP, ZINB และเทคนิค RF โดยใช้ข้อมูลจริงและข้อมูลจำลองที่สร้างจากค่าประมาณพารามิเตอร์ของข้อมูลจริง การคำนวณเกณฑ์วัดประสิทธิภาพการพยากรณ์ของตัวแบบใช้วิธีการเค-โฟลด์ตรวจสอบแบบไขว้ (K-Fold Cross validation) ร่วมกับการจำลองด้วยวิธีมอนติคาร์โล (Monte Carlo simulation)

1.2 วัตถุประสงค์ของการศึกษา

เพื่อเปรียบเทียบประสิทธิภาพในการพยากรณ์ของตัวแบบสำหรับข้อมูลจำนวนนับที่มีปัญหาการกระจายและค่าศูนย์เฟ้อ

กรณีการกระจายต่ำกว่าเกณฑ์ ได้แก่ ตัวแบบการถดถอย CMP และเทคนิค RF

กรณีการกระจายเกินเกณฑ์ ได้แก่ ตัวแบบการถดถอย QP, CMP, ZIP, ZINB และเทคนิค RF

1.3 ขอบเขตการศึกษา

1.3.1 ขอบเขตของตัวแบบและโปรแกรมที่ใช้

1. ตัวแบบที่ใช้ในการศึกษา ได้แก่ ตัวแบบการถดถอย QP, CMP, ZIP, ZINB และเทคนิค RF

2. การจำลองข้อมูลและการวิเคราะห์ข้อมูลโดยใช้โปรแกรม R เวอร์ชัน 4.0.0 และโปรแกรม Easy Fit เวอร์ชัน 5.5

1.3.2 ขอบเขตของข้อมูลจริง

1. ข้อมูลที่ใช้ในการศึกษานี้เป็นข้อมูลอุบัติเหตุทางถนนที่ส่งผลให้มีผู้เสียชีวิตทั่วประเทศไทย เดือนเมษายน ปี พ.ศ. 2558 มีจำนวนทั้งหมด 4,666 เหตุการณ์ และจำนวนอุบัติเหตุที่ไม่มีผู้เสียชีวิต 3,836 ครั้ง จำนวนอุบัติเหตุที่มีผู้เสียชีวิตหนึ่งคนและสองคน คือ 780 ครั้ง และ 50 ครั้ง ตัวแปรตอบสนอง คือ จำนวนผู้เสียชีวิตจากอุบัติเหตุทางถนน ตัวแปรเชิงคุณภาพ ได้แก่ ประเภทสายทาง, พื้นผิวถนน, ลักษณะสายทาง, สภาพอากาศ และ ช่วงเวลา ข้อมูลนี้ได้ถูกรวบรวมมาจากฐานข้อมูลสถิติอุบัติเหตุของกรมป้องกันและบรรเทาสาธารณภัยเท่านั้น ไม่รวมข้อมูลสถิติอุบัติเหตุทางถนนของประเทศไทย จากฐานข้อมูลอื่น ๆ เช่น สำนักงานตำรวจแห่งชาติ, ศูนย์ข้อมูลอุบัติเหตุ Thai RSC และกระทรวงสาธารณสุข เป็นต้น โดยตัวแปรที่ใช้ในการศึกษามีดังนี้

ตารางที่ 1.1 ตัวแปรของข้อมูลการเกิดอุบัติเหตุทางถนนทั่วประเทศไทย ปี พ.ศ. 2558 (n = 4,666)

ชื่อตัวแปร	**คำอธิบายตัวแปร**
Y: HUMAN_DEAD	จำนวนผู้ที่เสียชีวิตหรือเสียชีวิตในเวลาต่อมาจากอุบัติเหตุทางถนน (คน/จุดเกิดเหตุ) 0 ไม่มีผู้เสียชีวิต 1 มีผู้เสียชีวิตจำนวน 1 คน 2 มีผู้เสียชีวิตจำนวน 2 คน
X_1: ROADTYPE_ID	ประเภทสายทาง : 1 ถนนกรมทางหลวง 2 ถนนกรมทางหลวงชนบท 3 ถนนเทศบาล 4 ถนนใน อบต., หมู่บ้านและอื่น ๆ
X_2: ROADSKIN_ID	พื้นผิวถนน : 1 แห้ง 2 เปียก, หลุมบ่อและอื่น ๆ
X_3: ACDPOINT_ID	ลักษณะสายทาง : 1 ทางตรง 2 ทางโค้ง 3 ทางแยก 4 ทางคนข้าม, มีสิ่งกีดขวางและอื่น ๆ
X_4: ATMOSPHEE_ID	สภาพอากาศ : 1 แจ่มใส 2 มีหมอก, ฝนตก, ฝุ่นควันและอื่น ๆ
X_5: LIGHT_ID	ช่วงเวลา : 1 กลางวัน 2 กลางคืนมีแสงไฟฟ้า 3 กลางคืนไม่มีแสงไฟฟ้า 4 อื่น ๆ

2. ข้อมูลการเรียกร้องค่าสินไหมกรมธรรม์ประกันภัยรถยนต์จากบริษัทประกันภัยแห่งหนึ่งในประเทศไทย ปี พ.ศ. 2560 จำนวน 2,991 กรมธรรม์ (คน) ตัวแปรตอบสนอง คือ จำนวนครั้งของการเรียกร้องค่าสินไหมทดแทนของลูกค้า ตัวแปรเชิงคุณภาพ ได้แก่ เพศและการต่อกรมธรรม์ ตัวแปรเชิงปริมาณ ได้แก่ เบี้ยประกันภัยและส่วนลดประวัติดี โดยตัวแปรที่ใช้ในการศึกษามีดังนี้

ตารางที่ 1.2 ตัวแปรของข้อมูลเรียกร้องค้าสินไหมทดแทนของลูกค้า ปี พ.ศ. 2560 (n = 2,991)

ชื่อตัวแปร	**คำอธิบายตัวแปร**
Y: Claim	จำนวนครั้งของการเรียกร้องค่าสินไหมทดแทน (ครั้ง/คน)
X_1: Premium	เบี้ยประกันภัย (1 : 100 บาท)
X_2: CarYear	อายุรถยนต์ที่รับประกัน
X_3: Gender	เพศ
	0 ชาย
	1 หญิง
X_4: No Claim Bonus (NCB)	ส่วนลดประวัติดี (1 : 100 บาท)
X_5: Renew	การต่อกรมธรรม์
	0 ไม่ต่อกรมธรรม์
	1 ต่อกรมธรรม์

1.3.3 ขอบเขตของข้อมูลจำลอง

การกำหนดขอบเขตในการสร้างข้อมูลจำลองโดยใช้ค่าตั้งต้นมาจากข้อมูลจริง

1. ขนาดตัวอย่าง (n) คือ 250 500 1,000 3,000 และ 5,000

2. ตัวแปรอิสระ

2.1 ตัวแปรเชิงคุณภาพจะจำลองข้อมูลโดยใช้การแจกแจงพหุนาม (Multinomial distribution) โดยกำหนดความน่าจะเป็นที่จะเกิดลักษณะที่สนใจ (q_i) ในแต่ละกลุ่มจากสัดส่วนของกลุ่มนั้น ๆ ในข้อมูลจริงเป็นตังตั้งต้น ยกตัวอย่าง เช่น ตัวแปร X_1 : ROADTYPE_ID: ประเภทสายทาง: 1 ถนนกรมทางหลวง, 2 ถนนกรมทางหลวงชนบท, 3 ถนนเทศบาล และ 4 ถนนใน อบต., หมู่บ้านและอื่น ๆ พบว่าในข้อมูลจริง ถนนกรมทางหลวงมีทั้งหมด 1,864 กรณีจากข้อมูลทั้งหมด 4,666 เหตุการณ์ ดังนั้นจะกำหนดค่า q_1 สำหรับตัวแปร X_1 เท่ากับ q_1 = 1,864/4,666 = 0.3995 และ q_2 = 536/4,666 = 0.1149 ในการกำหนดค่า q_i ของตัวแปรอิสระอื่น ๆ จะทำในลักษณะเดียวกัน

2.2 ตัวแปรเชิงปริมาณจะจำลองข้อมูลโดยใช้การแจกแจงของข้อมูลจริง ซึ่งข้อมูลจริงจะมีการตรวจสอบการแจกแจงด้วยโปรแกรม Easy Fit เวอร์ชัน 5.5 โดยนำค่าพารามิเตอร์ของการแจกแจงที่ได้ไปใช้ในการจำลองตัวแปรอิสระ

3. ตัวแปรตอบสนอง

3.1 ทำการวิเคราะห์ข้อมูลจริงโดยใช้ตัวแบบการถดถอย ZICMP เพื่อประมาณค่าสัมประสิทธิ์การถดถอยของค่าเฉลี่ย ($\boldsymbol{\beta}$) โดย $\beta_0, \beta_1, ..., \beta_p$ สัมประสิทธิ์การถดถอยของการกระจาย ($\boldsymbol{\gamma}$) โดย $\gamma_0, \gamma_1, ..., \gamma_p$ และสัมประสิทธิ์การถดถอยความน่าจะเป็นที่จะเกิดศูนย์ ($\boldsymbol{\varphi}$) โดย $\varphi_0, \varphi_1, ..., \varphi_p$

3.1.1 ข้อมูลจำลองชุดที่ 1 จะใช้ค่าสัมประสิทธิ์การถดถอยจากข้อมูลจริง ในที่นี้ $\delta_1(\boldsymbol{\beta}, \boldsymbol{\gamma}, \boldsymbol{\varphi})$ คือ เวกเตอร์สัมประสิทธิ์การถดถอยจากหัวข้อ 3.1

3.1.2 ข้อมูลจำลองชุดที่ 2 จะใช้ค่าสัมประสิทธิ์การถดถอยจากข้อมูลจริงที่เพิ่มขึ้น 50 เปอร์เซ็นต์ ในที่นี้ $\delta_2(\boldsymbol{\beta}, \boldsymbol{\gamma}, \boldsymbol{\varphi}) = 1.5 \times \delta_1(\boldsymbol{\beta}, \boldsymbol{\gamma}, \boldsymbol{\varphi})$ ซึ่งเป็นเวกเตอร์สัมประสิทธิ์การถดถอยของข้อมูลจำลองชุดที่ 2

3.2 ประมาณค่าเฉลี่ย ($\boldsymbol{\lambda}$), ค่าการกระจาย ($\boldsymbol{\nu}$) และความน่าจะเป็นที่จะเกิดศูนย์ ($\boldsymbol{\pi}$) โดยใช้เวกเตอร์สัมประสิทธิ์การถดถอยที่ได้จากข้อมูลจำลองชุดที่ 1 และข้อมูลจำลองชุดที่ 2

3.3 สร้างตัวแปรตอบสนองที่มีการแจกแจง ZICMP สำหรับข้อมูลจำลองแต่ละชุด นั่นคือ $\mathbf{Y} \sim ZICMP(\boldsymbol{\lambda}, \boldsymbol{\nu}, \boldsymbol{\pi})$

4. งานวิจัยนี้กำหนดระดับนัยสำคัญทางสถิติ เท่ากับ 0.05

5. วิธีการเปรียบเทียบประสิทธิภาพการพยากรณ์ของตัวแบบจะใช้วิธีเค-โฟลด์ตรวจสอบแบบไขว้ (K-Fold Cross Validation) ที่ K=5 โดยแบ่งข้อมูลออกเป็น 5 ส่วนเท่า ๆ กัน ซึ่งข้อมูล 1 ส่วนจะเป็นชุดทดสอบสำหรับพยากรณ์และข้อมูล 4 ส่วนที่เหลือจะเป็นชุดฝึกสอนสำหรับสร้างตัวแบบ

1.4 เกณฑ์ที่ใช้ในการเปรียบเทียบ

1.4.1 ค่าเฉลี่ยของรากของค่าคลาดเคลื่อนกำลังสองเฉลี่ย (Mean of Root Mean Square Error: mRMSE)

เนื่องจากใช้วิธีการเปรียบเทียบประสิทธิภาพการพยากรณ์ของตัวแบบด้วยวิธีเค-โฟลด์ตรวจสอบไขว้ จึงต้องคำนวณค่ารากที่สองของค่าคลาดเคลื่อนกำลังสองเฉลี่ย (RMSE) ในแต่ละโฟลด์ มีสูตรการคำนวณดังสมการ (1.1) โดยที่ค่าคาดเคลื่อน คือ ความแตกต่างของค่าจริง

ของตัวแปรตอบสนองกับค่าพยากรณ์ หลังจากนั้นจะการคำนวณค่า mRMSE ดังสมการ (1.2) และ (1.3) ซึ่งตัวแบบที่ให้ค่า mRMSE ต่ำสุดจะเป็นตัวแบบที่มีประสิทธิภาพการพยากรณ์ดีที่สุด

$$\mathrm{RMSE}_l = \sqrt{\frac{\sum_{i=1}^{n_l}\left(y_i - \hat{y}_i\right)^2}{n_l}} \tag{1.1}$$

โดยที่ $RMSE_l$ คือ ค่าคลาดเคลื่อนกำลังสองเฉลี่ยในโฟลด์ ที่ l

y_i คือ ค่าจริงของตัวแปรตอบสนอง โดยที่ $i = 1, 2, \ldots, n_l$

$\hat{y}_i$ คือ ค่าพยากรณ์ โดยที่ $i = 1, 2, \ldots, n_l$

n_l คือ ขนาดตัวอย่างในแต่ละโฟลด์ โดยที่ $l = 1, 2, \ldots, K$

การคำนวน mRMSE ในกรณีข้อมูลจริง ดังสูตรต่อไปนี้

$$\mathrm{mRMSE} = \frac{\sum_{l=1}^{K} RMSE_l}{K} \tag{1.2}$$

โดยที่ K คือ จำนวนโฟลด์ที่ l ในงานวิจัยนี้กำหนด $K = 5$

การคำนวณ mRMSE ในกรณีข้อมูลจำลอง ดังสูตรต่อไปนี้

$$\mathrm{mRMSE} = \frac{\sum_{h=1}^{M}\sum_{l=1}^{K} RMSE_{hl}}{M \times K} \tag{1.3}$$

โดยที่ M คือ จำนวนรอบในการทำซ้ำ ในงานวิจัยนี้กำหนด $M = 500$

$RMSE_{hl}$ คือ ค่า RMSE ในรอบที่ h และโฟลด์ที่ l

1.4.2 ค่าเฉลี่ยของค่าเบี่ยงเบนมาตรฐาน (Mean of Standard Deviation: mS.D.)

เนื่องจากใช้วิธีการเปรียบเทียบประสิทธิภาพการพยากรณ์ของตัวแบบด้วยวิธีเค-โฟลด์ตรวจสอบไขว้จึงต้องคำนวณค่าเบี่ยงเบนมาตรฐาน (S.D.) ของค่า mRMSE ในแต่ละโฟลด์ มีสูตรการคำนวณดังสมการที่ (1.4) โดยที่ค่าเบี่ยงเบนมาตรฐาน คือ ค่าการกระจายของตัวแปรตอบสนอง หลังจากนั้นจะทำการคำนวณค่า mS.D. ดังสมการที่ (1.5) และ (1.6) ซึ่งตัวแบบที่ให้ค่า mS.D. ต่ำกว่าจะมีประสิทธิภาพที่แม่นยำกว่า

$$S.D._{\cdot l} = \sqrt{\frac{\sum_{l=1}^{n_l}(y_i - \bar{y})^2}{n_l - 1}} \tag{1.4}$$

โดยที่ $S.D._{\cdot l}$ คือ ค่าเบี่ยงเบนมาตรฐาน ในโฟลด์ที่ l โดยที่ $l = 1, 2, ..., K$

$\bar{y}$ คือ ค่าเฉลี่ยของกลุ่มตัวอย่าง

n_l คือ ขนาดตัวอย่างในแต่ละโฟลด์

การคำนวณ mS.D. ในกรณีข้อมูลจริง ดังสูตรต่อไปนี้

$$\text{mS.D.} = \frac{\sum_{l=1}^{K} S.D._{\cdot l}}{K} \tag{1.5}$$

โดยที่ K คือ จำนวนโฟลด์ในงานวิจัยนี้กำหนด $K = 5$

การคำนวณ mS.D. ในกรณีข้อมูลจำลอง ดังสูตรต่อไปนี้

$$\text{mS.D.} = \frac{\sum_{h=1}^{M}\sum_{l=1}^{K} S.D._{\cdot hl}}{M \times K} \tag{1.6}$$

โดยที่ M คือ จำนวนรอบในการทำซ้ำ ในงานวิจัยนี้กำหนด $M = 500$

$S.D._{\cdot hl}$ คือ ค่า S.D. ในรอบที่ h และโฟลด์ที่ l

ในงานวิจัยนี้ใช้เกณฑ์การเปรียบเทียบประสิทธิภาพการพยากรณ์ของตัวแบบจะเลือกพิจารณาจากค่า mRMSE เป็นหลักและค่า mS.D. ใช้เพื่อประกอบการตัดสินใจเพียงเท่านั้น

1.5 ประโยชน์ที่คาดว่าจะได้รับ

เพื่อเป็นแนวทางในการเลือกใช้ตัวแบบในการพยากรณ์ข้อมูลจำนวนนับที่มีผลกระทบจากค่าศูนย์เฟ้อและปัญหาการกระจายต่ำกว่าเกณฑ์หรือเกินเกณฑ์ โดยใช้ขนาดตัวอย่างต่าง ๆ ได้อย่างเหมาะสม

บทที่ 2
วรรณกรรมและงานวิจัยที่เกี่ยวข้อง

การวิจัยครั้งนี้เป็นการศึกษาเปรียบเทียบประสิทธิภาพในการพยากรณ์ของตัวแบบสำหรับข้อมูลจำนวนนับที่มีการกระจายต่ำกว่าเกณฑ์และการกระจายเกินเกณฑ์ โดยส่วนนี้จะกล่าวถึงทฤษฎีและงานวิจัยต่าง ๆ ที่เกี่ยวข้อง ดังนี้

2.1 ตัวแบบ/เทคนิคที่เกี่ยวข้อง

2.1.1 ตัวแบบการถดถอยปัวซง (Poisson: P regression model)

การแจกแจงความน่าจะเป็น P เป็นการแจกแจงที่อธิบายถึงจำนวนครั้งของเหตุการณ์หรือจำนวนสิ่งที่สนใจที่เกิดขึ้นในช่วงเวลาหรือขอบเขตที่กำหนด ให้ Y แทนตัวแปรตอบสนองของค่าสังเกตที่ i โดยเป็นจำนวนเต็มบวก (Non negative integer) สามารถเขียนฟังก์ชันมวลความน่าจะเป็น (Probability mass function) ของ Y ได้ดังนี้

$$P\left(Y=y_i\middle|\lambda_i\right)=\frac{e^{-(\lambda_i)}\lambda_i^{y_i}}{y_i!} \quad ;y_i=0,1,2,\ldots;i=1,2,\ldots,n \tag{2.1}$$

โดย λ_i เป็นค่าพารามิเตอร์สำหรับค่าสังเกตที่ i ซึ่งหมายถึงค่าเฉลี่ยหรือค่าคาดหวัง (Expected number) ของจำนวนครั้งในเหตุการณ์ที่ i

ฟังก์ชันที่ใช้แสดงความสัมพันธ์ระหว่าง Y และ X คือ ฟังก์ชันเชื่อมโยงล็อก (Log link function) ดังสมการต่อไปนี้

$$\ln(\lambda_i)=\beta_0+\sum_{j=1}^{p}\beta_j x_{ij} \tag{2.2}$$

จะได้ว่า

$$\lambda_i=\exp\left(\beta_0+\sum_{j=1}^{p}\beta_j x_{ij}\right)=\left(e^{\beta_0}\right)\left(e^{\beta_1}\right)^{x_{i1}}\ldots\left(e^{\beta_p}\right)^{x_{ip}} \tag{2.3}$$

โดยที่ x_{ij} คือ ตัวแปรอิสระที่ j ($j=1,2,...,p$) จากค่าสังเกตที่ i

β_0 คือ ค่าคงที่สัมประสิทธิ์การถดถอย สำหรับการประมาณเฉลี่ย

β_j คือ สัมประสิทธิ์การถดถอย สำหรับการประมาณค่าเฉลี่ย

แต่ในทางปฏิบัติข้อมูลจำนวนนับมักมีความแปรปรวนสูงกว่าหรือต่ำกว่าค่าเฉลี่ย หรือกล่าวอีกนัยหนึ่ง คือ ข้อมูลมีการกระจายเกินเกณฑ์หรือข้อมูลมีการกระจายต่ำกว่าเกณฑ์ ทำให้ตัวแบบการถดถอยปัวซงไม่เหมาะสมในการวิเคราะห์ข้อมูลดังกล่าว

2.1.2 ตัวแบบการถดถอยควอไซปัวซง

(Quasi Poisson: QP regression model)

ตัวแบบการถดถอย QP เป็นตัวแบบที่มีฟังก์ชันมวลน่าจะเป็นเหมือนกับการแจกแจงปัวซง แต่ความแปรปรวนแตกต่างกัน โดยที่กำหนดให้ Y เป็นตัวแปรตอบสนองของค่าสังเกตที่ i โดยที่ $i=1,2,...,n$ และเป็นตัวแปรสุ่มที่มีการแจกแจงปัวซงที่เป็นอิสระต่อกัน มีพารามิเตอร์ 2 ตัว คือ พารามิเตอร์ค่าเฉลี่ย λ โดยที่ $\lambda_i > 0$ และพารามิเตอร์การกระจายเกินเกณฑ์ θ โดยที่ $\theta_i > 1$ และความแปรปรวนเท่ากับ $Var(Y_i) = f(\lambda_i) = \theta\lambda_i$ โดยที่ $f(\lambda_i)$ คือ ฟังก์ชันของความแปรปรวนที่ถูกกำหนดโดยค่าเฉลี่ย การประมาณค่าพารามิเตอร์ใช้วิธีภาวะน่าจะเป็นควอไซ (Quasi-likelihood function) หรือเรียกว่า Full Quasi – likelihood ซึ่งถูกพัฒนาโดย Wedderburn (1974) เป็นทางเลือกหนึ่งของฟังก์ชันภาวะน่าจะเป็นในกรณีที่ไม่ทราบการแจกแจงความน่าเป็นที่แท้จริง ซึ่งจะประมาณค่าสัมประสิทธิ์การถดถอยและส่วนประกอบของความแปรปรวน (Variance Components) พร้อมกัน โดยเจาะจงเพียงรูปแบบของ Linear parameter $(\boldsymbol{\beta})$ หรือเป็นเวกเตอร์ของสัมประสิทธิ์การถดถอย และในการประมาณ $f(\lambda_i)$ จะใช้ฟังก์ชันของค่าเฉลี่ยและความแปรปรวนของฟังก์ชันของค่าเฉลี่ย (ศิรินทิพย์, 2553 และนพวรรณ, 2561)

Quasi-score กำหนดได้ดังนี้

$$\frac{\partial Q(Y_i;\lambda_i)}{\partial \lambda_i} = \frac{Y_i - \lambda_i}{f(\lambda_i)} \tag{2.4}$$

หรือหาได้จากสมการ

$$Q(Y_i;\lambda_i) = \int_{y_i}^{\lambda_i} \frac{y-t}{f(t)} dt \tag{2.5}$$

โดย $Q(Y_i;\lambda_i)$ คือ ฟังก์ชันภาวะน่าจะเป็นควอไซ (Quasi-likelihood function) ฟังก์ชันเชื่อมโยงล็อกของ QP เหมือนกับการแจกแจงปัวซง ซึ่งแสดงไว้ดังสมการที่ (2.2) และ (2.3)

θ คือ พารามิเตอร์การกระจายเกินเกณฑ์ โดยที่ $\theta>1$ ถ้าหาก $\theta=1$ จะเป็นการแจกแจงปัวซงและ Quasi – likelihood สำหรับการประมาณพารามิเตอร์ $(\boldsymbol{\beta})$ สามารถหาได้จาก $Q(Y_i;\lambda_i)$ อาจเขียนอยู่ในรูปของ $Q(\boldsymbol{\beta})=\mathbf{0}$ ซึ่งจะได้

$$Q_j(\boldsymbol{\beta})=\sum_{i=1}^{n}\frac{(y_i-\lambda_i)x_{ij}}{f(\lambda_i)}\left(\frac{\partial\lambda_i}{\partial\eta_i}\right)=0,\quad j=1,2,...,p \tag{2.6}$$

และ

$$\theta=\frac{1}{n-p}\sum_{i}^{n}\frac{y_i-\lambda_i}{\sqrt{f(\lambda_i)}} \tag{2.7}$$

2.1.3 ตัวแบบการถดถอยปัวซงค่าศูนย์เฟ้อ

(Zero-inflated Poisson: ZIP regression model)

ตัวแบบการถดถอย ZIP ถูกนำเสนอโดย Lambert (1992) เป็นส่วนขยายของตัวแบบการถดถอย P ในกรณีที่ตัวแปรตอบสนองมีแนวโน้มที่จะเกิดค่าศูนย์เฟ้อ ซึ่งเรียกว่า Problem of Excess Zero เหตุการณ์ดังกล่าวจึงนำไปสู่การพัฒนาเป็นตัวแบบผสม (Mixture model) ที่ข้อมูลเกิดจาก 2 แหล่ง คือ การเกิดค่าศูนย์ที่แท้จริง (Certain zero) และการเกิดค่าที่เป็นจำนวนนับ ซึ่งจะมีค่าศูนย์หรือไม่มีค่าศูนย์ก็ได้ สามารถอธิบายได้ดังนี้

การแจกแจง ZIP สร้างมาจากพารามิเตอร์ 2 ตัว คือ พารามิเตอร์ค่าเฉลี่ย λ และพารามิเตอร์แสดงความน่าจะเป็นที่จะเกิดศูนย์ π โดยที่ $0\le\pi\le1$ ตัวแปรสุ่ม Y ที่มีการแจกแจง ZIP แทนด้วย $Y\sim ZIP(\lambda,\pi)$ ฟังก์ชันมวลความน่าจะเป็นของ Y เป็นดังนี้

$$P(Y_i=y_i\mid\lambda_i,\pi_i)=\begin{cases}\pi_i+(1-\pi_i)e^{-\lambda_i} & ;y_i=0,\\ (1-\pi_i)\left(\dfrac{e^{-\lambda_i}\lambda_i^{y_i}}{y_i!}\right) & ;y_i>0\end{cases} \tag{2.8}$$

ฟังก์ชันเชื่อมโยงล็อกของ ZIP สำหรับพารามิเตอร์ λ_i เหมือนกับการแจกแจงปัวซง ซึ่งแสดงไว้ดังสมการที่ (2.2) และ (2.3) และฟังก์ชันเชื่อมโยงสำหรับพารามิเตอร์ π_i ใช้ฟังก์ชันเชื่อมโยงลอจิต (Logit link function) ดังสมการต่อไปนี้

$$\text{logit}(\pi_i) = \ln\left(\frac{\pi_i}{1-\pi_i}\right) = \varphi_0 + \sum_{j=1}^{p} \varphi_j x_{ij} \tag{2.9}$$

$$\pi_i = \frac{\exp\left(\varphi_0 + \sum_{j=1}^{p} \varphi_j x_{ij}\right)}{1+\exp\left(\varphi_0 + \sum_{j=1}^{p} \varphi_j x_{ij}\right)} \tag{2.10}$$

โดยที่ φ_0 คือ ค่าคงที่สัมประสิทธิ์การถดถอยลอจิต สำหรับการประมาณค่าความน่าจะเป็นที่จะเกิดศูนย์

φ_j คือ สัมประสิทธิ์การถดถอยลอจิต สำหรับการประมาณค่าความน่าจะเป็นที่จะเกิดศูนย์

x_{ij} คือ ตัวแปรอิสระที่ j ($j = 1,2,...,p$) จากค่าสังเกตที่ i

การหาค่าคาดหมายและความแปรปรวนของตัวแปรสุ่ม $Y \sim ZIP(\lambda,\pi)$ อธิบายได้ดังนี้ กำหนดให้ตัวแปร z_i เป็นตัวแปรแฝง (Latent variable) ที่ไม่สามารถสังเกตค่าได้ ซึ่งมีค่าเป็น 0 หรือ 1 เมื่อ $z_i = 0$ จะได้ว่า $Y_i = 0$ เมื่อ $z_i = 1$จะได้ว่า Y_i เป็นตัวแปรที่มีการแจกแจงปัวซง ($Y_i \sim Poisson(\lambda_i)$) จะได้ว่า

$$E(Y_i \mid z_i) = \begin{cases} \pi_i & ;Y_i=0 \\ 1\text{-}\pi_i & ;Y_i:\text{Poisson}(\lambda_i) \end{cases} \tag{2.11}$$

ซึ่งมีค่าเฉลี่ยและความแปรปรวนจากการแจกแจงผสม (Mixed distribution) ของ Y_i คือ

$$\begin{aligned} E(Y_i) &= E[E(Y_i \mid z_i)] \\ &= \pi_i(0) + (1\text{-}\pi_i)\lambda_i \\ &= (1\text{-}\pi_i)\lambda_i \end{aligned} \tag{2.12}$$

เนื่องจาก

$$E[Var(Y_i \mid z_i)] = \pi_i(0-(1-\pi_i)\lambda_i)^2 + (1-\pi_i)(\lambda_i-(1-\pi_i)\lambda_i)^2$$

$$= \pi_i (1-\pi_i)^2 \lambda_i^2 + (1-\pi_i)\left(\lambda_i \left(1-(1-\pi_i)\right)\right)^2$$
$$= \pi_i (1-\pi_i)^2 \lambda_i^2 + (1-\pi_i)(\lambda_i \pi_i)^2$$
$$= \pi_i (1-\pi_i)\lambda_i^2 + \left((1-\pi_i)+\pi_i\right)$$
$$= \pi_i (1-\pi_i)\lambda_i^2$$

จะได้ว่า

$$Var(Y_i) = E[Var(Y_i \mid z_i)] + Var[E(Y_i \mid z_i)]$$

$$Var(Y_i) = (1-\pi_i)\lambda_i(1+\pi_i\lambda_i) \quad (2.13)$$

สามารถเขียนฟังก์ชันล็อกภาวะน่าจะเป็น (Log-likelihood function) ของการแจกแจง ZIP ได้ดังนี้

$$\log L(\lambda_i, \pi_i; y_i) = \sum_{y_i=0} \log\left\{\pi_i + (1-\pi_i)e^{-\lambda_i}\right\} + \sum_{y_i>0} \log\left\{(1-\pi_i)\left(\frac{e^{-\lambda_i}\lambda_i^{y_i}}{y_i!}\right)\right\} \quad (2.14)$$

2.1.4 ตัวแบบการถดถอยทวินามเชิงลบค่าศูนย์เฟ้อ
(Zero-inflated Negative Binomial: ZINB regression model)

การแจกแจง ZINB ถูกพัฒนาโดย Ridout, Demetrio และ Hinde (1998) สร้างมาจากพารามิเตอร์ 3 ตัว คือ พารามิเตอร์ค่าเฉลี่ย λ โดยที่ $\lambda > 0$ พารามิเตอร์แสดงความน่าจะเป็นที่จะเกิดศูนย์ π โดยที่ $0 \le \pi \le 1$ และพารามิเตอร์การกระจาย α โดยที่ $\alpha > 0$ แสดงถึงข้อมูลมีการกระจายเกินเกณฑ์ นอกจากนี้ยังพบว่า $\alpha \to 0$ จะทำให้กลายเป็นการแจกแจง ZIP ตัวแปรสุ่ม Y ที่มีการแจกแจง ZINB แทนด้วยแทนด้วย $Y \sim ZINB(\lambda, \pi, \alpha)$ มีฟังก์ชันมวลความน่าจะเป็นของ Y เป็นดังนี้

$$\Pr(Y_i = y_i \mid \lambda_i, \pi_i, \alpha) = \begin{cases} \pi_i + (1-\pi_i)(1+\alpha\lambda_i)^{-\alpha^{-1}} & ;y_i = 0 \\ (1-\pi_i)\dfrac{\Gamma(y_i+\alpha^{-1})}{y_i!\Gamma(\alpha^{-1})}\left(\dfrac{\alpha^{-1}}{\alpha^{-1}+\lambda_i}\right)^{\alpha^{-1}}\left(\dfrac{\lambda_i}{\alpha^{-1}+\lambda_i}\right)^{y_i} & ;y_i > 0 \end{cases} \quad (2.15)$$

โดยที่ Γ คือ แกมมาฟังก์ชัน (Gamma function) และ $\lambda > 0,\ \alpha > 0,\ 0 \le \pi \le 1$

งานวิจัยนี้จะใช้ตัวแบบการถดถอยทวินามเชิงลบที่มีความแปรปรวนเป็นฟังก์ชันพหุนามกำลังสองของค่าเฉลี่ย (Quadratic mean-variance negative binomial model: NB2)

ฟังก์ชันเชื่อมโยงล็อกของ ZINB สำหรับพารามิเตอร์ λ_i เหมือนกับการแจกแจง P ซึ่งแสดงไว้ดังสมการที่ (2.2) และ (2.3) และฟังก์ชันเชื่อมโยงสำหรับพารามิเตอร์ π_i ใช้ฟังก์ชันเชื่อมโยงลอจิต (Logit link function) เหมือนกับการแจกแจง ZIP ที่แสดงไว้ดังสมการ (2.9) และ (2.10) สำหรับพารามิเตอร์การกระจาย α สามารถประมาณได้ดังนี้

$$\alpha = \sum_{i=1}^{n} \left\{ \frac{(y_i - \lambda_i)^2 - \lambda_i}{\lambda_i^2} \right\} \tag{2.16}$$

ค่าเฉลี่ยและความแปรปรวนของ Y_i คือ

$$E(Y_i) = (1-\pi_i)\lambda_i \tag{2.17}$$

$$Var(Y_i) = (1-\pi_i)\lambda_i(1+\alpha\lambda_i + \pi_i\lambda_i) \tag{2.18}$$

สามารถเขียนฟังก์ชันล็อกภาวะน่าจะเป็นของการแจกแจง ZINB ได้ดังนี้

$$\log L(\lambda_i, \alpha_i, \pi_i; y_i) = \sum_{y_i=0} \log \left\{ \pi_i + (1+\pi_i) \left(\frac{\alpha^{-1}{}_i}{\alpha_i^{-1} + \lambda_i} \right)^{\alpha_i} \right\}$$

$$+ \sum_{y_i>0} \log \left\{ (1-\pi_i) \frac{\Gamma(y_i + \alpha_i^{-1})}{\Gamma(y_i+1)\Gamma(\alpha_i^{-1})} \left(\frac{\alpha_i^{-1}}{\alpha^{-1}{}_i + \lambda_i} \right)^{\alpha^{-1}} \left(1 - \frac{\lambda_i}{\alpha_i^{-1} + \lambda_i} \right)^{y_i} \right\} \tag{2.19}$$

2.1.5 ตัวแบบการถดถอยคอนเวย์แม็กซ์เวลล์ปัวซง (Conway Maxwell Poisson: CMP regression model)

การแจกแจง CMP ถูกนำเสนอโดย Conway และ Maxwell (1962) เป็นการแจกแจงนัยทั่วไปและครอบคลุมข้อมูลที่มีการกระจายต่ำกว่าเกณฑ์และการกระจายเกินเกณฑ์ โดยกำหนดให้ตัวแปรสุ่ม Y เป็นตัวแปรสุ่มที่มีการแจกแจง CMP และเป็นอิสระต่อกัน มีพารามิเตอร์ 2 ตัว คือ พารามิเตอร์ค่าเฉลี่ย λ โดยที่ $\lambda > 0$ ซึ่งเป็นค่าคาดหวังของตัวแปรตอบสนองและพารามิเตอร์การกระจาย ν โดยที่ $\nu > 0$ ตัวแปรสุ่ม Y ที่มีการแจกแจง CMP แทนด้วย $Y \sim CMP(\lambda, \nu)$ มีฟังก์ชันมวลความน่าจะเป็นของ Y ดังนี้

$$P(Y_i = y_i \mid \lambda_i, \nu_i) = \frac{1}{Z(\lambda_i, \nu_i)} \frac{\lambda_i^{y_i}}{(y_i!)^{\nu_i}} \; ; y_i = 0,1,2,... \tag{2.20}$$

$$Z(\lambda_i, \nu_i) = \sum_{n=0}^{\infty} \frac{\lambda_i^{n}}{(n!)^{\nu_i}} \tag{2.21}$$

โดยที่ $\lambda > 0$, $\nu > 0$ และ $Z(\lambda, \nu)$ (Normalizing constant) เป็นค่าคงที่ซึ่งได้จากการประมาณ ดังสมการ (2.21) ฟังก์ชันเชื่อมโยงล็อกของ CMP สำหรับพารามิเตอร์ λ_i เหมือนกับการแจกแจง P ซึ่งแสดงไว้ดังสมการที่ (2.2) และ (2.3) และฟังก์ชันเชื่อมโยงล็อกสำหรับพารามิเตอร์ ν_i แสดงไว้ดังสมการที่ (2.22) และ (2.23)

$$\ln(\nu_i) = \gamma_0 + \sum_{j=1}^{p} \gamma_j x_{ij} \tag{2.22}$$

$$\nu_i = \exp\left(\gamma_0 + \sum_{j=1}^{p} \gamma_j x_{ij} \right) = \left(e^{\gamma_0}\right)\left(e^{\gamma_1}\right)^{x_{i1}} ... \left(e^{\gamma_p}\right)^{x_{ip}} \tag{2.23}$$

โดยที่ γ_0 คือ ค่าคงที่สัมประสิทธิ์การถดถอย สำหรับการประมาณค่าการกระจาย

γ_j คือ สัมประสิทธิ์การถดถอย สำหรับการประมาณค่าการกระจาย

กรณีพิเศษของการแจกแจง CMP ได้แก่ ในกรณีที่ $\nu = 1$ การแจกแจง CMP จะลดรูปเป็นการแจกแจง P ในกรณี $\nu \to \infty$ ทำให้ $Z(\lambda, \nu)$ ลู่เข้าใกล้ $1+\lambda$ ซึ่งสอดคล้องกับการแจกแจงแบร์นูลลี (Bernoulli distribution) ที่มีพารามิเตอร์ $\lambda/(1+\lambda)$ ในกรณีที่ $\nu = 0$ และ $\lambda < 1$ จะทำให้ $Z(\lambda, \nu) = 1/(1-\lambda)$ จะสอดคล้องกับการแจกแจงเรขาคณิต (Geometric distribution) ที่มีพารามิเตอร์ $(1-\lambda)$

ในงานวิจัยนี้ได้ทำการคำนวณ ν ตามวิธีการของ Hauer (2001) และ Miaou และ Lord (2003) ที่ทำให้สามารถประมาณได้โดยตรงด้วย $\frac{E(Y)}{V(Y)} \approx \nu$ ซึ่งเป็นอัตราส่วนของค่าเฉลี่ยและความแปรปรวน จะได้ว่ากรณีที่ $\nu < 1$ หมายถึง การกระจายเกินเกณฑ์ และกรณีที่ $\nu > 1$ หมายถึง การกระจายต่ำกว่าเกณฑ์

การแจกแจง CMP มีค่าเฉลี่ยและความแปรปรวนดังสมการต่อไปนี้

$$E(Y_i) \approx \lambda_i^{1/\nu_i} - \frac{1-\nu_i}{2\nu_i} \tag{2.24}$$

$$Var(Y_i) \approx \frac{\lambda^{1/\nu_i}}{\nu_i} \tag{2.25}$$

ต่อมา Guikema และ Cofelt ปี ค.ศ. 2008 ได้นำเสนอพารามิเตอร์รูปแบบใหม่ของการแจกแจง CMP โดยการแทนที่ $\lambda^{1/\nu}$ ด้วยพารามิเตอร์ λ_{CMP} โดยที่ λ_{CMP} เป็นฟังก์ชันของ λ และ ν เรียกว่า Mean Parametrized CMP โดยรูปแบบใหม่ของฟังก์ชันมวลความน่าจะเป็น (PMF) เป็นดังนี้

$$P(Y_i = y_i \mid \lambda_{CMP_i}, \nu_i) = \frac{1}{S(\lambda_{CMP_i}, \nu_i)} \left(\frac{{\lambda_{CMP_i}}^{y_i}}{y_i!} \right)^{\nu_i} \quad ; y = 0,1,2,... \tag{2.26}$$

$$S(\lambda_i, \nu_i) = \sum_{n=0}^{\infty} \left(\frac{{\lambda_{CMP_i}}^{n}}{n!} \right)^{\nu_i} \tag{2.27}$$

เนื่องจากการแจกแจง CMP ไม่ได้อยู่ในตระกูลเลขชี้กำลัง (Exponential family) จึงจำเป็นต้องปรับพารามิเตอร์ใหม่ โดยกำหนดให้ $\lambda_{CMP} = \lambda^{1/\nu}$ เพื่อสะดวกต่อการประมาณและง่ายต่อการอธิบายตัวแบบการถดถอยสำหรับค่าเฉลี่ยมากขึ้น (Huang, 2017) ซึ่งในงานวิจัยนี้จะใช้สมการฟังก์ชันเชื่อมโยงล็อกของ Mean Parametrized CMP สำหรับพารามิเตอร์ λ_i ดังแสดงไว้ดังสมการที่ (2.28) และฟังก์ชันเชื่อมโยงล็อกสำหรับพารามิเตอร์ ν_i เหมือนกับการแจกแจง CMP ที่แสดงไว้ดังสมการ (2.22) และ(2.23)

$$\lambda_{CMP_i} = \exp\left(\beta_0 + \sum_{j=1}^{p} \beta_j x_{ij} \right) = \left(e^{\beta_0} \right) \left(e^{\beta_1} \right)^{x_{i1}} ... \left(e^{\beta_p} \right)^{x_{ip}} \tag{2.28}$$

โดยที่ β_0 คือ ค่าคงที่สัมประสิทธิ์การถดถอย สำหรับการประมาณค่าเฉลี่ย

β_j คือ สัมประสิทธิ์การถดถอย สำหรับการประมาณค่าเฉลี่ย

สามารถเขียนฟังก์ชันล็อกภาวะน่าจะเป็นของการแจกแจง CMP ได้ดังนี้

$$\log L(\lambda_i, \nu_i; y_i) = \sum_{y_i=0} \log \left\{ \frac{1}{Z(\lambda_i, \nu_i)} \frac{\lambda_i^{y_i}}{(y_i!)^{\nu_i}} \right\} \tag{2.29}$$

หรือสามารถเขียนฟังก์ชันล็อกภาวะน่าจะเป็นของ Mean Parametrized CMP ได้ดังนี้

$$\log L(\lambda_i, \nu_i; y_i) = \sum_{y_i=0} \log \left\{ \frac{1}{S(\lambda_{CMP_i}, \nu_i)} \left(\frac{\lambda_{CMP_i}^{y_i}}{y_i!} \right)^{\nu_i} \right\} \tag{2.30}$$

2.1.6 ตัวแบบการถดถอยคอนเวย์แม็กซ์เวลล์ปัวซงค่าศูนย์เฟ้อ
(Zero inflated Conway Maxwell Poisson: ZICMP regression model)

การแจกแจง ZICMP ถูกพัฒนาโดย Sellers และ Raim (2016) เป็นส่วนขยายของการแจกแจง CMP โดยสร้างมาจากพารามิเตอร์ 3 ตัว คือ พารามิเตอร์ค่าเฉลี่ย λ พารามิเตอร์การกระจาย ν โดยที่ $\nu > 0$ และพารามิเตอร์แสดงความน่าจะเป็นที่จะเกิดศูนย์ π โดยที่ $0 \le \pi \le 1$ ตัวแปรสุ่ม Y ที่มีการแจกแจง ZICMP แทนด้วย $Y \sim ZICMP(\lambda, \nu, \pi)$ ฟังก์ชันมวลความน่าจะเป็นของ Y เป็นดังนี้

$$\mathrm{P}(Y_i = y_i \mid \lambda_i, \nu_i, \pi_i) = \begin{cases} \pi_i + (1-\pi_i)\left(\dfrac{1}{Z(\lambda_i, \nu_i)} \right) & ; y_i = 0 \\ (1-\pi_i)\left(\dfrac{1}{Z(\lambda_i, \nu_i)} \cdot \dfrac{\lambda^{y_i}}{(y_i!)^{\nu_i}} \right) & ; y_i > 0 \end{cases} \tag{2.31}$$

โดยที่ $\lambda_i > 0,\ \nu > 0,\ 0 \le \pi_i < 1$ และ $i = 1, 2, 3, ..., n$

ฟังก์ชันเชื่อมโยงล็อกของ ZICMP สำหรับพารามิเตอร์ λ_i เหมือนกับการแจกแจง P ซึ่งแสดงไว้ดังสมการที่ (2.2) และ (2.3) ฟังก์ชันเชื่อมโยงล็อก ZICMP สำหรับพารามิเตอร์ ν_i เหมือนกับการแจกแจง CMP ซึ่งแสดงไว้ดังสมการที่ (2.22) และ (2.23) และฟังก์ชันเชื่อมโยงสำหรับพารามิเตอร์ π_i ใช้ฟังก์ชันเชื่อมโยงลอจิตเหมือนกับการแจกแจง ZIP ที่แสดงไว้ดังสมการ (2.9) และ (2.10)

ค่าเฉลี่ยและความแปรปรวนของ Y_i คือ

$$E(Y_i) = (1-\pi_i)\frac{1}{Z(\lambda_i,\nu_i)}\sum_{j=0}^{p}\frac{j\lambda_i^{\,j}}{(j!)^{\nu_i}} \tag{2.32}$$

$$Var(Y_i) = (1-\pi_i)\frac{1}{Z(\lambda_i,\nu_i)}\sum_{j=0}^{p}\frac{j^2\lambda_i^{\,j}}{(j!)^{\nu_i}} - E(Y_i) \tag{2.33}$$

สามารถเขียนฟังก์ชันล็อกภาวะน่าจะเป็นของ ZICMP ได้ดังนี้

$$\log L(\lambda_i,\nu_i,\pi_i;y_i) = \sum_{y_i=0}\log\left\{\pi_i + (1-\pi_i)\left(\frac{1}{Z(\lambda_i,\nu_i)}\right)\right\}$$

$$+\sum_{y_i>0}\log\left\{(1-\pi_i)\left(\frac{1}{Z(\lambda_i,\nu_i)}\cdot\frac{\lambda_i^{\,y_i}}{(y_i!)^{\nu_i}}\right)\right\} \tag{2.34}$$

2.1.7 การเรียนรู้ของเครื่อง (Machine learning)

บุญเสริม (2548) กล่าวว่า “ปัญญาประดิษฐ์ (Artificial intelligence: AI) หมายถึง ศาสตร์แขนงหนึ่งของวิทยาศาสตร์คอมพิวเตอร์ที่เกี่ยวข้องกับวิธีการทำให้คอมพิวเตอร์มีความสามารถคล้ายมนุษย์หรือเลียนแบบพฤติกรรมมนุษย์ คือโปรแกรม Software (ซอฟแวร์) ต่าง ๆ ที่ใช้กับคอมพิวเตอร์ โดยเฉพาะความสามารถในการคิดเองได้ โดยผ่านกระบวนการเรียนรู้ต่าง ๆ ที่เรียกว่า Machine learning” ซึ่งคำนิยาม AI ตามความสามารถที่มนุษย์ต้องการแบ่งได้ 4 กลุ่ม ดังนี้

1. การกระทำคล้ายมนุษย์ Acting humanly
2. การคิดคล้ายมนุษย์ Thinking humanly
3. คิดอย่างมีเหตุผล Thinking rationally
4. กระทำอย่างมีเหตุผล Acting rationally

Machine learning คือ ส่วนการเรียนรู้ของเครื่องถูกใช้งานเสมือนเป็นสมองของ AI เราอาจพูดได้ว่า AI ใช้ Machine learning ในการสร้างความฉลาดมักจะใช้เรียกโมเดลที่เกิดจากการเรียนรู้ของปัญญาประดิษฐ์ไม่ได้เกิดจากการเขียนโดยใช้มนุษย์ ซึ่งมนุษย์มีหน้าที่เขียนโปรแกรมให้

AI (เครื่อง) เรียนรู้จากข้อมูลเท่านั้น ลักษณะการเรียนรู้ของ Machine learning นั้นสามารถแบ่งออกเป็น 3 ประเภท ดังนี้

1. Supervised learning: เป็นกลุ่มของอัลกอริทึมที่มีลักษณะการพยากรณ์ผลลัพธ์ โดยอ้างอิงจากตัวอย่างที่ได้ทำการสอนหรือที่เรียกว่าข้อมูลชุดฝึกสอน (Training data) เช่น จะพยากรณ์ว่าตัวอักษรในภาพคืออะไรหรือรายได้ที่ควรคาดหวังจากการลงทุนเป็นเท่าใด โดยการที่มนุษย์มาคอยแยกประเภทหรือบอกผลลัพธ์ (Label) ที่ควรจะเป็น จากนั้นจะนำข้อมูลชุดฝึกสอนไปผ่านอัลกอริทึมสำหรับสร้างตัวแบบพยากรณ์ (Prediction model) เมื่อได้ตัวแบบพยากรณ์แล้วจะนำข้อมูลใหม่ที่เครื่องไม่เคยเห็น (New data) กล่าวคือไม่ใช่ข้อมูลชุดเดียวกันกับข้อมูลฝึกหัด เครื่องจะต้องพยากรณ์ (Prediction) ว่าคำตอบที่ได้ควรจะเป็นสิ่งใด โดยส่วนใหญ่ในการใช้ Supervised learning ในชีวิตจริงมักถูกนำไปใช้แทนการทำงานแบบ Rule base คือ มีกฎหรือรูปแบบการทำงานที่ตายตัวหรือสามารถอธิบายเหตุผลออกมาได้อย่างชัดเจน

2. Unsupervised learning: เป็นกลุ่มของอัลกอริทึมแบบไม่มีการสอนโดยจะไม่มีการระบุผล (target variable) ที่ต้องการให้คอมพิวเตอร์หาความสัมพันธ์จากข้อมูลด้วยตนเอง

3. การเรียนรู้แบบเสริมกำลัง (Reinforcement learning) เป็นการเรียนรู้ที่คอมพิวเตอร์จะสนใจต่อสิ่งแวดล้อมเป็นพิเศษ โดยคอมพิวเตอร์มีปฏิสัมพันธ์กับสิ่งแวดล้อมที่เปลี่ยนไปตลอดเวลาโดยคอมพิวเตอร์จะต้องทำงานบางอย่าง (เช่น ขับรถ) โดยที่ไม่มี “ผู้สอน” คอยบอกอย่างจริงจังว่าวิธีการที่ทำอยู่นั้นเข้าใกล้เป้าหมายแล้วหรือไม่

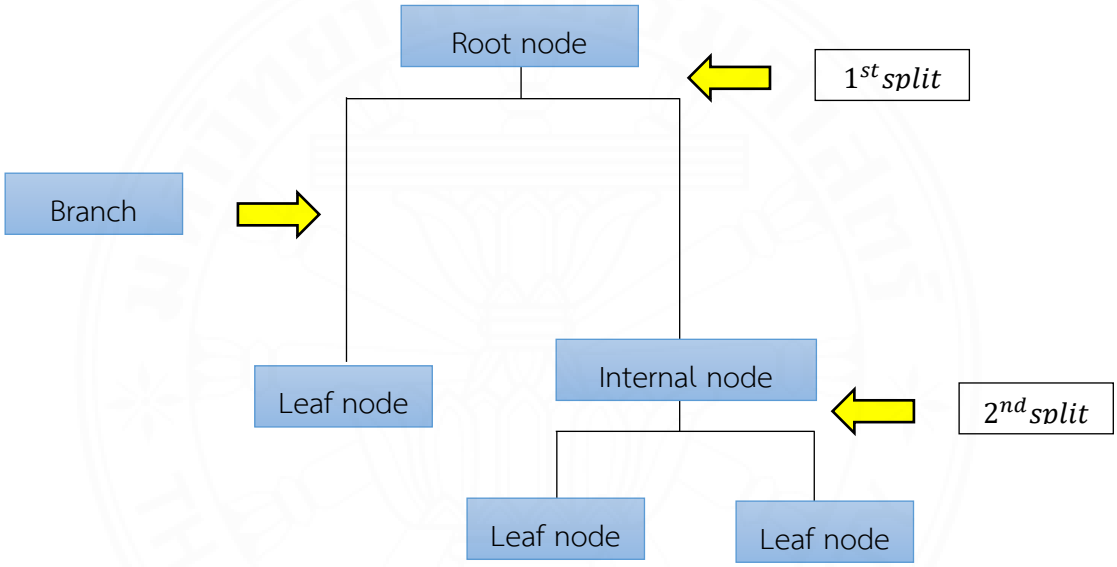
2.1.7.1 ต้นไม้ตัดสินใจ (Decision tree)

ต้นไม้ตัดสินใจ หรือ Decision tree ถูกคิดค้นโดย Kam (1995) เป็นเทคนิคการพยากรณ์ข้อมูลในลักษณะหนึ่งเพื่อการหาทางเลือกที่ดีที่สุด ซึ่งไม่จำเป็นที่จะต้องมีข้อตกลง เงื่อนไข หรือข้อกำหนดสำหรับการแจกแจงใด ๆ นอกจากนี้เทคนิคต้นไม้ตัดสินใจสามารถนำไปใช้กับรูปแบบความสัมพันธ์ของข้อมูลที่เป็นเชิงเส้นและไม่ใช่เป็นเชิงเส้นได้ โดยการนำข้อมูลมาสร้างตัวแบบพยากรณ์ในรูปแบบของโครงสร้างต้นไม้ซึ่งมีการเรียนรู้ข้อมูลแบบมีผู้สอน สามารถสร้างตัวแบบการจัดหมวดหมู่ (Clustering) ได้จากกลุ่มตัวอย่างของข้อมูลที่กำหนดไว้ล้วงหน้าและสามารถพยากรณ์กลุ่มของรายการที่ยังไม่เคยนำมาจัดหมวดหมู่ได้อีกด้วย สำหรับโครงสร้างของต้นไม้ตัดสินใจมีลักษณะคล้ายกับโครงสร้างต้นไม้จริง ซึ่งประกอบด้วย ราก กิ่ง และใบ โดยการแตกแขนงไปตามเงื่อนไขหรือเส้นทางของกิ่งไม้และจะได้ค่าพยากรณ์ออกมาในที่สุดซึ่งก็คือใบ ในการกำหนดเงื่อนไขใช้กระบวนการในรูปแบบ “ถ้า (เงื่อนไข) แล้ว (ผลลัพธ์)” (If-then Rule) มาประกอบโครงสร้างต้นไม้ตัดสินใจ (ภาพที่ 2.1) โครงสร้างต้นไม้ตัดสินใจจะประกอบด้วย

โหนดภายใน (Internal node) คือ โหนดที่แสดงถึงคุณลักษณะ (Feature) ที่นำมาใช้ในการแบ่งกลุ่มของข้อมูลซึ่งมี โหนดราก (Root node) เป็นส่วนที่อยู่บนสุดของโครงสร้าง ซึ่งเป็นโหนดที่มีอิทธิพลต่อการจำแนกกลุ่มมากที่สุด

กิ่ง (Branch) เป็นตัวเชื่อมระหว่างโหนดที่ใช้เป็นเงื่อนไขหรือทางเลือกของกระบวนการ ซึ่งมาจากอิทธิพลหรือคุณสมบัติของตัวแปรอิสระแต่ละตัว

โหนดใบ (Leaf node) เป็นโหนดที่แสดงผลลัพธ์ของเงื่อนไขหรือกระบวนการตามเงื่อนไขที่เกิดขึ้น ซึ่งในที่นี้ผลลัพธ์ก็คือ ค่าพยากรณ์



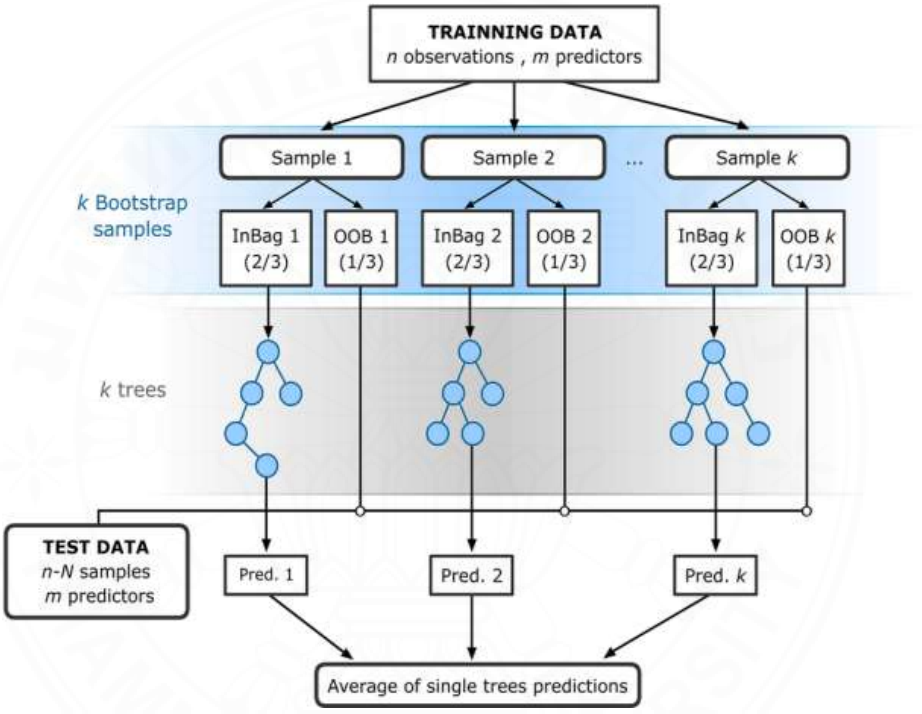
ภาพที่ 2.1 โครงสร้างต้นไม้ตัดสินใจ

จากภาพที่ 2.1 เป็นโครงสร้างสำหรับการสร้างต้นไม้ตัดสินใจต้นเพียงต้นเดียว ซึ่งต้นไม้ตัดสินใจแต่ละต้นจะมีลักษณะที่คล้ายกัน โดยผ่านกระบวนการการบรรจุถุง (Bagging) คือ การนำข้อมูลไปสุ่มเลือกตัวอย่างแบบคืนที่ (Sampling with replacement) ซึ่งประกอบไปด้วยตัวอย่างที่ถูกเลือก เรียกว่า In-Bag และตัวอย่างส่วนหนึ่งที่ไม่ถูกเลือกเรียกว่า Out-Of-Bag (OOB) หลังจากนั้นจะเริ่มต้นทำการคัดเลือกตัวแปรอิสระที่มีความสัมพันธ์กับตัวแปรตอบสนองมากที่สุดขึ้นมาเป็น Root node และจะทำการแตกกิ่งตัวแปรอิสระออกไปเรื่อย ๆ จนถึงเกณฑ์ที่ระบุไว้เพื่อจบกระบวนการ

2.1.7.2 เทคนิคป่าสุ่ม (Random forest : RF)

เทคนิค RF ถูกพัฒนาโดย Breiman (2001) เป็นตัวแบบที่ไม่มีข้อตกลงหรือเงื่อนไขในการใช้งานเหมือนตัวแบบทางสถิติเช่นเดียวกับต้นไม้ตัดสินใจ โดยใช้เทคนิคการสุ่มเลือกใช้ข้อมูลและคุณลักษณะจากกระบวนการต้นไม้ตัดสินใจ ซึ่งถูกสร้างจากการนำข้อมูลไปสุ่มเลือก

ตัวอย่างแบบคืนที่ แล้วนำมาสร้างเป็นต้นไม้ โดยเทคนิคป่าสุ่มมีฐานความคิดจากอัลกอริทึมต้นไม้ตัดสินใจที่เน้นการสร้างตัวแบบด้วยวิธีการ Decision tree ขึ้นมาหลายๆ ตัวแบบที่ไม่ซ้ำกัน โดยตัวอย่างที่ถูกเลือกเข้ามาเป็นข้อมูลชุดฝึกสอน (Training data) คือ In-Bag และ OOB จะถูกนำมาใช้ในการทดสอบต้นไม้ตัดสินใจด้วยวิธีการบรรจุถุง (Bagging) ผลลัพธ์ที่ได้อย่างอิสระจากต้นไม้ตัดสินใจในแต่ละต้นจะถูกนำมาคิดเป็นผลการโหวตที่มากที่สุดหรือค่าเฉลี่ย (ดังภาพที่ 2.2) ซึ่งค่าดังกล่าวจะถูกเลือกนำมาเป็นตัวแทนของกระบวนการต้นไม้



ภาพที่ 2.2 กระบวนการของเทคนิคป่าสุ่ม

(ที่มา : Rodriguez-Galiano, Chica-Olmo & Chica-Rivas (2014))

2.1.7.3 กระบวนการของเทคนิคป่าสุ่ม

ขั้นที่ 1 ทำการแบ่งข้อมูลเป็นออก 2 ชุด คือ ชุดฝึกสอนและชุดทดสอบ (Testing data) โดยข้อมูล 1 ใน 2 ส่วนนี้จะถูกใช้เป็นข้อมูลชุดฝึกสอนและข้อมูลส่วนที่เหลือเป็นชุดทดสอบหรือที่เรียกว่า “OOB” ส่วนการพิจารณาว่าข้อมูลส่วนใดจะเป็นข้อชุดทดสอบถูกกำหนดด้วยความน่าจะเป็นที่เท่ากับ $\frac{1}{3}$ จากนั้นทำการสร้างต้นไม้ตัดสินใจจากข้อมูลชุดฝึกสอนด้วยการนำเข้าตัวแปรอิสระตามจำนวนที่กำหนดไว้ m ตัว

ขั้นที่ 2 สมมติมีตัวแปรอิสระทั้งหมด p ตัว นำตัวแปรอิสระที่มีความสำคัญ (Variable importance: VIMP) กับตัวแปรตอบสนองมากที่สุดมาเป็นโหนดรากหรือจุดเริ่มต้น

ขั้นที่ 3 จะสุ่มเลือกตัวแปรอิสระ m ตัว จากตัวแปรอิสระทั้งหมด p ตัว โดยที่ $m < p$ ซึ่งในแต่ละชุดจะนำมาสร้างต้นไม้ตัดสินใจที่มีสองทางเลือก (Binary tree) หรือ กิ่ง (branch) โดยใช้จุดแบ่งที่ดีที่สุด (Best Split Point: C_i) และจะพิจารณาตัวแปรอิสระทีละตัว ลำดับในการพิจารณาขึ้นอยู่กับความสำคัญของตัวแปรอิสระนั้นกับตัวแปรตอบสนอง ตัวแปรที่มีความสำคัญมากที่สุด (รองลงมาจากโหนดราก) จะนำมาพิจารณาเป็นโหนดภายใน (Internal node) ลำดับแรก จากนั้นสร้างทางเลือกแบบสองทางเพื่อนำตัวแปรอิสระที่มีความสำคัญกับตัวแปรตอบสนองลำดับที่สองเข้ามาเป็นโหนดภายในลำดับที่สอง ทำเช่นนี้จนกระทั่งครบ m ตัว ในแต่ละโหนด (แต่ละตัวแปรอิสระ) จะหาจุดแบ่งที่แยกข้อมูลออกเป็นสองส่วน (YES กับ NO) ในที่นี้จุดแบ่งที่ 1 คือ C_1 และ จุดแบ่งที่ 2 คือ C_2 ตามลำดับจนถึงจุดแบ่งที่ m คือ C_m

ขั้นที่ 4 จากขั้นตอนที่ 3 เมื่อพิจารณาตัวแปรอิสระจนครบ m ตัวแล้ว จะได้จุดแบ่งทั้งหมด m จุด ($c_1, c_2, ..., c_m$) จากนั้นนำจุดแบ่งข้างต้นไปใช้แบ่งข้อมูล OOB เพื่อทำการหาค่าพยากรณ์

ขั้นที่ 5 ทำซ้ำในขั้นตอนที่ 1 ถึง 4 จำนวนทั้งหมด ntree ครั้ง เพื่อสร้างต้นไม้ตัดสินใจ ntree ต้น

ขั้นที่ 6 จากต้นไม้ตัดสินใจทั้ง ntree ต้น จะคำนวณค่าคลาดเคลื่อนกำลังสองเฉลี่ยของ OOB (MSE_{OOB_i}) ของแต่ละต้น โดยการคำนวณ OBB นี้จะนำตัวแบบทั้ง ntree ต้นมาใช้กับชุดของตัวแปรอิสระในข้อมูล OOB ของขั้นตอนที่ 1

ขั้นที่ 7 กิ่งใดที่ให้ค่า MSE_{OOB_i} ต่ำที่สุดกิ่งนั้นจะถูกเลือกเป็นตัวแบบพยากรณ์ของต้นไม้นั้น ๆ

ขั้นที่ 8 ค่าพยากรณ์ที่ได้จากต้นไม้ทั้ง ntree ต้น จะถูกนำมาหาค่าเฉลี่ยเพื่อให้ได้ค่าพยากรณ์เพียงค่าเดียว

2.1.7.4 การระบุพารามิเตอร์ในเทคนิคป่าสุ่ม

เป็นการกำหนดค่าเริ่มต้นสำหรับโปรแกรม R โดยมีพารามิเตอร์ต่าง ๆ ดังนี้

1. จำนวนต้นไม้ที่จะเติบโตในป่า (ntree)

เป็นพารามิเตอร์ที่สำคัญที่สุดของป่าสุ่มจำนวนต้นไม้เริ่มต้นคือ 500 แม้ว่าผลลัพธ์จะมีความเสถียรภาพมากขึ้นแต่ก็ส่งผลให้ไม่สามารถอธิบายความสัมพันธ์ระหว่างตัวแปรตอบสนองกับปัจจัยจากตัวแบบ

2. จำนวนสุ่มเลือกตัวแปรตัวอิสระในต้นไม้แต่ละต้น ($mtry:m$)

ซึ่งค่าเริ่มต้นของ m คือ $p/3$ สำหรับการถดถอย (Regression) และ $p^{1/2}$ สำหรับการจำแนกประเภท (Classification) โดยที่ p คือจำนวนของตัวแปรอิสระ

3. จำนวนขนาดของโหนดขั้วของต้นไม้ที่น้อยที่สุด (Node size)

โหนดเริ่มต้นคือ n=5 สำหรับการถดถอย (Regression) และ n=1 สำหรับการจำแนกประเภท (Classification) (James และคณะ, 2013)

2.1.7.5 ค่าคลาดเคลื่อนกำลังสองเฉลี่ยของ OOB

(Out of Bag Mean Square Error)

Out of Bag (OOB) คือ ตัวอย่างส่วนหนึ่งที่ไม่ถูกเลือก โดยถูกแบ่งออกคิดเป็น $\frac{1}{3}$ ของข้อมูลทั้งหมด (Breiman, 2001 และ Grömping, 2009) เพื่อนำมาใช้ในการทดสอบต้นไม้ตัดสินใจที่ได้จากข้อมูลชุดทดสอบ ซึ่งวิธีการนี้เรียกว่า Bagging โดยการคำนวณค่าคลาดเคลื่อนกำลังสองเฉลี่ยของ OOB สามารถคำนวณได้ดังนี้

$$MSE_{OOB} = \frac{1}{n}\sum_{i=1}^{n}\left\{y_i - \bar{\hat{y}}_{OOB_i}\right\}^2 \qquad (2.33)$$

โดยที่ $\bar{\hat{y}}_{OOB_i}$ คือ ค่าเฉลี่ยของการพยากรณ์ที่ i ที่ได้จากต้นไม้ทุกต้นในข้อมูล OOB

y_i คือ ค่าสังเกตที่ i ในข้อมูล OOB

ในงานวิจัยนี้จะทำการเลือกชุดพารามิเตอร์รวมถึง ntree, mtry และ node size สำหรับเทคนิคป่าสุ่มที่ทำการเปรียบเทียบชุดพารามิเตอร์ที่แตกต่างกันแล้ว หากชุดพารามิเตอร์ใดให้ค่าเฉลี่ยของค่าคลาดเคลื่อนกำลังสองเฉลี่ย ($mMSE_{OOB}$) ต่ำที่สุดจะถูกเลือกใช้เป็นพารามิเตอร์ของเทคนิคป่าสุ่ม

2.1.7.6 หลักการเลือกตัวแปรอิสระเพื่อแบ่งโหนดการตัดสินใจ

หลักการเลือกตัวแปรที่ใช้ในการแบ่งโหนดการตัดสินใจ (Node) เพื่อสร้างจำนวนข้อมูลในแต่ละกลุ่มที่แบ่งออกมา (Leaf node) ของอัลกอริทึมเทคนิค RF จะใช้ตัวแปรทุกตัวในการประมวลผลไม่ต้องดำเนินการหาตัวแปรที่มีความสำคัญผ่านการพิจารณาค่า p-value หรือการทดสอบนัยสำคัญ เช่น ตัวแบบการถดถอยอื่น ๆ แต่จะประมวลผลจากสุ่มหาตัวแปรอิสระที่สามารถใช้แยกกลุ่มของกลุ่มตัวอย่างออกจากกันอย่างมีนัยสำคัญได้ โดยวิธีการในการแบ่งแยก

คัดเลือกตัวแปร คือ การพิจารณาค่า Variable importance (VIMP) ซึ่งสามารถสรุปรายละเอียดได้ดังนี้

ความสำคัญของตัวแปร (Variable importance: VIMP)

ความสำคัญของตัวแปร (VIMP) คือ ค่าส่วนต่างของค่าคลาดเคลื่อนกำลังสองเฉลี่ยของการพยากรณ์ชุดข้อมูลที่ตัวแปรอิสระไม่ได้ถูกสุ่มมาใช้ (OOB Prediction Error) กับ ชุดที่ตัวแปรอิสระถูกสุ่มนำมาใช้ (Breiman, 2001) พิจารณาตัวอย่าง OOB ที่เกี่ยวข้อง (ข้อมูลไม่รวมอยู่ในตัวอย่าง boostrap ที่ใช้สร้างต้นไม้แต่ละต้น) โดย X^j สำหรับต้นไม้แต่ละต้น (t) ถูกกำหนดไว้ดังนี้

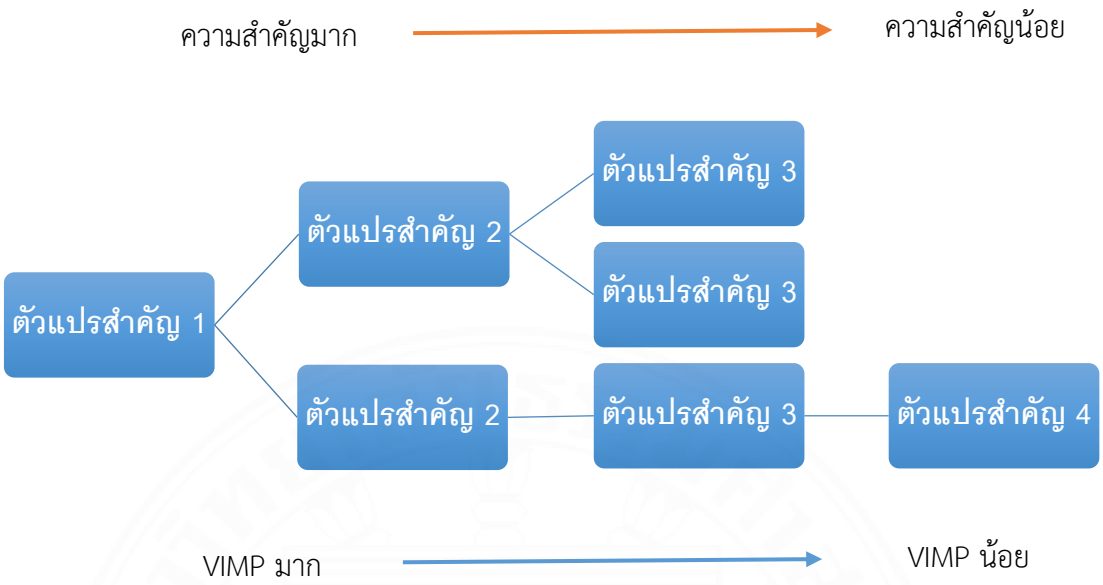
$$VIMP(X_j) = \frac{1}{ntree}\sum_t (MSE^j{}_{OOB_t} - MSE_{OOB_t}) \qquad (2.35)$$

โดยที่ OOB_t คือ ข้อมูลที่ไม่รวมอยู่ในตัวอย่าง bootstrap ที่ใช้สร้างต้นไม้ต้นที่ t

MSE_{OOB_t} คือ ค่าอัตราความผิดพลาดของ OOB จาก OOB_t ที่ตัวแปรอิสระไม่ได้ถูกสุ่มมาใช้

$MSE^j{}_{OOB_t}$ คือ ค่าอัตราความผิดพลาดของ OOB จาก OOB_t ที่ตัวแปรอิสระที่ j ที่ถูกสุ่มมาใช้

หากตัวแปรที่มีค่า VIMP มาก เมื่อมีการสุ่มข้อมูลใหม่จะส่งผลให้ความแปรปรวนของผลลัพธ์มาก กล่าวคือ เมื่อนำตัวแปรไปพิจารณาในชุดข้อมูลกับกรณีที่ตัวแปรไม่ถูกนำไปพิจารณาในชุดข้อมูลจะมีค่าคลาดเคลื่อนของผลลัพธ์ต่างกันสูง สะท้อนถึงความสำคัญของตัวแปรนั้น ๆ ที่มีต่อผลลัพธ์การพยากรณ์สำหรับตัวแปรที่มีค่า VIMP เข้าใกล้ศูนย์และค่าเป็นลบบ่งบอกว่า เมื่อมีการสุ่มข้อมูลใหม่ผลลัพธ์ที่ได้ค่าคลาดเคลื่อนก็ยังคงเหมือนเดิม แสดงถึงตัวแปรดังกล่าวไม่ได้ส่งผลต่อความแปรปรวนของผลลัพธ์สะท้อนการมีความสำคัญต่ำ (มีตัวแปรนี้หรือไม่ค่าผลลัพธ์ไม่ต่างกันมาก) โดยการนำมาใช้นั้นตัวแบบมักจะหลีกเลี่ยงตัวแปรที่มีค่า VIMP ที่เข้าใกล้ศูนย์และค่าเป็นลบโดยจะเลือกใช้ตัวแปรที่มีค่าเป็นบวก ซึ่งเมื่อมีการสุ่มข้อมูลใหม่แล้วจะส่งผลต่อการแปรปรวนของผลลัพธ์พยากรณ์เพื่อสะท้อนถึงความสำคัญของตัวแปรนั้น ๆ ที่มีต่อการพยากรณ์ผลลัพธ์ ตามภาพที่ 2.3 ที่แสดงไว้ดังนี้



ภาพที่ 2.3 ความสำคัญของตัวแปร

2.1.7.7 การหาจุดแบ่งที่ดีที่สุด (Best split point: c)

1. ผลบวกกำลังสองของความคลาดเคลื่อน

(Residual sum square: RSS)

การหาจุดแบ่งที่ดีที่สุดของข้อมูล (Best Split point: C_i) สำหรับการแบ่งข้อมูลชุดฝึกสอนที่มีจำนวนตัวแปร m ตัวแปรและมีขนาดตัวอย่าง n ตัว $(x_{ij}; i=1,2,...,m; j=1,2,...,n)$ ซึ่งในแต่ละครั้งที่พิจารณาตัวแปรอิสระ 1 ตัวนั้นจะสามารถหาจุดแบ่งที่เป็นไปได้ n จุด โดยการคำนวณค่า RSS จากค่าพยากรณ์ในแต่ละกลุ่มที่ถูกแบ่งเพื่อสร้างทางเลือก ค่าพยากรณ์ในที่นี้ คือ ค่าเฉลี่ยของตัวแปรตอบสนองในแต่ละกลุ่มย่อย (R_{bj}) ซึ่งจะเลือกจุด Split point ที่ให้ค่า RSS น้อยที่สุด

$$RSS = \sum_{j=1}^{n} \sum_{i \in R_{bj}} \left(y_{ij} - \bar{y}_{R_{bj}} \right)^2 \qquad (2.36)$$

โดยที่ R_{bj} คือ ค่าสังเกต j ของตัวแปรอิสระ i ที่ถูกแบ่งออกมาในการสร้างทางเลือกทั้งหมด b กลุ่ม ในที่นี้ $b=1,2$ (Yes และ No) โดยที่ $i=1,2,...,m$ และ $j=1,2,...,n$

y_{ij} คือ ค่าตัวแปรตอบสนอง

$\bar{y}_{R_{bj}}$ คือ ค่าเฉลี่ยในแต่ละกลุ่มซึ่งคำนวณมากจาก y_{ij} ใน R_{bj}

2. ขั้นตอนการหาจุดแบ่งที่ดีที่สุด (C_i)

ขั้นที่ 1 หา VIMP จากตัวอิสระทั้งหมด P ตัว

ขั้นที่ 2 ทำการหาจุดแบ่งข้อมูล โดยการพิจารณาตัวแปรอิสระที่ให้ค่า VIMP มากที่สุดถูกเลือกมาเป็นโหนดราก (Root node) สมมติว่าตัวแปร X_1 เป็นตัวแปรที่ให้ค่า VIMP มากที่สุด (เป็นตัวแปรที่มีความสำคัญมากที่สุด) ซึ่งตัวแปร X_1 จะมีค่าทั้งหมด n ค่า $(x_{11}, x_{12}, ..., x_{1n})$ และนำค่าของ $x_{1j}; j = 1, 2, ..., n$ แต่ละตัวมาเป็นจุดแบ่ง (Split point: c_{1j}) โดยนำจุดแบ่ง c_{1j} มาแบ่งข้อมูลออกเป็นสองส่วนตามเงื่อนไข ($x_{1j} \le c_{1j}$ และ $x_{1j} > c_{1j}$) กำหนดให้ R_{1j} เป็นกลุ่มของค่าสังเกตที่มีค่าของตัวแปร X_1 น้อยกว่าหรือเท่ากับ c_{1j} และ R_{2j} เป็นกลุ่มของค่าสังเกตที่มีค่าของตัวแปร X_1 มากกว่า c_{1j} จากนั้นคำนวณค่าเฉลี่ยของค่าสังเกต จะได้ $\bar{y}_{R_{1j}}$ และ $\bar{y}_{R_{2j}}$ ในข้อมูล R_{1j} และ R_{2j} ตามลำดับ นั้นคือจะได้จุดแบ่งที่เป็นไปได้ทั้งหมด คือ $c_{11}, c_{12}, ..., c_{1n}$

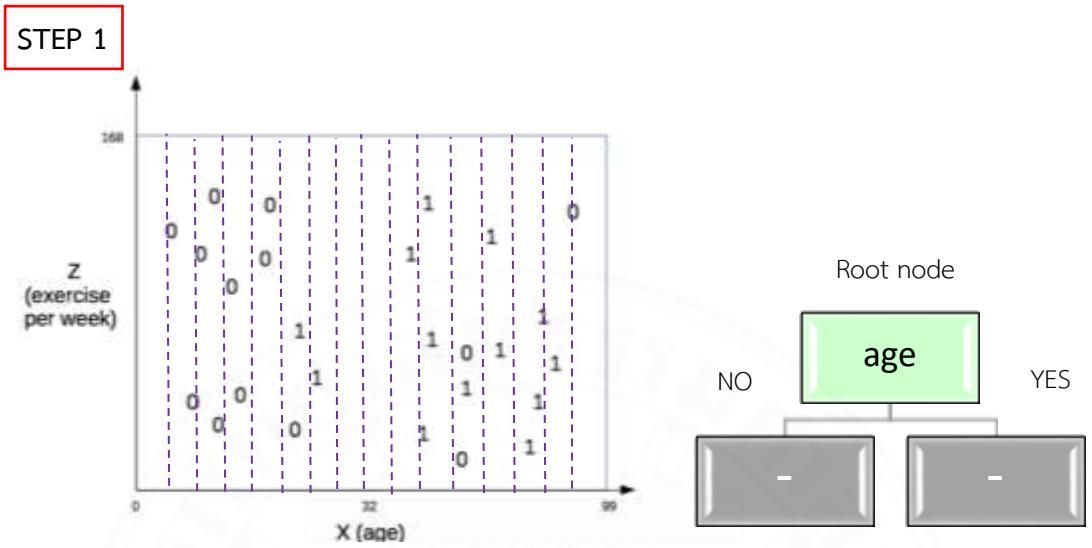
ขั้นที่ 3 จากขั้นตอนที่ 2 จะคำนวณค่าผลรวมความคลาดเคลื่อนกำลังสอง (RSS) ดังสมการที่ (2.32) ของแต่ละจุดแบ่ง $c_{11}, c_{12}, ..., c_{1n}$ จากข้อมูลทั้งสอง (R_{1j} และ R_{2j}) โดยเลือก split point ณ จุด c_{1j} ที่ให้ค่า RSS น้อยที่สุดมาเป็นจุดแบ่งที่ดีที่สุด (c_1) การพิจารณาหาจุดแบ่งที่ดีที่สุด (c_i) ของตัวแปรอิสระตัวอื่น ๆ (X_i) ก็จะพิจารณาในทำนองเดียวกันกับ X_1

ขั้นที่ 4 เมื่อได้จุดแบ่งที่ดีที่สุดสำหรับโหนดแล้ว เช่น กำหนดให้ $x_{1j} \le c_1$ เป็นทางเลือกที่ไม่สอดคล้องกับเงื่อนไข (NO) จะแสดงผลลัพธ์ค่าพยากรณ์ในที่นี้คือค่าเฉลี่ย $\bar{y}_{R_{1j}}$ ในทางกลับกันเมื่อ $x_{1j} > c_1$ เป็นทางเลือกที่สอดคล้องกับเงื่อนไข (YES) จะดำเนินการสร้างโหนดถัดไป โดยพิจารณาตัวแปรอิสระที่มีความสำคัญรองลงมา เพื่อหาจุดแบ่งที่ดีที่สุดในขั้นถัดไป

ขั้นที่ 5 ในการสร้างโหนดต่อไปนั้น จะทำการสุ่มตัวแปรอิสระ m ที่มี VIMP รองลงมาเพื่อเข้าไปเป็นโหนดที่สอง ซึ่งทำการหาจุดแบ่งที่ดีที่สุด (c_2) เช่นเดียวกับขั้นตอนที่ 1 ถึง ขั้นตอนที่ 3 จนกว่าจะครบ m ตัว $(c_1, c_2, ..., c_m)$ หลังจากนั้นทำการแสดงค่าพยากรณ์ของขั้นสุดท้าย

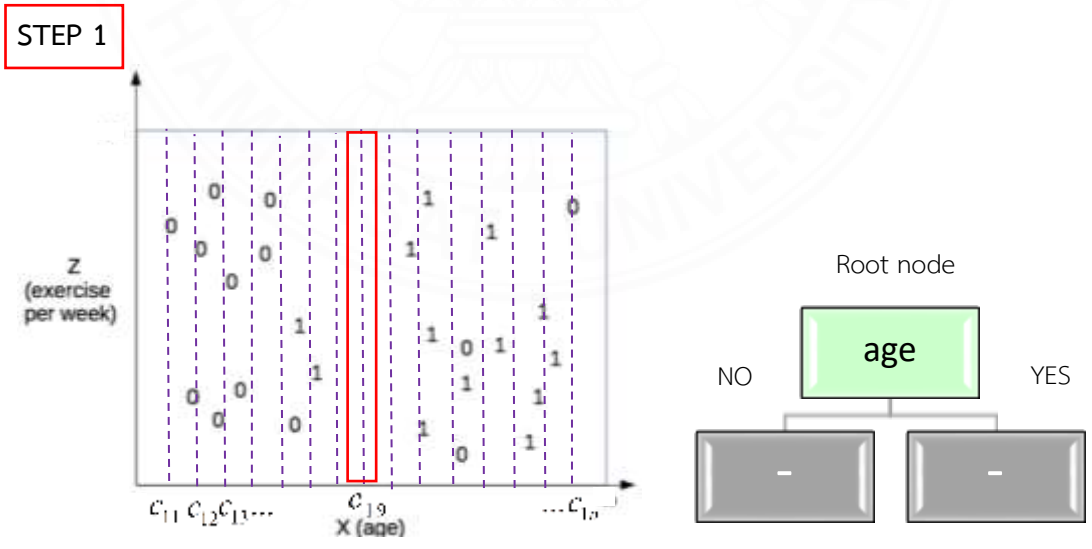
ขั้นที่ 6 ทำซ้ำขั้นตอนที่ 1 ถึง 5 ในต้นไม้ตัดสินใจจนครบ ntree ต้น

แผนภาพแสดงตัวอย่างขั้นตอนการหาจุดแบ่งที่ดีที่สุด



ภาพที่ 2.4 (a) ขั้นตอนการหาจุดแบ่งที่ดีที่สุด (Step 1)

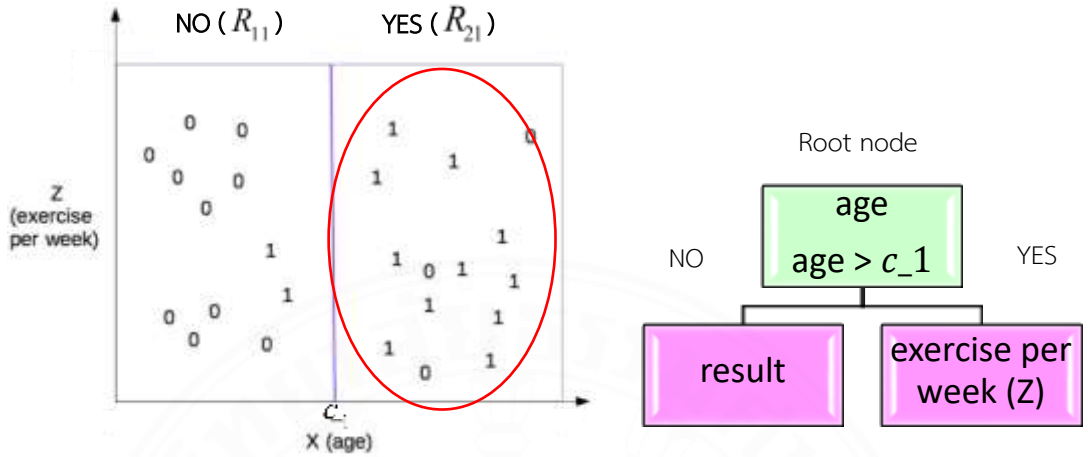
แบ่งกลุ่ม (R_{bj}) ออกเป็น 2 กลุ่มตามค่าของ $X(age)$ เพื่อหาจุดแบ่ง c_{ij} ที่เป็นไปได้ทั้งหมดจากสูตร $RSS = \sum_{j=1}^{n} \sum_{i \in R_{bj}}^{b} \left(y_{ij} - \bar{y}_{R_{bj}} \right)^2$



ภาพที่ 2.4 (b) ขั้นตอนการหาจุดแบ่งที่ดีที่สุด (Step 1)

ทำการเปรียบเทียบ c_{ij} ที่ให้ค่าต่ำสุดมาเป็นจุดแบ่งที่ดีที่สุด (สมมติ c_1 ในที่นี้ คือ c_{19}) ดังภาพที่ 2.4 (b)

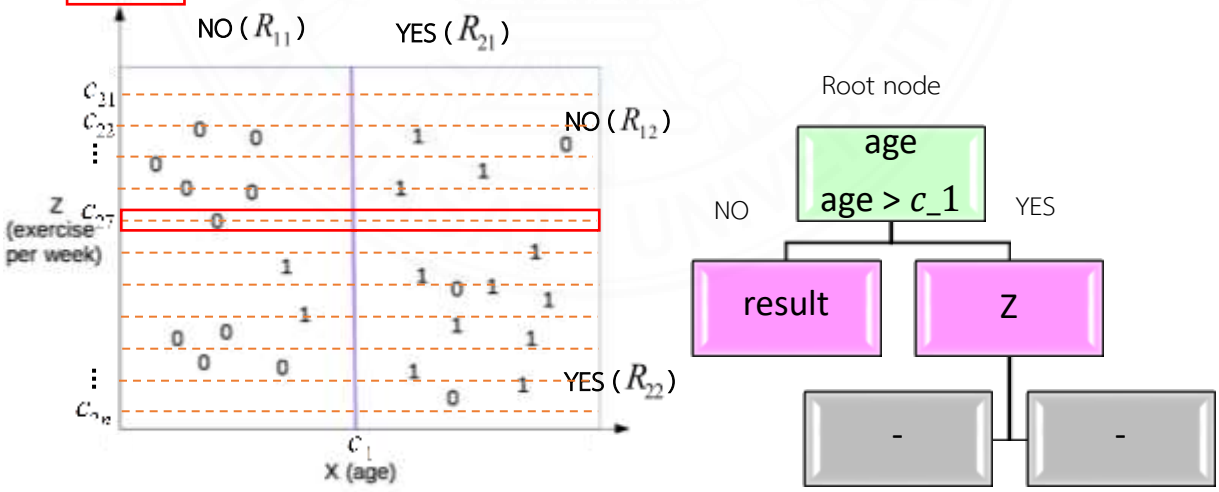
STEP 2



ภาพที่ 2.4 (c) ขั้นตอนการหาจุดแบ่งที่ดีที่สุด (Step 2)

สมมติให้ $X(age) = c_1$ เป็นจุดตัดที่ดีที่สุด ทำการแบ่งข้อมูลออกเป็นสองกลุ่ม ได้แก่ R_{11} และ R_{21} ดังรูปที่ 2.4 (c) สำหรับทางเลือกที่สอดคล้องกับเงื่อนไข YES จะดำเนินการสร้างโหนดที่สองต่อไป

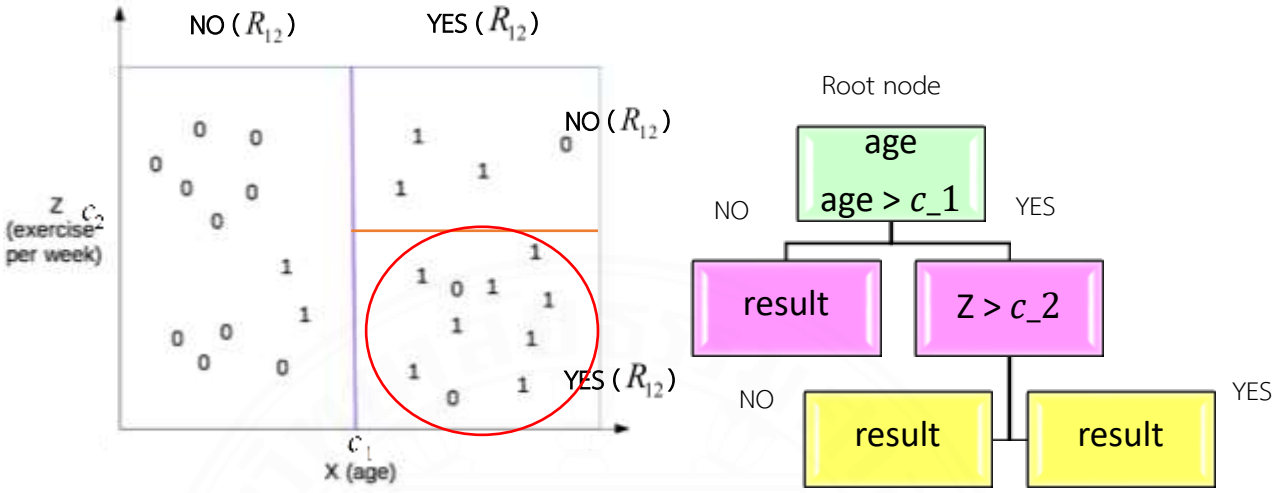
STEP 3



ภาพที่ 2.4 (d) ขั้นตอนการหาจุดแบ่งที่ดีที่สุด (Step 3)

จากภาพ 2.4 (d) ทำการแบ่งกลุ่ม (c_{21}) กลุ่มตามค่าของ $X(Z)$ เพื่อหาจุดแบ่ง c_{ij} ที่เป็นไปได้ทั้งหมด n จุด และทำการเปรียบเทียบ c_{ij} ที่ให้ค่าต่ำสุดมาเป็นจุดแบ่งที่ดีที่สุด (สมมติ c_2 ในที่นี้ คือ c_{27})

STEP 4



ภาพที่ 2.4 (e) ขั้นตอนการหาจุดแบ่งที่ดีที่สุด (Step 4)

สมมติให้ $X(Z)=c_2$ เป็นจุดตัดที่ดีที่สุด ทำการแบ่งข้อมูลออกเป็นสองกลุ่ม ได้แก่ R_{21} และ R_{22} ดังภาพที่ 2.4 (e) โดยจะทำการแบ่งในชุดข้อมูลที่สอดคล้องกับเงื่อนไข YES (R_{12}) ของ $X(age)$ ก่อนหน้านี้ เมื่อครบเกณฑ์ที่กำหนดจึงแสดงผลลัพธ์สุดท้าย

สำหรับในการสร้างโหนดต่อไปจะทำเช่นนี้ไปเรื่อยๆจนจำนวนตัวแปรอิสระครบ m ตัว จึงจะแสดงผลลัพธ์สุดท้าย และกระบวนการนี้จะถูกนำไปใช้ในต้นไม้ตัดสินใจต้นอื่น ๆ จนครบ ntree ต้น

2.2 สถิติทดสอบสกอร์ (Score test statistic)

สถิติทดสอบสกอร์ถูกนำเสนอโดย Van den Broek (1995) สำหรับทดสอบค่าศูนย์เฟ้อในการแจกแจงปัวซง ภายใต้สมมติฐานว่างที่ว่าข้อมูลไม่ได้มีสัดส่วนของค่าศูนย์เฟ้อ ($H_0:\pi=0$) ซึ่งสถิติทดสอบสกอร์จะมีการแจกแจงไคกำลังสอง (χ_1^2) ที่มีองศาเสรีเท่ากับ 1 เมื่อสมมติฐานว่างเป็นจริง หากปฏิเสธสมมติฐานว่างจะบ่งบอกว่าข้อมูลมีการแจกแจง ZIP โดยมีสูตรการคำนวณดังนี้

$$S=\frac{(n_0-n\hat{\pi})^2}{n\hat{\pi}(1-\hat{\pi})-n\bar{y}\hat{\pi}^2} \qquad (2.37)$$

โดยที่ $\hat{\lambda} = \bar{y}$

$\hat{\pi} = P(Y_i = 0) = e^{-\hat{\lambda}}$

n_0 คือ จำนวนของค่าศูนย์

n คือ จำนวนของค่าสังเกตทั้งหมด

$\bar{y}$ คือ ค่าเฉลี่ยของข้อมูล

โดยจะปฏิเสธสมมติฐานว่างเมื่อ S มากกว่า $\chi^2_{\omega,df=1}$ ที่ระดับนัยสำคัญ ω และองศาเสรีเท่ากับ 1 หรือค่า p-value มีค่าน้อยกว่าระดับนัยสำคัญ ω

2.3 การทดสอบการกระจาย (Dispersion test)

สถิติการทดสอบการกระจายถูกนำเสนอโดย Cameron และ Trivedi (1986) การทดสอบดังกล่าวเป็นวิธีที่พิจารณาว่าความชันของสมการถดถอย (Regression slope) เบี่ยงเบนไปจากศูนย์เพียงใด โดย α คือ พารามิเตอร์การกระจายเกินเกณฑ์ของตัวแบบการถดถอย NB ภายใต้สมมติฐานว่างที่ว่าค่าเฉลี่ยไม่แตกต่างจากความแปรปรวน ($H_0 : \alpha = 0$) หากปฏิเสธสมมติฐานว่างจะบ่งบอกว่าข้อมูลมีปัญหาการกระจาย โดยมีสูตรการคำนวณดังนี้

$$t = \frac{(y_i - \hat{\lambda}_i)^2 - y_i}{\hat{\lambda}_i} \tag{2.38}$$

โดยที่ $\hat{\lambda}_i$ คือ ค่าประมาณค่าเฉลี่ยจากตัวแบบการถดถอย P

y_i คือ ค่าสังเกต

โดยจะปฏิเสธสมมติฐานว่างเมื่อ เมื่อ t มากกว่า $t_{\omega,df=n-1}$ ที่ระดับนัยสำคัญ ω และองศาเสรีเท่ากับ 1 หรือค่า p-value มีค่าน้อยกว่าระดับนัยสำคัญ ω

2.4 เกณฑ์สารสนเทศของอะกะอิเกะ (Akaike's information criterion: AIC)

เกณฑ์ AIC ถูกพัฒนาขึ้นโดย Akaike (1974) จนถึงปัจจุบันมักถูกใช้เป็นตัวสถิติเพื่อแสดงผลในส่วนของตัวแบบทางสถิติ ตัวแบบที่ให้ค่า AIC ต่ำสุดจะเป็นตัวแบบที่ดีที่สุด โดยทั่วไปแล้วเกณฑ์ AIC จะถูกพบได้ 2 รูปแบบดังสมการที่ (2.38) และ (2.39)

$$\text{AIC} = \frac{-2L + 2k}{n} = \frac{-2(L - k)}{n} \tag{2.39}$$

หรือ

$$\mathrm{AIC} = -2L + 2k = -2(L - k) \qquad (2.40)$$

โดยที่ L คือ ตัวแบบล็อกภาวะน่าจะเป็น

k คือ จำนวนพารามิเตอร์และ

n คือ จำนวนของค่าสังเกตในตัวแบบ

เมื่อพิจารณาจากทั้งสองสมการจะพบว่า เมื่อจำนวนพารามิเตอร์ยิ่งเพิ่มขึ้นจะทำให้ $-2L$ มีค่าน้อยลง ดังนั้นจึงทำการปรับความเอนเอียงนี้โดยการเพิ่มส่วนของ $2k$ เข้าไปในสมการ นอกจากนี้ยังพบว่า n ส่งผลต่อตัวสถิติ $-2L$ เช่นเดียวกัน เมื่อพิจารณาจากสมการ (2.39) ซึ่งหารด้วย n ทำให้ได้ส่วนปรับความเอนเอียงต่อหน่วยสังเกตต่อตัวสถิติ $-2L$ สำหรับการเปรียบเทียบภายใต้เงื่อนไขที่เหมือนกัน

2.5 วิธีวัดประสิทธิภาพ

เค-โฟลด์ตรวจสอบไขว้ (K – Folds cross validation)

สำหรับการวัดประสิทธิภาพการพยากรณ์ของตัวแบบด้วยวิธีเค-โฟลด์ตรวจสอบไขว้ เป็นวิธีที่นิยมในการเรียนรู้ของเครื่องและในงานวิจัยต่าง ๆ เพราะง่ายในการทำความเข้าใจและทำให้มีความเอนเอียงต่ำ โดยแบ่งข้อมูลออกเป็น K ส่วนเท่า ๆ กัน (K โฟลด์) โดยการทำเค-โฟลด์ตรวจสอบไขว้ในงานวิจัยครั้งนี้จะใช้วิธีเลือกสุ่มแบบเที่ยงตรง ดังภาพที่ 2.5

ภาพที่ 2.5 กระบวนการเค-โฟลด์ตรวจสอบไขว้

จากภาพที่ 2.5 จะเลือกสุ่มข้อมูลออกเป็น K ชุดเท่า ๆ กัน ในที่นี้ $K=5$ กำหนดให้ K_l หมายถึง ข้อมูลชุดทดสอบที่ l โดยที่ $l=1,2,...,K$ การทดลองครั้งแรกข้อมูลชุดที่ 1 (โฟลด์ 1) เป็นข้อมูลชุดทดสอบ และข้อมูลชุดที่เหลือ (โฟลด์ 2 + โฟลด์ 3 + โฟลด์ 4 + โฟลด์ 5) เป็นข้อมูลชุดฝึกสอน ในการทดลองครั้งที่สองข้อมูลที่ 2 (โฟลด์ 2) เป็นข้อมูลชุดทดสอบและข้อมูลชุดที่เหลือเป็นข้อมูลชุดฝึกสอนทำจนกระทั่งข้อมูลทุกชุดได้ถูกนำมาเป็นข้อมูลชุดทดสอบซึ่งมีการทดลองทั้งหมด K ครั้ง ข้อดีของการเลือกสุ่มแบบความเที่ยงตรง K กลุ่ม คือ ข้อมูลทุกตัวจะถูกนำมาเป็นข้อมูลชุดฝึกสอนและข้อมูลชุดทดสอบ แต่ข้อเสียคือใช้เวลานานในการทดลองเนื่องจากต้องทดลองข้อมูลทั้งหมด K ครั้ง

2.6 งานวิจัยที่เกี่ยวข้อง

2.6.1 งานวิจัยที่เกี่ยวกับการเปรียบเทียบภาวะสารูปดีของตัวแบบ

สุภัทรา โนนคล้อ และคณะ (2561) ประยุกต์ใช้ตัวแบบการถดถอย P ในการวิเคราะห์ข้อมูลจำนวนนับทั่วไป เพื่อศึกษาลักษณะเสี่ยงทางคลินิกที่มีความสัมพันธ์กับการเกิดปอดบวมในผู้ป่วยโรคหลอดเลือดสมองที่เข้ารับการรักษาในโรงพยาบาลน่าน จังหวัดน่าน ประเทศไทย ทำการสุ่มตัวอย่างจำนวน 464 ราย จากเวชระเบียนผู้ป่วยที่เข้ารับการรักษาในช่วงวันที่ 1 มกราคม ปี พ.ศ. 2556 ถึง 31 ธันวาคม ปี พ.ศ. 2559 ในการวิเคราะห์หลายตัวแปรด้วยวิธีคัดเลือกตัวแปรพยากรณ์ออกแบบไปข้างหน้า ผลการวิจัยพบว่าลักษณะเสี่ยงทางคลินิกมีความสัมพันธ์กับการเกิดปอดบวมอย่างมีนัยสำคัญทางสถิติ

งานวิจัยที่ประยุกต์ใช้ตัวแบบในการวิเคราะห์ข้อมูลจำนวนนับที่มีปัญหาการกระจายต่ำกว่าเกณฑ์ Ridout, Demétrio และ Hinde (1998) ได้ทำการเปรียบเทียบตัวแบบการถดถอย P, NB, ZIP และ ZINB โดยใช้ข้อมูลจากการทดลองของ Marin และคณะ ปี ค.ศ. 1993 เกี่ยวกับการยิงแสงไปยังรากของแอปเปิลภายใต้ระดับความเข้มข้นของไซโตยานิน BAP ที่แตกต่างกันร่วมกับปัจจัยระยะเวลาของการยิงแสง ซึ่งเป็นข้อมูลที่มีการกระจายเกินเกณฑ์ เมื่อทำการเปรียบเทียบตัวแบบการถดถอยโดยพิจารณาจากเกณฑ์สารสนเทศของอะกะอิเกะและเกณฑ์สารสนเทศของเบส์ (BIC) พบว่าตัวแบบการถดถอยการถดถอย ZINB มีความเหมาะสมที่สุด

Thakali (2008) เปรียบเทียบประสิทธิภาพของตัวแบบการเกิดอุบัติเหตุบนทางหลวงในประเทศไทย ใช้ข้อมูลในอดีตย้อนหลัง 7 ปี ค.ศ. (2001 ถึง 2006) ซึ่งเป็นข้อมูลที่มีการกระจายเกินเกณฑ์ ตัวแปรตอบสนอง คือ ปริมาณจราจร ตัวแปรอิสระ คือ ร้อยละปริมาณรถบรรทุก ข้อมูลทางเรขาคณิต (จำนวนช่องจราจร ชนิดเกาะกลางไหล่ทาง) จำนวนทางเชื่อมต่อกิโลเมตร จำนวนทางแยกต่อกิโลเมตร จำนวนทางโค้งต่อกิโลเมตร ปริมาณน้ำฝน และเดือนเกิดเหตุ ในการวิเคราะห์ได้ใช้ตัว

แบบการถดถอย P และ NB เพื่อสร้างตัวแบบการถดถอยจำนวนการเกิดอุบัติเหตุจำนวนผู้เสียชีวิต จำนวนผู้บาดเจ็บและจำนวนอุบัติเหตุ โดยความเหมาะสมของตัวแบบทดสอบจากค่า Goodness of fit ผลการศึกษาพบว่าตัวแบบการถดถอย P เหมาะสมกับข้อมูลจำนวนอุบัติเหตุกับจำนวนผู้บาดเจ็บส่วนตัวแบบการถดถอย NB เหมาะสมกับข้อมูลจำนวนผู้เสียชีวิต นอกจากนี้พบว่าการมีไหล่ทางส่งผลให้มีจำนวนอุบัติเหตุลดลงมากที่สุด ช่วงเดือนเมษายนส่งผลให้มีจำนวนผู้เสียชีวิตและจำนวนผู้บาดเจ็บสูงสุดและชนิดเกาะกลางมีอิทธิพลต่อจำนวนอุบัติเหตุมากที่สุด

Annafari (2010) ได้ทำการวิเคราะห์ปัจจัยที่มีอิทธิพลต่อความต้องการในการใช้โทรศัพท์ของชาวสวีเดน ตัวแปรตอบสนอง คือ จำนวนโทรศัพท์ต่อคนหนึ่งคน มีค่าระหว่าง 0 ถึง 9 ตัวแปรอิสระ คือ ข้อมูลส่วนตัวข้องผู้เข้าร่วม เช่น อายุ เพศ การศึกษา อาชีพ รายได้ เป็นต้น ซึ่งมีแบบสอบถามที่ได้รับการตอบรับทั้งหมด 2,245 ฉบับ ผลการวิจัยพบว่า ตัวแบบการถดถอย GP เป็นตัวแบบที่เหมาะสมที่จะใช้ในการวิเคราะห์ข้อมูลนี้ โดยอ้างอิงจากเกณฑ์การวัดประสิทธิภาพของตัวแบบโดยใช้ Deviance, Pearson Chi-Square และ Log Likelihood

อดิเทพ ไชยวรรณ, วสันต์ บุญโฮ้ และ พิษณุ ทองขาว (2555) ต้องการหาปัจจัยเสี่ยงของการผลิตสินค้าบกพร่องในโรงงานอุตสาหกรรมผลิตชิ้นส่วนรถยนต์ โดยใช้ข้อมูลจำนวนสินค้าที่บกพร่องในโรงงานอุตสาหกรรมผลิตชิ้นส่วนรถยนต์แห่งหนึ่งในจังหวัดปทุมธานี ตั้งแต่เดือน มกราคม พ.ศ. 2555 ถึงเดือน กรกฎาคม พ.ศ. 2555 โดยตัวแปรตอบสนอง คือ จำนวนสินค้าบกพร่อง ตัวแปรอิสระ คือ คือ ขั้นตอนการผลิต เครื่องจักร พนักงานผู้รับผิดชอบ และประเภทสินค้าที่ผลิต โดยใช้การวิเคราะห์การถดถอย P เปรียบเทียบกับตัวแบบการถดถอย NB ในการเลือกตัวแบบที่เหมาะสมไปใช้ในการวิเคราะห์ข้อมูลด้วยวิธี Goodness of fit ผลการศึกษาพบว่า ตัวแบบการถดถอย NB เหมาะสมมากกว่าตัวแบบการถดถอย P ทุกกรณีของทุกสินค้า

Seyoum, Ndlovu และ Zewotir (2016) ได้ทำการศึกษาเพื่อเปรียบเทียบประสิทธิภาพของตัวแบบการถดถอย QP และ NB โดยใช้ข้อมูลจำนวนผู้ป่วยที่มีการเปลี่ยนแปลงจำนวนเซลล์ CD4 เริ่มต้นเนื่องจากการรักษาด้วยยาต้านไวรัสที่ให้กับผู้ใหญ่ที่ติดเชื้อเอชไอวีในเอธิโอเปียเหนือ - ตะวันตก (ภูมิภาค Amhara) มีจำนวนผู้ป่วยทั้งหมดที่เข้าร่วม 792 คน ตัวแปรตอบสนอง คือ จำนวนเซลล์ CD4 ที่เปลี่ยนไปต่อปริมาตรเนื้อเยื่อลูกบาศก์มิลลิเมตร ตัวแปรอิสระ คือ อายุ น้ำหนักของเซลล์ CD4 และเพศ มีเกณฑ์การเปรียบเทียบประสิทธิภาพ ได้แก่ ค่าดีเวียนซ์ (Deviance), AIC, BIC, ค่าล็อกภาวะน่าจะเป็น (Log-likelihood) และค่าเพียร์สันไคกำลังสอง (Pearson chi-square) ผลการศึกษาพบว่า ตัวแบบการถดถอย QP เหมาะสมกว่า NB

สำหรับงานวิจัยที่ประยุกต์ใช้ตัวแบบในการวิเคราะห์ข้อมูลจำนวนนับที่มีปัญหาการกระจายตัวและค่าศูนย์เฟ้อ Xie, He และ Goh (2001) ทำการศึกษาการใช้การแจกแจงสำหรับข้อมูลที่มีค่าศูนย์เฟ้อ โดยเปรียบเทียบตัวแบบการถดถอย P กับ ZIP จากข้อมูลจำลองด้วยวิธีมอนติคาร์โล

โดยกำหนดขนาดตัวอย่าง 3 ระดับ คือ ขนาดตัวอย่างเท่ากับ 10 20 และ 50 และค่าเฉลี่ย 2 ระดับ คือ 5 และ 10 ผลการศึกษาพบว่า ตัวแบบการถดถอย ZIP มีความเหมาะสมกว่า P เนื่องจากสามารถจัดการกับข้อมูลที่มีค่าศูนย์เฟ้อได้ดีกว่า

กษมะ นิจจันทร์พันศรี (2554) ได้ศึกษาเปรียบเทียบความเหมาะสมของ 3 ตัวแบบ คือ ตัวแบบการถดถอย ZIP, ZINB และ ZIGP ภายใต้ความน่าจะเป็นที่จะเกิดศูนย์และระดับความเบ้ของข้อมูลที่ไม่เป็นศูนย์หลายระดับ โดยใช้วิธีการจำลองข้อมูลด้วยวิธีมอนติคาร์โลและการทดสอบความเหมาะสมของตัวแบบด้วยสถิติ Deviance จากการศึกษาพบว่า ทุกตัวแบบมีความเหมาะสมกับข้อมูลที่มีความน่าจะเป็นที่จะเกิดศูนย์ในสัดส่วนที่สูง (0.85-0.95) ที่ทุกระดับของความเบ้ของข้อมูลที่ไม่เป็นศูนย์ เมื่อความเบ้ของข้อมูลที่ไม่เป็นศูนย์มีค่าน้อยหรือไม่เบ้ พิจารณาจากเกณฑ์ AIC พบว่าตัวแบบการถดถอย ZINB มีความเหมาะสมที่สุด แต่เมื่อพิจารณาจากเกณฑ์ BIC พบว่าส่วนใหญ่ตัวแบบที่เหมาะสม คือ ตัวแบบการถดถอย ZIGP

Ayati และ Abbasi (2014) ได้ดำเนินงานวิจัยที่มีวัตถุประสงค์เพื่อการเปรียบเทียบประสิทธิภาพการพยากรณ์การชนของรถยนต์ โดยใช้ข้อมูลอุบัติเหตุรถชนบนทางหลวง ในเมืองหลวง Mashhad Urban ประเทศอิหร่าน จำนวน 156 เขต ระหว่างปี ค.ศ. 2006 ถึง 2009 ตัวแปรตอบสนอง คือ จำนวนครั้งของการชน ตัวแปรอิสระ คือ การคล่องตัวของการจราจร ระดับการติดขัดของจราจรและความเร็วของรถ ตัวแปรทางเรขาคณิตของถนน เช่น จำนวนเลน โครงสร้างเส้นโค้งและแนวนอนของถนน เป็นต้น ตัวแบบการถดถอยที่ใช้ ได้แก่ P, NB, ZIP และ ZINB เกณฑ์การเปรียบเทียบประสิทธิภาพ คือ ค่า AIC และ BIC พบว่า ตัวแบบการถดถอย ZINB เป็นตัวแบบที่ดีที่สุดและเหมาะสมที่สุดทั้งข้อมูลอุบัติเหตุที่ไม่มีผู้ได้รับบาดเจ็บและอุบัติเหตุที่มีผู้ได้รับบาดเจ็บ

Yang และคณะ (2017) ทำการเปรียบเทียบประสิทธิภาพของตัวแบบ ได้แก่ ตัวแบบการถดถอยกำลังสองน้อยที่สุด (Ordinary least-squares regression: LST) ตัวแบบถดถอย P, NB, ZIP, ZINB, ตัวแบบการถดถอยปัวซงค่าศูนย์ผันแปร (Zero-altered Poisson regression: ZAP) และตัวแบบการถดถอยทวินามเชิงลบค่าศูนย์ผันแปร (Zero-altered Negative binomial regression: ZANB) เพื่อเปรียบเทียบประสิทธิภาพของแต่ละตัวแบบภายใต้เงื่อนไขที่แตกต่างกันของสัดส่วนค่าศูนย์และการกระจายเกินเกณฑ์ของข้อมูล โดยผู้วิจัยได้ทำการจำลองข้อมูลแบบผสมระหว่างความน่าจะเป็นที่จะเกิดศูนย์ของค่าศูนย์ (ร้อยละ 20 40 60 และ 80) และพารามิเตอร์การกระจายที่แตกต่างกัน (10 50 และ 100) จากการแจกแจง NB รวมถึงการใช้ข้อมูลจริงเกี่ยวกับการสำรวจด้านสุขภาพจาก Behavioral Risk Factor Surveillance System จากนั้นทำการวิเคราะห์ถึงการเกิดค่าศูนย์เฟ้อ และสำรวจความสัมพันธ์ระหว่างกิจกรรมทางกายภาพและสุขภาพที่ส่งผลต่อคุณภาพของชีวิต โดยใช้เกณฑ์สารสนเทศ AIC และสถิติทดสอบวอง (Vuong test) ในการประเมิน

ตัวแบบข้างต้น ผลการศึกษาจากข้อมูลจำลอง พบว่า ZINB และ ZANB มีประสิทธิภาพที่ดีกว่าตัวแบบอื่น ๆ สำหรับผลการศึกษาจากข้อมูลจริงพบว่า ZANB เป็นตัวแบบที่มีประสิทธิภาพดีที่สุด

นวพรรณ เชื้ออ่ำ, บุญอ้อม โฉมที และ อภิญญา หิรัญวงษ์ (2561) มีวัตถุประสงค์เพื่อเปรียบเทียบความเหมาะสมของตัวแบบการถดถอย QP และ ZINB สำหรับตัวแปรตอบสนองที่มีลักษณะเป็นข้อมูลจำนวนนับและมีปัญหาการกระจายเกินเกณฑ์ ข้อมูลจริงที่ใช้ คือ ข้อมูลการเกิดอุบัติเหตุบนท้องถนนที่เป็นถนนทางหลวงโดยเก็บรวบรวมข้อมูลจากสำนักอำนวยความปลอดภัยกรมทางหลวง ตั้งแต่เดือนมกราคมถึงเดือนธันวาคม ปี พ.ศ. 2559 โดยการศึกษากับข้อมูลจริงที่ตัวแปรตอบสนอง คือ จำนวนผู้บาดเจ็บในการเกิดอุบัติเหตุแต่ละครั้ง ซึ่งแบ่งตามลักษณะของข้อมูลเป็น 3 กรณี คือ ตัวอย่างขนาดเล็ก (n =17) กลาง (n =32) และใหญ่ (n =56) และตัวแปรอิสระ คือ ปริมาณการจราจรต่อวัน โครงสร้างของถนน โดยตัวอย่างขนาดเล็กมีค่าความน่าจะเป็นที่จะเกิดศูนย์ของตัวแปรตอบสนองเท่ากับ 0.25 และสำหรับตัวอย่างขนาดกลางและใหญ่มีค่าความน่าจะเป็นที่จะเกิดศูนย์ของตัวแปรตอบสนองเท่ากับ 0.50 ตัวแบบการถดถอย QP มีความเหมาะสมมากกว่า ZINB เกือบทุกกรณี

Sim, Gupta และ Ong (2018) ทำการศึกษาเพื่อเปรียบเทียบความเหมาะสมของตัวแบบการถดถอย ZIP, ZIGP และ CMP ซึ่งใช้กับข้อมูลจำลองด้วยวิธีมอนติคาร์โล ที่ขนาดตัวอย่างเท่ากับ 100, 500 และ 1,000 โดยจำลองข้อมูลกรณีการกระจายเกินเกณฑ์และการกระจายต่ำกว่าเกณฑ์ มีการทำซ้ำ 1,000 รอบ และข้อมูลจริงที่ใช้ คือ ข้อมูลจำนวนรากที่เกิดจากหน่อที่ขยายพันธุ์ขนาดเล็ก 270 ยอดของแอปเปิลที่ได้รับการปรับปรุงสายพันธุ์ ซึ่งเป็นข้อมูลของ Ridout และคณะ ปี ค.ศ. 2001 โดยทำการเก็บรวบรวมข้อมูลในช่วงระยะเวลาเติบโตของรากที่มีการบำรุงรักษาสภาพที่เหมือนกันแต่มียอดอยู่ ซึ่งเพาะเลี้ยงบนอาหารเลี้ยงเชื้อที่มีความเข้มข้นต่างกันของไซโตไคนิน BAP ภายใต้ระยะเวลาการรับแสง 8 หรือ 16 ชั่วโมง ตัวแปรตอบสนอง คือ จำนวนราก ตัวแปรอิสระ คือ ช่วงแสง จำนวน 140 และ 130 ช็อต โดยแบ่งเป็นกลุ่มที่ 1 คือ 8 ชั่วโมง และกลุ่มที่ 2 คือ 16 ชั่วโมง โดยใช้เกณฑ์เปรียบเทียบประสิทธิภาพ AIC ผลการศึกษาพบว่า ตัวแบบการถดถอย CMP ให้ค่าเกณฑ์ประสิทธิภาพที่ดีที่สุด

Alqawba และ Diawara (2020) ได้วิเคราะห์ข้อมูลจำนวนครั้งของการเกิดพายุทรายจำแนกรายเดือนที่บันทึกโดยสถานีสนามบิน AQI ในจังหวัดตะวันออกซาอุดีอาระเบีย มีจำนวนพายุทรายที่รุนแรงจำนวน 348 ครั้งต่อเดือน ตั้งแต่เดือนมกราคม ค.ศ. 1978 ถึงเดือนธันวาคม ค.ศ. 2013 โดยใช้ตัวแบบการถดถอย ZIP, ZINB และ ZICMP และเกณฑ์เปรียบเทียบประสิทธิภาพ AIC BIC และรากของค่าคลาดเคลื่อนการพยากรณ์กำลังสองเฉลี่ย ผลการศึกษาพบว่า ตัวแบบที่ให้ค่าเกณฑ์การเปรียบเทียบประสิทธิภาพที่ดีที่สุด คือ ตัวแบบการถดถอย ZINB

Prasetijio และคณะ (2020) มีวัตถุประสงค์ในการเปรียบเทียบประสิทธิภาพการพยากรณ์ของตัวแบบการถดถอย P, NB, ZIP และ ZINB โดยใช้ข้อมูลอุบัติเหตุทางถนนของเส้นทาง F0050 Kluang-Air Hitam-Batu Pahat ในประเทศมาเลเซีย ตั้งแต่กิโลเมตรที่ 0 ถึง กิโลเมตรที่ 58 ระหว่างปี ค.ศ. 2010 ถึง ปี ค.ศ. 2014 โดยกองบังคับการตำรวจจราจร สำนักงานตำรวจแห่งชาติมาเลเซีย เป็นหน่วยงานที่จัดการเพื่อรวบรวมข้อมูล ตัวแปรตอบสนอง คือ ผู้เสียชีวิต, ผู้ได้รับบาดเจ็บสาหัสและผู้ได้รับบาดเจ็บเล็กน้อย ตัวแปรอิสระ คือ ความเสียหายจากอุบัติเหตุ, ปริมาณการใช้ข้อมูลเฉลี่ยรายปี และความเร็วของยานพาหนะ ในการศึกษาครั้งนี้ได้เปรียบเทียบประสิทธิภาพของตัวแบบจำนวนอุบัติเหตุที่มีส่งผลให้มีการเสียชีวิต จากการทดสอบการกระจายพบว่าข้อมูลมีปัญหาการกระจายตัวเกินเกณฑ์ โดยที่ ZINB ให้ค่า AIC ต่ำสุด และพบว่าตัวแบบการถดถอย ZINB และ NB มีประสิทธิภาพการพยากรณ์ดีกว่า P และ ZIP

2.6.2 งานวิจัยที่เกี่ยวกับการเปรียบเทียบประสิทธิภาพการพยากรณ์ของตัวแบบ

สำหรับงานวิจัยที่ประยุกต์ใช้ตัวแบบในการวิเคราะห์ข้อมูลจำนวนนับที่มีปัญหาการกระจายเกินเกณฑ์ Lord, Guikema และ Geedipally (2008) ได้ใช้ตัวแบบการถดถอย CMP และ NB เพื่อสร้างสมการพยากรณ์การชนของรถยนต์ ตัวแปรตอบสนอง คือ จำนวนครั้งของการชน และตัวแปรอิสระ คือ ปริมาณการจราจร ข้อมูลที่ใช้ในการศึกษา มีจำนวน 2 ชุด ได้แก่ ข้อมูลบริเวณแยกสัญญาณจราจรสี่เลนที่ตั้งอยู่ในรัฐ Toronto Ontario ปี ค.ศ. 1995 มี 868 เหตุการณ์ที่ถูกบันทึก จำนวนการชนทั้งหมด 52,273 ครั้ง และข้อมูลย้อนหลัง 5 ปี บริเวณถนนสี่เลนในเขตชนบทของรัฐ Texas ซึ่งบันทึกโดย Texas Department of Public Safety (DPS) และ Texas Depart ment of Transportation (TxDOT) มี 3,220 เหตุการณ์ที่ถูกบันทึก จำนวนการชนทั้งหมด 79,455 ครั้ง งานวิจัยมีวัตถุประสงค์เพื่อเปรียบเทียบประสิทธิภาพการพยากรณ์ของตัวแบบการถดถอย CMP และ NB โดยใช้เกณฑ์การวัดประสิทธิภาพ ได้แก่ ค่า AIC และค่าเบี่ยงเบนสัมบูรณ์เฉลี่ย (Mean Absolute deviation: MAD) ผลการศึกษาพบว่า ตัวแบบการถดถอย CMP และ NB มีประสิทธิภาพใกล้เคียงกัน

Pott และ Elith (2006) ทำการศึกษาและเปรียบเทียบประสิทธิภาพการพยากรณ์ความอุดมสมบูรณ์ของพืชสายพันธุ์ Leionema ralstonii โดยใช้ภาพถ่ายทางอากาศ จากบริเวณที่พืชเติบโตตามโขดหิน ตัวแปรตอบสนอง คือ จำนวนโขดหินที่พบ Leionema ralstonii ในรัสมี 400 เมตร และพบว่าไม่มีพืชที่เติบโตตามโขดหินคิดเป็นร้อยละ 37 (ความน่าจะเป็นที่จะเกิดศูนย์เท่ากับ 0.37) และตัวแปรอิสระ ได้แก่ ลอกธรรมชาติของพื้นที่โผล่ขึ้นมา (m^2) และปริมาณน้ำฝนรายปีเฉลี่ยบริเวณและ โขดหินจากโขดหินที่ทำการถ่ายเก็บรวบรวมมา 1,100 โขดหิน โดยใช้ตัวแบบการถดถอย P, QP, NB, ZIP และ เฮอร์เดิล (Hurdle) เกณฑ์การเปรียบเทียบ ได้แก่ RMSE และ R^2 ผลลัพธ์จากงานวิจัยพบว่า ตัวแบบเฮอร์เดิลเป็นตัวแบบที่มีปะสิทธิภาพดีที่สุด รองลงมา คือ ตัวแบบการถดถอย P และ QP ตามลำดับ

สำหรับงานวิจัยที่ประยุกต์ใช้ตัวแบบทางสถิติเปรียบเทียบกับการเรียนรู้ของเครื่องในการวิเคราะห์ข้อมูลจำนวนนับ Chang และ Chen (2005) ศึกษาข้อมูลอุบัติเหตุระหว่างปี ค.ศ. 2001-2002 ในไต้หวัน บนทางด่วน National Freeway 1 ซึ่งมีความยาว 373 กิโลเมตร บันทึกโดย Ministry of Transportation and Communications มีอุบัติเหตุเกิดขึ้นทั้งหมด 2,968 เหตุการณ์ ตัวแปรตอบสนอง คือ จำนวนครั้งการเกิดอุบัติเหตุ ตัวแปรอิสระ ได้แก่ สถานที่เกิดเหตุ, ข้อมูลส่วนบุคคลของคนขับ, สภาพความเสียหายทางร่างกายที่ได้รับบาดเจ็บ และข้อมูลการจราจร รวมไปถึงข้อมูลลักษณะโครงสร้างของถนน เช่น จำนวนเลน ความโค้งแนวนอนของถนน แนวตั้งของถนน เป็นต้น และข้อมูลสภาพอากาศนำมาจากรายงานประจำปีของภูมิอากาศวิทยา เช่น อุณหภูมิ ความชื้น ปริมาณน้ำฝน ความเร็วลมและความหน้าแน่นของเมฆ เป็นต้น โดยเปรียบเทียบความเหมาะสมของตัวแบบด้วยวิธีภาวะสารูปดีและเปรียบเทียบประสิทธิภาพการพยากรณ์ของตัวแบบด้วยร้อยละความถูกต้อง (Accuracy percentage) ตัวแบบการถดถอยที่ใช้ ได้แก่ P และ NB และเทคนิค Classification and Regression Tree (CART) พบว่าตัวแบบการถดถอย NB และเทคนิค CART มีประสิทธิภาพใกล้เคียงกันและมีความเหมาะสมกว่าตัวแบบการถดถอย P โดยปัจจัยที่มีผลต่อจำนวนครั้งการเกิดอุบัติเหตุบนทางด่วน คือ ปริมาณการจราจรต่อวันและปริมาณน้ำฝน และตัวแบบการถดถอย NB ให้ค่าความถูกต้องร้อยละ 52.3 ส่วน CART ให้ค่าความถูกต้องร้อยละ 52.6

Ma และ Yuan (2018) ทำการเปรียบเทียบประสิทธิภาพการพยากรณ์การเกิดอุบัติเหตุบนท้องถนนแห่งหนึ่งในประเทศจีน มีขนาดตัวอย่างเท่ากับ 200 และมีตัวแปรอิสระเชิงปริมาณทั้งหมด ระยะช่วงผิวของถนน, จำนวนรถรายวันเฉลี่ยต่อปี, คะแนนสภาพการจราจรและร้อยละความหนาแน่นของรถ ตัวแปรอิสระเชิงคุณภาพ ได้แก่ ถนน 2 เลนหรือน้อยกว่า 2 เลน, ถนนที่มากกว่า 2 เลน, สายทางของถนนชนบทหรือถนนในเมือง โดยนักวิจัยใช้ตัวแปรอิสระ 6 ตัวแปรในวิเคราะห์และพยากรณ์ มีเกณฑ์วัดประสิทธิภาพ คือ AIC และ BIC สำหรับตัวแบบการถดถอยและ RSS สำหรับเปรียบเทียบตัวแบบการถดถอยและเทคนิค RF พบว่าตัวแบบการถดถอย ZINB มีประสิทธิภาพดีที่สุด รองลงมาคือ NB, RF และ P

Sadler และคณะ (2018) ทำการเปรียบเทียบประสิทธิภาพในการพยากรณ์จำนวนครั้งที่เกิดน้ำท่วมในเมือง Norfolk รัฐ Virginia ประเทศสหรัฐอเมริกา ตั้งแต่เดือนกันยายน 2010 ถึงตุลาคม 2016 เป็นข้อมูลที่ได้มาจากการได้รับแจ้งจากประชาชนในพื้นที่ มีอุทกภัยที่เกิดขึ้นในบันทึก 45 ครั้ง ตัวแปรตอบสนองคือ จำนวนครั้งการเกิดน้ำท่วม ตัวแปรอิสระมีทั้งหมด 19 ตัวแปร เช่น ปริมาณน้ำฝน น้ำขึ้นน้ำลง ระดับน้ำใต้ดินและสภาพลม เป็นต้น โดยใช้ตัวแบบการถดถอย P และเทคนิค RF เกณฑ์วัดประสิทธิภาพ คือ ค่ารากของค่าคลาดเคลื่อนกำลังสอง และค่าคลาดเคลื่อนสมบูรณ์ (Mean absolute error) ผลการศึกษาสรุปได้ว่าเทคนิค RF มีประสิทธิภาพการพยากรณ์ดีกว่าตัวแบบการถดถอย P ในการพยากรณ์จำนวนครั้งของการเกิดน้ำท่วม

บทที่ 3
วิธีการวิจัย

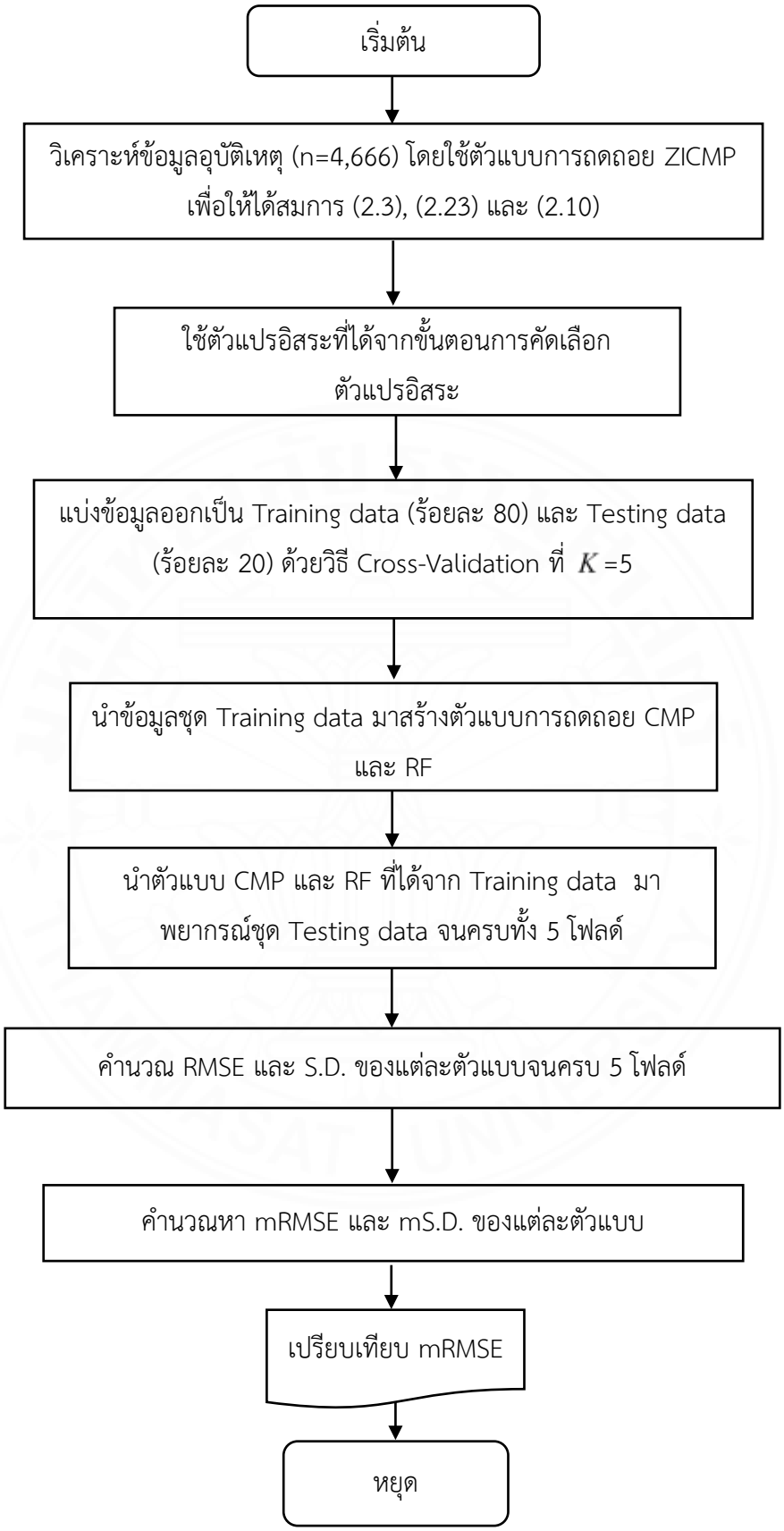
งานวิจัยชิ้นนี้มีวัตถุประสงค์เพื่อเปรียบเทียบประสิทธิภาพการพยากรณ์ของตัวแบบที่นิยมใช้วิเคราะห์ข้อมูลจำนวนนับต่าง ๆ โดยถูกแบ่งเป็นสองสถานการณ์ คือ กรณีจำนวนนับที่มีลักษณะการกระจายต่ำกว่าเกณฑ์และค่าศูนย์เฟ้อ ได้แก่ ตัวแบบการถดถอย CMP และเทคนิค RF อีกสถานการณ์หนึ่ง คือ กรณีจำนวนนับที่มีลักษณะการกระจายเกินเกณฑ์และค่าศูนย์เฟ้อ ได้แก่ ตัวแบบการถดถอย QP, CMP, ZIP, ZINB และเทคนิค RF การเปรียบเทียบความถูกต้องในการพยากรณ์ของตัวแบบใช้วิธีเค-โฟลด์ตรวจสอบไขว้ โดยทำการสุ่มแบ่งข้อมูลเป็นออก 5 ส่วนเท่า ๆ กัน ข้อมูล 1 ส่วนเป็นข้อมูลชุดทดสอบและข้อมูล 4 ส่วนที่เหลือเป็นข้อมูลชุดฝึกสอน ซึ่งข้อมูลในแต่ละชุดจะประกอบไปด้วยชุดทดสอบและชุดฝึกสอนที่ไม่ซ้ำกันทั้งหมด 5 ชุด สำหรับใช้สร้างสมการถดถอยของตัวแบบและพยากรณ์ และเกณฑ์ในการเปรียบเทียบประสิทธิภาพของตัวแบบ ได้แก่ ค่าเฉลี่ยของรากของค่าคลาดเคลื่อนกำลังสองเฉลี่ย (mRMSE) และค่าเฉลี่ยของค่าเบี่ยงเบนมาตรฐานของรากของค่าคลาดเคลื่อนกำลังสองเฉลี่ย (mS.D.) โดยมีรายละเอียดดังหัวข้อต่อไปนี้

3.1 การวิเคราะห์ข้อมูลกรณีข้อมูลมีการกระจายต่ำกว่าเกณฑ์และค่าศูนย์เฟ้อ

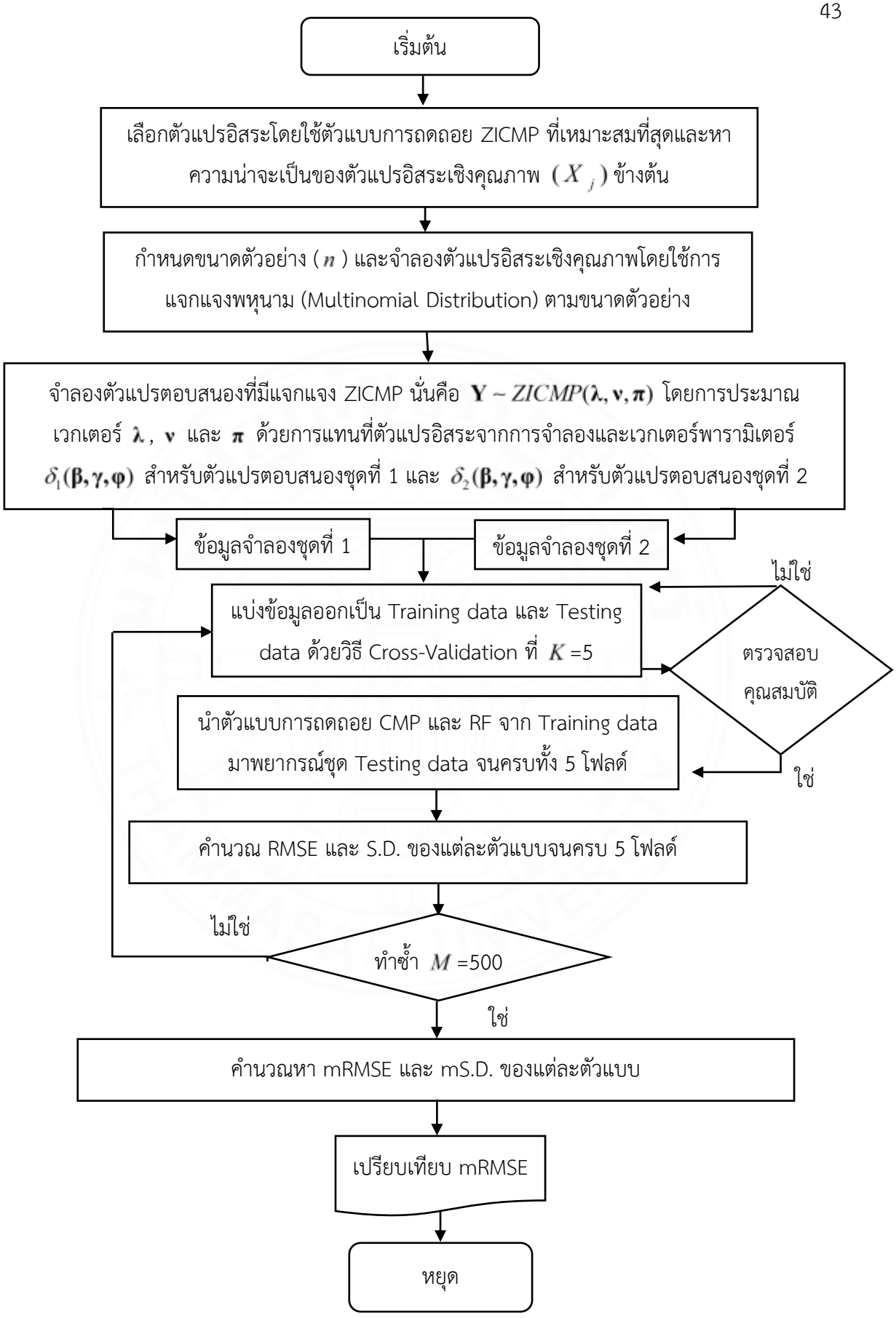
ขั้นตอนการวิจัยกรณีการกระจายต่ำกว่าเกณฑ์และค่าศูนย์เฟ้อ ดังภาพที่ 3.1 และ 3.2

3.1.1 การคัดเลือกตัวแปรอิสระโดยใช้ตัวแบบการถดถอย ZICMP

ทำการวิเคราะห์ข้อมูลจริงโดยใช้ตัวแบบการถดถอย ZICMP เพื่อให้ได้สมการถดถอยสำหรับการประมาณค่าเฉลี่ย (λ_i), การประมาณค่าการกระจาย (ν_i) และความน่าจะเป็นที่จะเกิดศูนย์ (π_i) สมการ (2.3), (2.23) และ (2.10) ตามลำดับ โดยจะเลือกตัวแบบที่มีความเหมาะสมที่สุด ซึ่งจะพิจารณาตัวแปรอิสระที่มีนัยสำคัญและให้ค่า AIC ต่ำที่สุด สมการ (2.40) และนำตัวแปรอิสระดังกล่าวมาสร้างตัวแบบการถดถอย CMP และเทคนิค RF เพื่อเปรียบเทียบประสิทธิภาพการพยากรณ์โดยใช้ข้อมูลจริงและข้อมูลจำลอง จากตารางที่ 3.1 พบว่า ประเภทสายทาง, ลักษณะสายทางและช่วงเวลาเป็นปัจจัยที่มีผลต่อการประมาณค่าพารามิเตอร์ค่าเฉลี่ย และพื้นผิวถนน, ลักษณะสายทางและช่วงเวลาเป็นปัจจัยที่มีผลต่อการประมาณค่าความน่าจะเป็นที่จะเกิดศูนย์ ในขณะที่สภาพอากาศไม่มีผลต่อการประมาณค่าพารามิเตอร์ โดยมีค่า AIC เท่ากับ 4,429.632 และ Log–likelihood เท่ากับ -2,195.816 เวกเตอร์ของพารามิเตอร์ $\delta_1(\boldsymbol{\beta},\boldsymbol{\gamma},\boldsymbol{\varphi})$ ได้จากสดมภ์ “Estimate” ของตารางที่ 3.1 จะถูกนำไปใช้ในการจำลองตัวแปรตอบสนองชุดที่ 1 สำหรับตัวแปรตอบสนองชุดที่ 2 จะใช้เวกเตอร์ของพารามิเตอร์ $\delta_2(\boldsymbol{\beta},\boldsymbol{\gamma},\boldsymbol{\varphi}) = 1.5 \times \delta_1(\boldsymbol{\beta},\boldsymbol{\gamma},\boldsymbol{\varphi})$ ในการจำลอง



ภาพที่ 3.1 ขั้นตอนการวิเคราะห์ข้อมูลจริงกรณีการกระจายต่ำกว่าเกณฑ์



ภาพที่ 3.2 ขั้นตอนการวิเคราะห์ข้อมูลจำลองกรณีการกระจายต่ำกว่าเกณฑ์

ตารางที่ 3.1 ตารางวิเคราะห์ข้อมูลจำนวนผู้เสียชีวิตจากอุบัติเหตุทางถนนทั่วประเทศไทย เดือนเมษายน ปี พ.ศ. 2558 จากตัวแบบการถดถอย ZICMP (n=4,666)

Coefficients	Estimate	S.E.	Z value	P-value
พารามิเตอร์ค่าเฉลี่ย				
ค่าคงที่	0.9345	0.3917	2.3859	0.0170
ประเภทสายทาง (1:กรมทางหลวง)				
2 ถนนกรมทางหลวงชนบท	-0.6681	0.2101	-3.1795	0.0015
3 ถนนเทศบาล	-2.0443	0.2838	-7.2025	0.0001
4 ถนนใน อบต. / หมู่บ้าน / อื่น ๆ	-2.0889	0.2301	-9.0794	0.0001
ลักษณะสายทาง (1: ทางตรง)				
2 ทางโค้ง	0.9000	0.3011	2.9887	0.0028
3 ทางแยก	-0.1771	0.3087	-0.5737	0.5662
4 ทางคนข้าม / มีสิ่งกีดขวาง / อื่น ๆ	-1.3811	0.3306	-4.1781	0.0001
ช่วงเวลา (1: กลางวัน)				
2 กลางคืนมีแสงไฟฟ้า	0.1440	0.2519	0.5716	0.5676
3 กลางคืนไม่มีแสงไฟฟ้า	-0.3740	0.2662	-1.4050	0.1600
4 อื่น ๆ	1.4457	0.2777	5.2059	0.0001
พารามิเตอร์การกระจาย				
ค่าคงที่	1.6813	0.1013	15.9737	0.0001
พารามิเตอร์ความน่าจะเป็นที่จะเกิดศูนย์				
ค่าคงที่	0.7221	0.1991	3.6269	0.0003
พื้นผิวถนน (1: แห้ง)				
2 เปียก, หลุมบ่อ และ อื่น ๆ	0.4091	0.1571	2.6038	0.0092
ลักษณะสายทาง (1: ทางตรง)				
2 ทางโค้ง	0.5717	0.1990	2.8730	0.0041
3 ทางแยก	-0.3107	0.2666	-1.1653	0.2439
4 ทางคนข้าม / มีสิ่งกีดขวาง / อื่น ๆ	-7.1086	6.1203	-1.1615	0.2454
ช่วงเวลา (1: กลางวัน)				
2 กลางคืนมีแสงไฟฟ้า	0.0051	0.1856	0.0272	0.9783
3 กลางคืนไม่มีแสงไฟฟ้า	-1.0289	0.2527	-4.0708	0.0001
4 อื่น ๆ	-0.0535	0.2972	-0.1800	0.8572

3.1.2 การวิเคราะห์ข้อมูลจริง

1. ทำการแบ่งข้อมูลออกเป็น 2 ชุด คือ ชุดฝึกสอนและชุดทดสอบ ด้วยวิธีเค-โฟลด์ตรวจสอบไขว้ ที่ K=5 ตามรายละเอียดดังหัวข้อ 2.5 หน้า 30 ข้อมูลอุบัติเหตุทางถนนที่ส่งผลให้มีผู้เสียชีวิตทั่วประเทศไทย เดือนเมษายน ปี พ.ศ. 2558 มีขนาดตัวอย่างทั้งหมด 4,666 เหตุการณ์ แบ่งเป็นข้อมูลชุดฝึกสอนจำนวน 3,733 เหตุการณ์ (ร้อยละ 80) และข้อมูลชุดทดสอบจำนวน 933 เหตุการณ์ (ร้อยละ 20)

2. ใช้ตัวแปรอิสระที่ได้จากหัวข้อ 3.1.1 เพื่อเตรียมสำหรับการสร้างตัวแบบ

3. นำข้อมูลชุดฝึกสอนมาสร้างตัวแบบการถดถอย CMP ดังสมการ (2.23) และ (2.28) และเทคนิค RF (หัวข้อ 2.1.7.2 หน้า 20) และนำตัวแบบที่ได้ไปพยากรณ์ข้อมูลชุดทดสอบจนครบทั้ง 5 โฟลด์

4. คำนวณ mRMSE สมการ (1.2) และค่า mS.D. สมการ (1.5) โดยตัวแบบที่เหมาะสมที่สุดคือตัวแบบที่ให้ค่า mRMSE ต่ำสุด

3.1.3 การวิเคราะห์ข้อมูลจำลอง

1. กำหนดขนาดตัวอย่าง (n) เท่ากับ 250 500 1,000 3,000 และ 5,000 และจำนวนรอบในการจำลอง (M) เท่ากับ 500

2. จำลองตัวแปรอิสระเชิงคุณภาพโดยใช้การแจกแจงพหุนาม (Multinomial distribution) ที่มีการกำหนดจำนวนกลุ่มและความน่าจะเป็นจากข้อมูลอุบัติเหตุทางถนนที่ส่งผลให้มีผู้เสียชีวิตทั่วประเทศไทย ตารางที่ 3.2 แสดงสัดส่วนของลักษณะที่สนใจในแต่ละตัวแปรอิสระที่มีผลต่อการประมาณค่าพารามิเตอร์ในข้อมูลชุดนี้

3. จำลองตัวแปรตอบสนองที่มีแจกแจง ZICMP นั่นคือ $\mathbf{Y} \sim ZICMP(\boldsymbol{\lambda}, \boldsymbol{\nu}, \boldsymbol{\pi})$ โดยการประมาณเวกเตอร์ค่าเฉลี่ย ($\boldsymbol{\lambda}$), ค่าการกระจาย ($\boldsymbol{\nu}$) และความน่าจะเป็นที่จะเกิดศูนย์ ($\boldsymbol{\pi}$) ด้วยการแทนที่ตัวแปรอิสระจากการจำลองและเวกเตอร์พารามิเตอร์ $\delta_1(\boldsymbol{\beta}, \boldsymbol{\gamma}, \boldsymbol{\varphi})$ สำหรับตัวแปรตอบสนองชุดที่ 1 ลงในสมการ (2.3), (2.23) และ (2.10) ตามลำดับ

4. ทำการแบ่งข้อมูลออกเป็น 2 ส่วน คือ ชุดฝึกสอนคิดเป็นร้อยละ 80 และชุดทดสอบคิดเป็นร้อยละ 20 ของขนาดตัวอย่างด้วยวิธีเค-โฟลด์ตรวจสอบไขว้ ที่ K=5 ตามรายละเอียดดังหัวข้อ 2.5 หน้า 30

5. ตรวจสอบปัญหาการกระจายของข้อมูลด้วย Dispersion test สมการที่ (2.38) หากชุดฝึกสอนชุดใดไม่เป็นไปตามเงื่อนไขจะไม่นำมาพิจารณา

6. นำข้อมูลชุดฝึกสอนมาสร้างตัวแบบการถดถอย CMP และเทคนิค RF และนำตัวแบบที่ได้ไปพยากรณ์ข้อมูลชุดทดสอบจนครบทั้ง 5 โฟลด์ เก็บค่า RMSE ในแต่ละโฟลด์

7. ทำซ้ำขั้นตอนที่ 4 ถึง 6 จนครบ 500 รอบ

8. คำนวณ mRMSE โดยใช้สมการ (1.3) และค่า mS.D. โดยใช้สมการ (1.6) และตัวแบบที่เหมาะสมที่สุดคือตัวแบบที่ให้ค่า mRMSE ต่ำสุด

9. ทำซ้ำในขั้นตอนที่ 3 ถึง 8 สำหรับตัวแปรตอบสนองชุดที่ 2 ($\delta_2(\boldsymbol{\beta}, \boldsymbol{\gamma}, \boldsymbol{\varphi})$)

ตารางที่ 3.2 สัดส่วนของลักษณะที่สนใจในแต่ละตัวแปรอิสระที่มีผลต่อการประมาณค่าพารามิเตอร์

ชื่อตัวแปร	**คำอธิบายตัวแปร**	**จำนวน (ร้อยละ)**
ROADTYPE_ID	ประเภทสายทาง :	
	1 ถนนกรมทางหลวง	1,864 (39.95)
	2 ถนนกรมทางหลวงชนบท	536 (11.49)
	3 ถนนเทศบาล	575 (12.32)
	4 ถนนใน อบต. / หมู่บ้าน / อื่น ๆ	1,691 (36.24)
ROADSKIN_ID	พื้นผิวถนน :	
	1 แห้ง	3,892 (83.41)
	2 เปียก, หลุมบ่อ และ อื่น ๆ	774 (16.59)
ACDPOINT_ID	ลักษณะสายทาง :	
	1 ทางตรง	2,841 (60.89)
	2 ทางโค้ง	835 (17.89)
	3 ทางแยก	710 (15.22)
	4 ทางคนข้าม / มีสิ่งกีดขวาง / อื่น ๆ	280 (6.00)
LIGHT_ID	ช่วงเวลา :	
	1 กลางวัน,	2,554 (54.73)
	2 กลางคืนมีแสงไฟฟ้า	1,055 (22.61)
	3 กลางคืนไม่มีแสงไฟฟ้า	849 (18.20)
	4 อื่น ๆ	208 (4.46)

3.2 การวิเคราะห์ข้อมูลกรณีข้อมูลมีการกระจายเกินเกณฑ์และค่าศูนย์เฟ้อ

ในการวิเคราะห์ข้อมูลการเรียกร้องสินไหมทดแทนของบริษัทประกันภัยแห่งหนึ่ง ปี พ.ศ. 2560 ซึ่งเป็นข้อมูลที่มีการกระจายเกินเกณฑ์และค่าศูนย์เฟ้อจะมีกระบวนการวิจัยเหมือนกับข้อมูลกรณีที่มีการกระจายต่ำกว่าเกณฑ์และค่าศูนย์เฟ้อที่ได้อธิบายไว้แล้วในหัวข้อ 3.1 และขั้นตอนการวิจัยแสดงดังภาพที่ 3.3 และ 3.4 โดยมีขั้นตอนเพิ่มเติม 2 ส่วน ได้แก่ จำนวนตัวแบบที่ใช้เปรียบเทียบประสิทธิภาพทั้งหมด 5 ตัวแบบ ประกอบด้วย ตัวแบบการถดถอย QP, CMP, ZIP, ZINB และเทคนิค RF และการจำลองตัวแปรเชิงปริมาณ

3.2.1 การจำลองตัวแปรเชิงปริมาณ

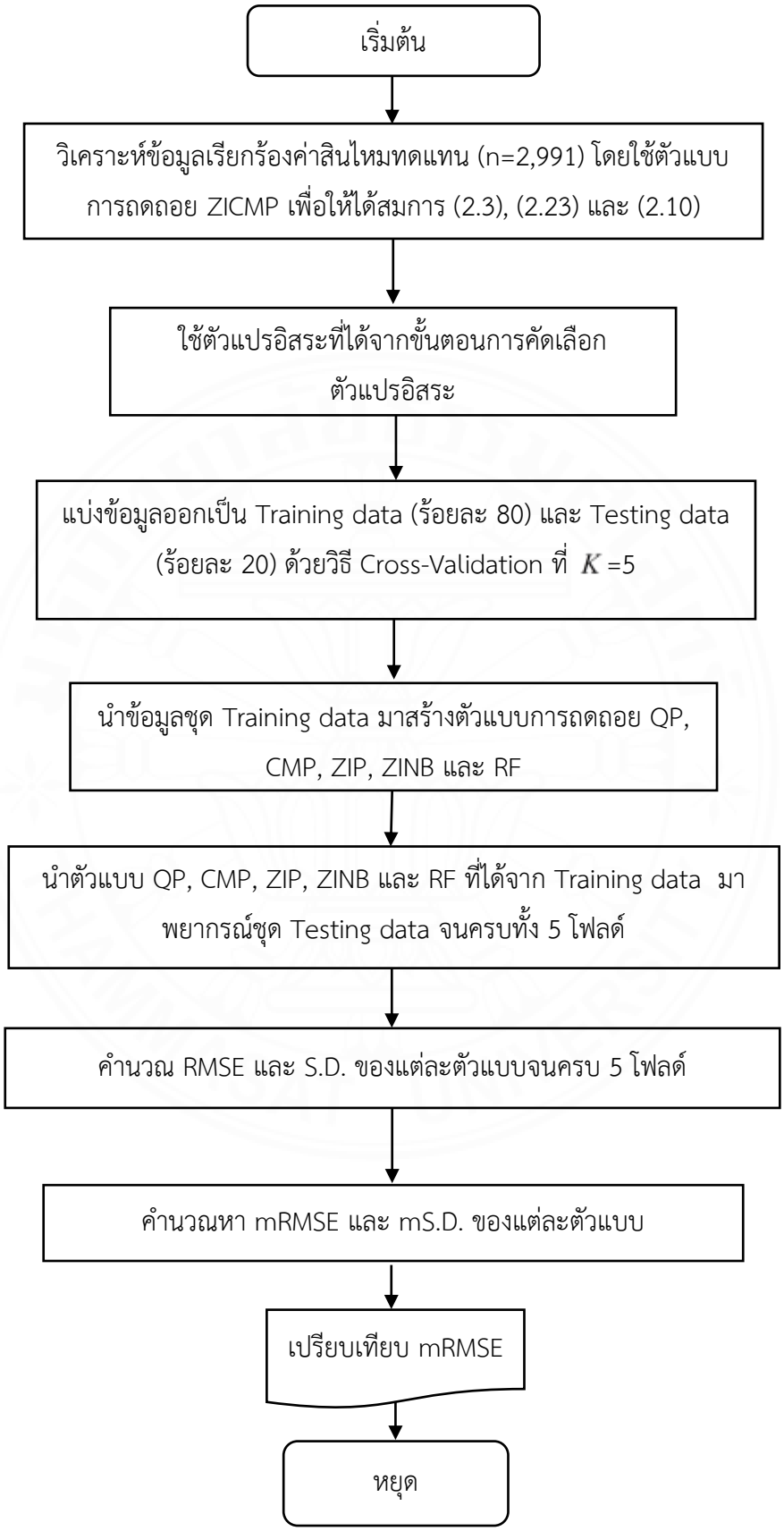
ทำการตรวจสอบการแจกแจงของตัวแปรอิสระเชิงปริมาณด้วยโปรแกรม Easy Fit เวอร์ชัน 5.5 โดยมีขั้นตอนดังต่อไปนี้

1. นำเข้าข้อมูลตัวแปรอิสระเชิงปริมาณในรูปแบบไฟล์ Excel เข้าสู่โปรแกรม โดยที่ตัวแปรนี้ถูกคัดเลือกมาจากตัวแบบการถดถอย ZICMP ดังตารางที่ 3.3 ได้แก่ เบี้ยประกันภัย และส่วนลดประวัติดีเท่านั้น ทั้งนี้ตัวแปรอิสระเชิงปริมาณด้านอายุรถยนต์ที่รับประกันและตัวแปรอิสระเชิงคุณภาพ ได้แก่ เพศและการต่อกรมธรรม์ ไม่มีผลต่อการประมาณค่าพารามิเตอร์ โดยที่ AIC เท่ากับ 5,541.207 และค่า Log-likelihood เท่ากับ -2,764.604 ตารางที่ 3.4 แสดงสถิติพรรณนาของตัวแปรอิสระที่มีผลต่อการประมาณค่าพารามิเตอร์

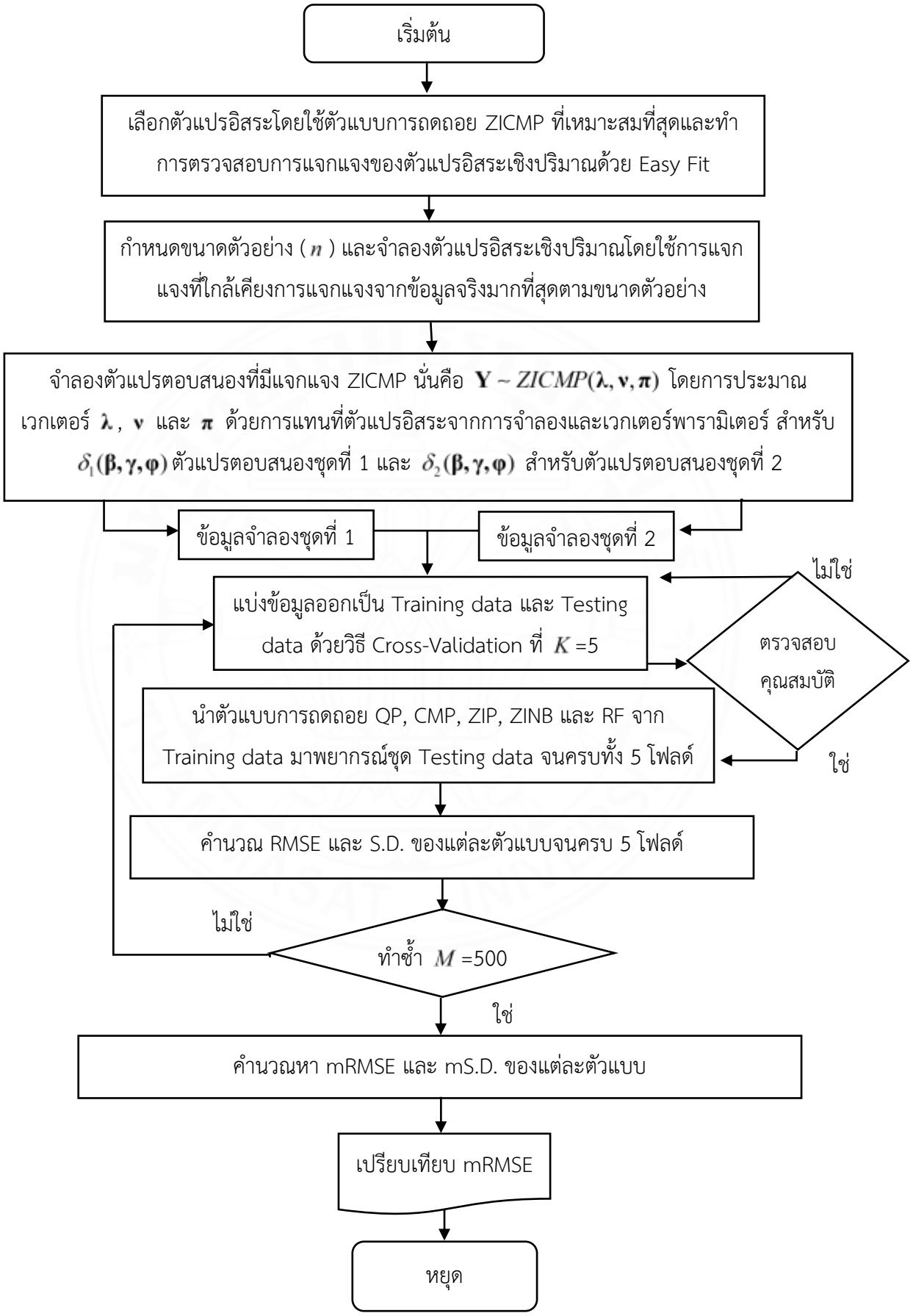
2. คลิกคำสั่ง Analysis บนแถบเครื่องมือและเลือกสดมภ์ที่ต้องการวิเคราะห์ จากนั้นเลือกชนิดการแจกแจง ในที่นี้เลือกการแจกแจงแบบต่อเนื่อง (Continuous distribution)

3. โปรแกรมจะแสดงผลโดยมีการแจกแจงที่ใกล้เคียงกับการแจกแจงของตัวแปรอิสระมากกว่า 10 รูปแบบ พร้อมทั้งให้ค่าพารามิเตอร์ของการแจกแจงนั้น

4. ผู้วิจัยเลือกการแจกแจง 3 อันดับแรก ทำการจำลองข้อมูลและสร้างฮิสโทแกรมด้วยโปรแกรม R เวอร์ชัน 4.4 จากการแจกแจงทั้ง 3 จากนั้นพิจารณาลักษณะการกระจายของข้อมูลที่ได้จากการแจกแจงทั้ง 3 เทียบกับการการกระจายของข้อมูลจริง การแจกแจงใดที่มีลักษณะการกระจายใกล้เคียงกับข้อมูลจริงมากที่สุดจะถูกเลือกเพื่อใช้สร้างข้อมูลจำลองตามขนาดตัวอย่างในหัวข้อ 3.1.3 ข้อ 1



ภาพที่ 3.3 ขั้นตอนการวิเคราะห์ข้อมูลจริงกรณีการกระจายเกินเกณฑ์



ภาพที่ 3.4 ขั้นตอนการวิเคราะห์ข้อมูลจำลองกรณีการกระจายเกินเกณฑ์

ตารางที่ 3.3 ตารางวิเคราะห์ข้อมูลการเรียกร้องค่าสินไหมทดแทนจากข้อมูลกรมธรรม์ประกันภัยรถยนต์ ปี พ.ศ. 2560 จากการตัวแบบการถดถอย ZICMP (n = 2,991)

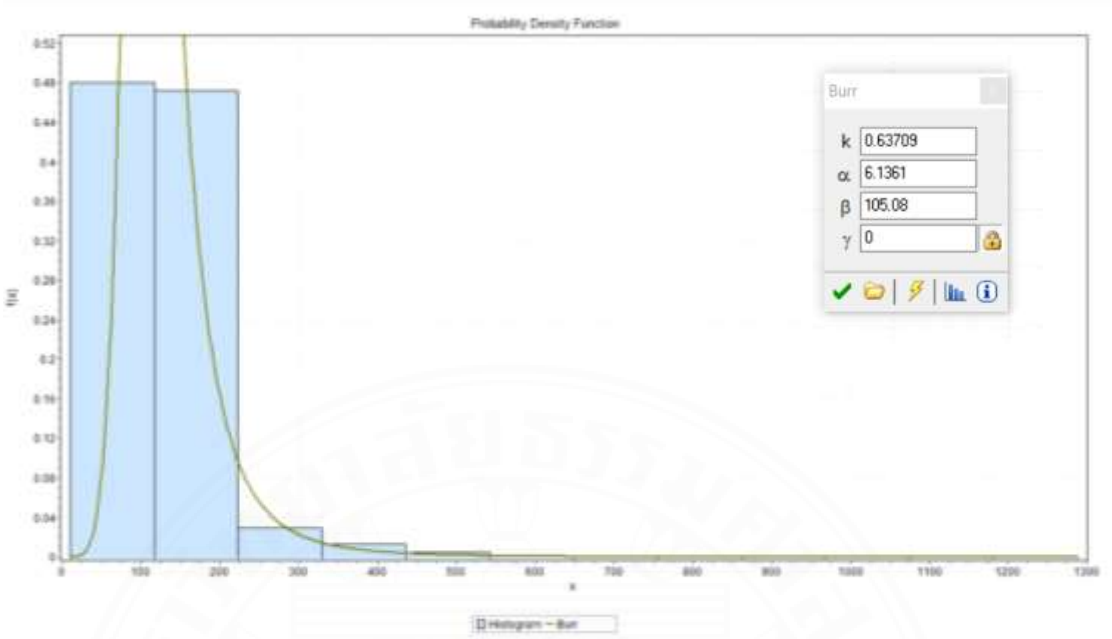
Coefficients	Estimate	S.E.	Z value	P-value
พารามิเตอร์ค่าเฉลี่ย				
ค่าคงที่	-0.3997	0.2082	-1.9194	0.0549
เบี้ยประกันภัย	0.0009	0.0004	2.4268	0.0152
ส่วนลดประวัติดี	-0.0032	0.0007	-4.7691	0.0001
พารามิเตอร์ค่าการกระจาย				
ค่าคงที่	-0.4013	0.2827	-1.4193	0.1558
พารามิเตอร์ความน่าจะเป็นที่จะเกิดศูนย์				
ค่าคงที่	0.6330	0.3883	1.6302	0.1031
เบี้ยประกันภัย	-0.0123	0.0052	-2.3642	0.0181

ตารางที่ 3.4 สถิติพรรณนาของตัวแปรอิสระที่มีผลต่อการประมาณค่าพารามิเตอร์

ชื่อตัวแปร	**คำอธิบายตัวแปร**	**ค่าเฉลี่ย (S.D.)**
Premium	เบี้ยประกันภัย (1 : 100 บาทต่อคน)	130.14 (63.70)
NCB	ส่วนลดประวัติดี (1:100 บาทต่อคน)	78.62 (47.42)

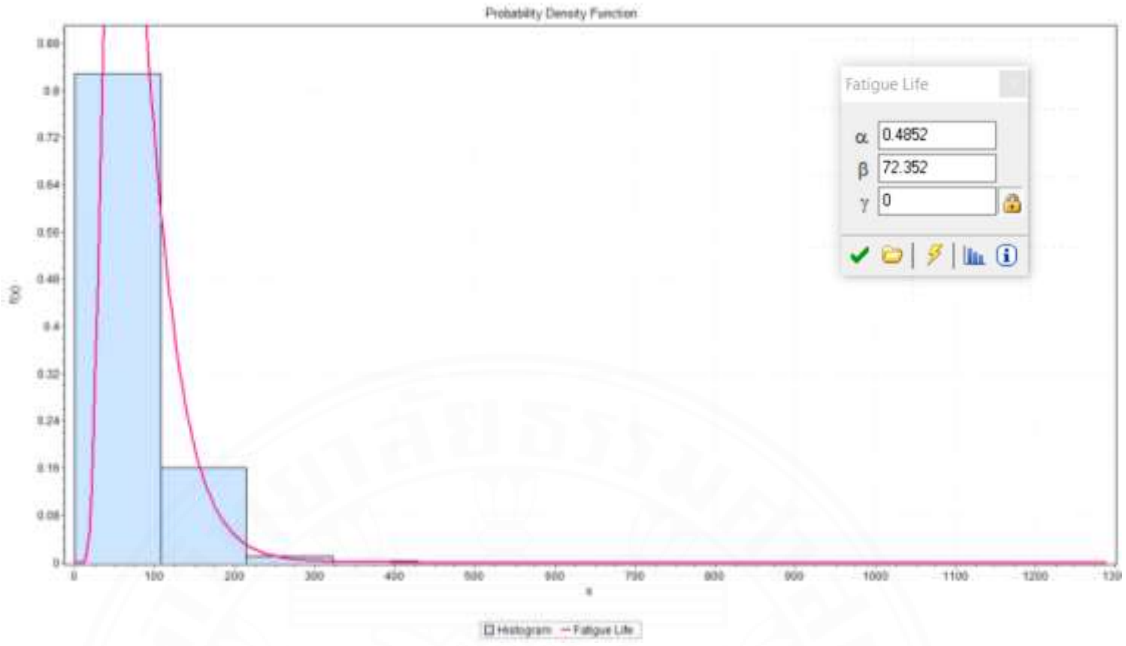
ในงานวิจัยครั้งนี้ พบว่า ตัวแปรเบี้ยประกันภัยมีการแจกแจงใกล้เคียงที่สุด 3 อันดับแรก คือ การแจกแจงเบอร์ (Burr distribution) การแจกแจงลอจิสติก 3 พารามิเตอร์ (Logistic 3P distribution) และการแจกแจงเบอร์ 4 พารามิเตอร์ 4P (Burr distribution) ตามลำดับ สำหรับตัวแปรส่วนลดประวัติดี มีการแจกแจงใกล้เคียงที่สุด 3 อันดับแรก คือ การแจกแจงค่าสุดขีดวางนัยทั่วไป (Generalized extreme value distribution) การแจกแจงแกมมาวางนัยทั่วไป (Generalized gamma distribution) และการแจกแจงเบิร์นบัม-แซนเดอร์ (Birnbuam-sander (Fatigue life) distribution)

การแจกแจงเบอร์เป็นการแจกแจงที่ใกล้เคียงการแจกแจงของตัวแปรเบี้ยประกันภัยที่สุด และการแจกแจงเบิร์นบัม-แซนเดอร์เป็นการแจกแจงที่ใกล้เคียงกับการแจกแจงของตัวแปรส่วนลดประวัติดีที่สุด ดังแสดงในภาพที่ 3.5 และ ภาพที่ 3.6 ตามลำดับ



ภาพที่ 3.5 การแจกแจงเบอร์และค่าพารามิเตอร์ของตัวแปรเบี้ยประกันภัยจากโปรแกรม Easy Fit

จากภาพที่ 3.5 จากการตรวจสอบการแจกแจงของตัวแปรเบี้ยประกันภัยด้วยโปรแกรม Easy Fit พบว่า $X_i \sim Burr(\beta = 105.08, \alpha = 6.1361, k = 0.63709)$ ในที่นี้ β หมายถึง พารามิเตอร์บ่งตำแหน่ง (Location parameter) และ α หมายถึง พารามิเตอร์การกระจาย (Dispersion parameter) และ k หมายถึง ตระกูลพารามิเตอร์ (Family parameter) จากนั้นทำการจำลองตัวแปรเบี้ยประกันภัยในโปรแกรม R โดยใช้แพ็คเกจ “rmutil” ซึ่งเป็นแพ็คเกจสำหรับการจำลองการแจกแจงเบอร์ โดยนำค่าพารามิเตอร์ที่ได้จากโปรแกรม Easy Fit มาใช้ในการจำลองเพื่อเลียนแบบการแจกแจงของตัวแปรเบี้ยประกันภัยที่ขนาดตัวอย่างต่าง ๆ ในขั้นตอนต่อไป



ภาพที่ 3.6 การแจกแจงเบิร์นบัม-แซนเดอร์และค่าพารามิเตอร์ของตัวแปรส่วนลดประวัติดีจากโปรแกรม Easy Fit

จากภาพที่ 3.6 จากการตรวจสอบการแจกแจงของตัวแปรส่วนลดประวัติดีด้วยโปรแกรม Easy Fit พบว่า $X_i \sim Fatigue(\alpha = 0.4852, \beta = 72.352, \gamma = 0)$ ในที่นี้ α หมายถึง พารามิเตอร์บ่งรูปร่าง (Shape parameter) และ β หมายถึง พารามิเตอร์มาตราส่วน (Scale parameter) และ γ หมายถึง พารามิเตอร์บ่งตำแหน่ง (Location parameter) จากนั้นทำการจำลองในโปรแกรม R โดยใช้แพ็คเกจ “extraDistr” เป็นแพ็คเกจสำหรับการจำลองการแจกแจงเบิร์นบัม-แซนเดอร์ โดยนำค่าพารามิเตอร์ที่ได้จากโปรแกรม Easy Fit มาใช้ในการจำลองเพื่อเลียนแบบการแจกแจงของตัวแปรส่วนลดประวัติดีที่ขนาดตัวอย่างต่าง ๆ ในขั้นตอนต่อไป

บทที่ 4
ผลการวิจัยและอธิบายผล

ในงานวิจัยนี้เป็นการศึกษาเพื่อเปรียบเทียบประสิทธิภาพการพยากรณ์ของตัวแบบสำหรับข้อมูลจำนวนนับที่มีปัญหาการกระจายและค่าศูนย์เฟ้อด้วยวิธีเค-โฟลด์ตรวจสอบไขว้ ซึ่งเป็นการสุ่มแบ่งข้อมูลที่ไม่ซ้ำกันออกเป็น 5 ชุดเท่า ๆ กัน โดยตัวแบบที่ใช้ในการพยากรณ์ข้อมูลแต่ละชุดได้แก่ ตัวแบบการถดถอย QP, CMP, ZIP, ZINB และเทคนิค RF มีเกณฑ์ในการพิจารณา คือ ค่าเฉลี่ยของรากของความคลาดเคลื่อนกำลังสองเฉลี่ยโดยเฉลี่ย (mRMSE) และค่าเฉลี่ยของค่าเบี่ยงเบนมาตรฐานของ mRMSE (mS.D.) เกณฑ์ดังกล่าวจะนำเสนอในรูปแบบของค่าเฉลี่ยจากการทำซ้ำทั้งหมด M = 500 รอบ แบ่งการนำเสนอออกเป็น 2 กรณี คือ

กรณีที่ 1 กรณีข้อมูลมีการกระจายต่ำกว่าเกณฑ์และค่าศูนย์เฟ้อและแสดงผลลัพธ์เกณฑ์วัดประสิทธิภาพจากข้อมูลจริงและข้อมูลจำลอง โดยใช้ตัวแบบ CMP และเทคนิค RF

กรณีที่ 2 กรณีข้อมูลมีการกระจายเกินเกณฑ์และค่าศูนย์เฟ้อและแสดงผลลัพธ์เกณฑ์วัดประสิทธิภาพจากข้อมูลจริงและข้อมูลจำลอง โดยใช้ตัวแบบ QP, CMP, ZIP, ZINB และเทคนิค RF

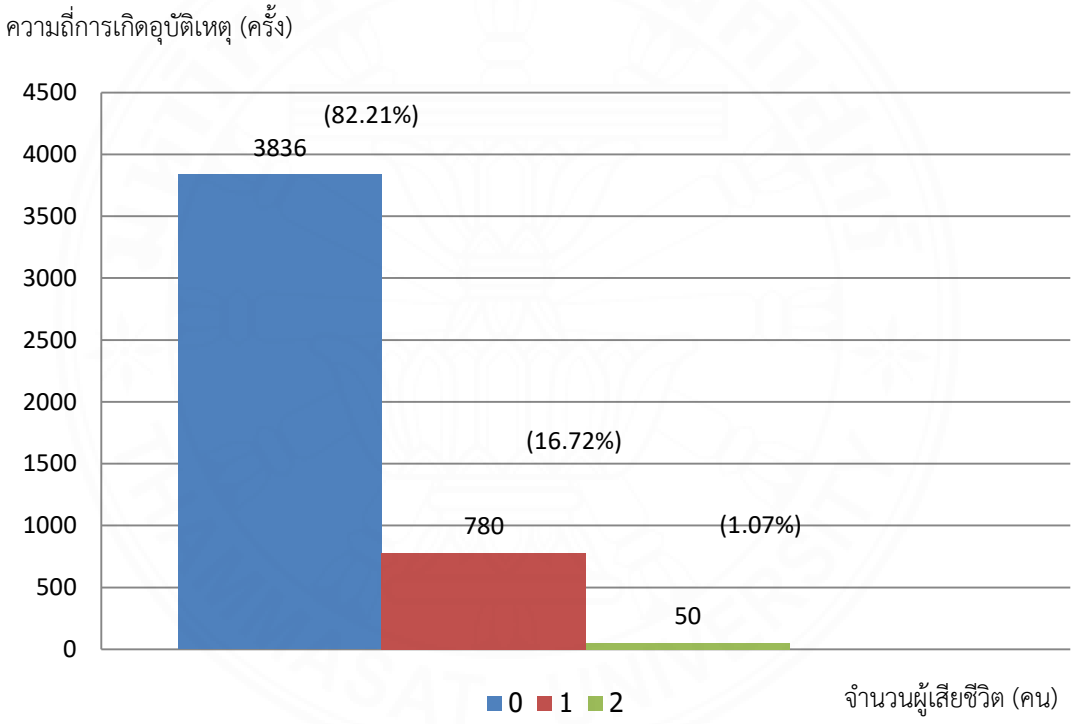
4.1. ผลการวิจัยกรณีข้อมูลมีการกระจายต่ำกว่าเกณฑ์และค่าศูนย์เฟ้อ

4.1.1 ผลการเปรียบเทียบประสิทธิภาพของตัวแบบสำหรับข้อมูลจริง

ข้อมูลที่ใช้ในการศึกษานี้เป็นข้อมูลสถิติอุบัติเหตุทางถนนที่ส่งผลให้มีผู้เสียชีวิตทั่วประเทศไทย เดือนเมษายน ปี พ.ศ. 2558 มีจำนวนทั้งหมด 4,666 เหตุการณ์ มีค่าเฉลี่ยของจำนวนผู้เสียชีวิต (λ) เท่ากับ 0.1886 คนต่อครั้ง มีค่าการกระจาย (ν) เท่ากับ 1.0808 เมื่อทำการทดสอบการกระจายของข้อมูลด้วย Dispersion test พบว่า ค่าสถิติทดสอบเท่ากับ -9.6372 และ p-value น้อยกว่า 0.0001 นั่นคือข้อมูลมีการกระจายต่ำกว่าเกณฑ์ และจำนวนอุบัติเหตุที่ไม่มีผู้เสียชีวิต 3,836 ครั้ง คิดเป็นร้อยละ 82.21 (ความน่าจะเป็นที่จะเกิดศูนย์ (π) เท่ากับ 0.8221) เมื่อทำการทดสอบค่าศูนย์เฟ้อด้วย Score test พบว่า ค่าสถิติทดสอบเท่ากับ 12.9274 และ p-value น้อยกว่า 0.0003 นั่นคือ ข้อมูลมีสัดส่วนของค่าศูนย์เฟ้อ จำนวนอุบัติเหตุที่มีผู้เสียชีวิตหนึ่งคนและสองคน คือ 780 ครั้ง (ร้อยละ 16.72) และ 50 ครั้ง (ร้อยละ 1.07) ตามลำดับ (ตารางที่ 4.1 และภาพที่ 4.1)

ตารางที่ 4.1 สถิติพรรณนาข้อมูลอุบัติเหตุทางถนนที่ส่งผลให้มีผู้เสียชีวิตทั่วประเทศไทย เดือนเมษายน ปี พ.ศ. 2558 (n = 4,666)

ชื่อตัวแปร	คำอธิบายตัวแปร	ความถี่ (ร้อยละ)
Y: HUMAN_DEAD	จำนวนผู้ที่เสียชีวิตหรือเสียชีวิตในเวลาต่อมาจากอุบัติเหตุทางถนน (ครั้ง)	
	0 ไม่มีผู้เสียชีวิต	3,836 (82.21)
	1 มีผู้เสียชีวิตจำนวน 1 คน	780 (16.72)
	2 มีผู้เสียชีวิตจำนวน 2 คน	50 (1.07)



ภาพที่ 4.1 แผนภูมิแสดงความถี่ของการเกิดอุบัติเหตุทางถนนที่ส่งผลให้มีผู้เสียชีวิตทั่วประเทศไทย เดือนเมษายน ปี พ.ศ. 2558

ผลการเปรียบเทียบประสิทธิภาพการพยากรณ์ของตัวแบบการถดถอย CMP และเทคนิค RF ในข้อมูลชุดทดสอบ โดยนำเสนอผลการศึกษาในรูปแบบตาราง ได้แก่ ค่า RMSE, mRMSE และ S.D. (ตารางที่ 4.2)

ตารางที่ 4.2 ผลการเปรียบเทียบประสิทธิภาพการพยากรณ์ของตัวแบบการถดถอย CMP และเทคนิค RF กรณีข้อมูลอุบัติเหตุทางถนนที่ส่งผลให้มีผู้เสียชีวิตทั่วประเทศไทย เดือนเมษายน ปี พ.ศ. 2558

	RMSE						
ตัวแบบ	โฟลด์ 1	โฟลด์ 2	โฟลด์ 3	โฟลด์ 4	โฟลด์ 5	mRMSE	S.D.
CMP	**0.3981**	**0.4015**	0.4260	**0.4021**	0.3912	**0.4038**	**0.0132**
RF	0.4000	0.4034	**0.4258**	0.4062	**0.3897**	0.4050	**0.0132**

หมายเหตุ **ตัวหนา** แทนวิธีที่ให้ค่าเฉลี่ยของรากของค่าคลาดเคลื่อนกำลังสองเฉลี่ยของตัวแบบที่ต่ำที่สุด

จากตารางที่ 4.2 พบว่า ค่า mRMSE คำนวณจากชุดทดสอบจำนวน 5 โฟลด์ (n = 933, 20%) ของตัวแบบมีความใกล้เคียงกัน โดยจะต่างกันที่ทศนิยมตำแหน่งที่สามเท่านั้น ซึ่งตัวแบบการถดถอย CMP มีประสิทธิภาพดีกว่าเทคนิค RF เพียงเล็กน้อย โดยมีค่า mRMSE เท่ากับ 0.4038 และ 0.4050 ตามลำดับ อีกทั้งพบว่า ตัวแบบทั้งสองมีค่าเบี่ยงเบนมาตรฐานไม่แตกต่างกัน

4.1.2 ผลการเปรียบเทียบประสิทธิภาพของตัวแบบสำหรับข้อมูลจำลอง

สำหรับการจำลองข้อมูลที่ 1 ใช้ค่าสัมประสิทธิ์การถดถอยจากตัวแบบการถดถอย ZICMP จากข้อมูลจริง (ตารางที่ 3.1) นั่นคือ เวกเตอร์ของพารามิเตอร์ $\delta_1(\boldsymbol{\beta},\boldsymbol{\gamma},\boldsymbol{\varphi})$ และสำหรับข้อมูลจำลองชุดที่ 2 ใช้เวกเตอร์ของพารามิเตอร์ $\delta_2(\boldsymbol{\beta},\boldsymbol{\gamma},\boldsymbol{\varphi}) = 1.5 \times \delta_1(\boldsymbol{\beta},\boldsymbol{\gamma},\boldsymbol{\varphi})$ ในการจำลอง (ค่าสัมประสิทธิ์การถดถอยจากข้อมูลจริงเพิ่มขึ้น 50 เปอร์เซ็นต์) จากนั้นทำการเปรียบเทียบประสิทธิภาพการพยากรณ์ของตัวแบบด้วยวิธีเค-โฟลด์ตรวจสอบไขว้โดยตัวแบบที่ใช้ ได้แก่ ตัวแบบการถดถอย CMP และเทคนิค RF และคำนวณเกณฑ์การเปรียบเทียบประสิทธิภาพ mRMSE และ mS.D. ในข้อมูลชุดทดสอบที่ขนาดตัวอย่าง ($\boldsymbol{n}$) = 250 500 1,000 3,000 และ 5,000 ทำซ้ำทั้งหมด 500 รอบ ซึ่งแสดงในรูปแบบภาพและตารางต่อไปนี้

ตารางที่ 4.3 ค่า mRMSE ของตัวแบบการถดถอย CMP และเทคนิค RF จากข้อมูลจำลองชุดที่ 1

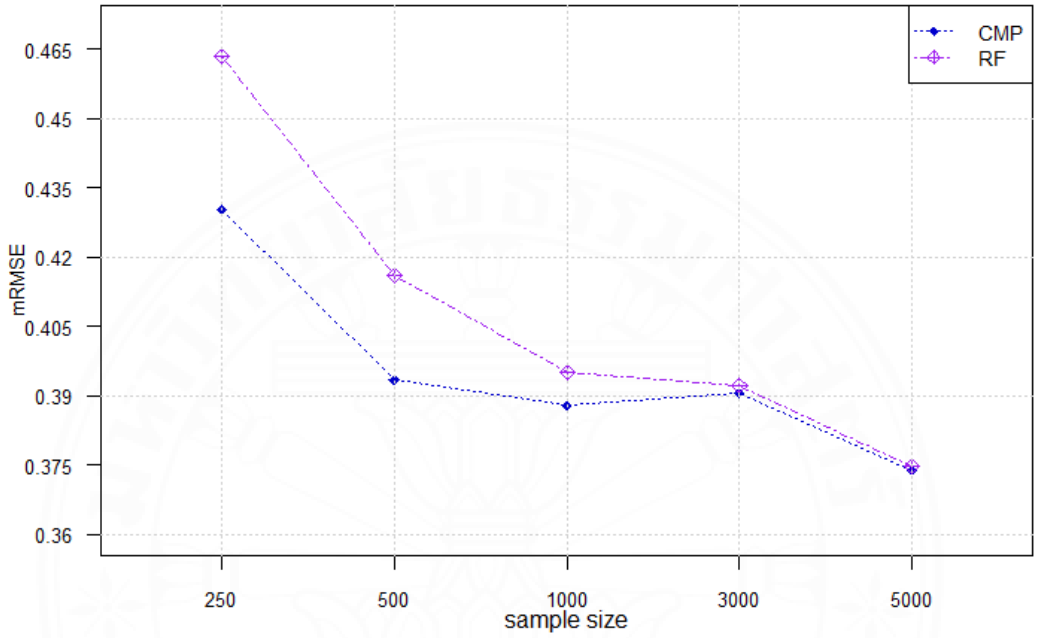
n	$\bar{\lambda}_i$	$\bar{\nu}_i$	$\bar{\pi}_i$	mRMSE (mS.D.)	
				CMP	RF
250	2.0125	5.0445	0.6246	**0.4303**	0.4636
				(0.0540)	(0.0605)
500	1.4612	5.0445	0.6168	**0.3934**	0.4160
				(0.0366)	(0.0362)
1,000	1.8094	5.0445	0.6184	**0.3880**	0.3951
				(0.0260)	(0.0264)
3,000	1.7723	5.0445	0.6176	**0.3906**	0.3922
				(0.0182)	(0.0180)
5,000	1.7553	5.0445	0.6220	**0.3740**	0.3749
				(0.0164)	(0.0161)

ตารางที่ 4.4 ค่า mRMSE ของตัวแบบการถดถอย CMP และเทคนิค RF จากข้อมูลจำลองชุดที่ 2

n	$\bar{\lambda}_i$	$\bar{\nu}_i$	$\bar{\pi}_i$	mRMSE (mS.D.)	
				CMP	RF
250	4.7271	11.3300	0.6728	**0.3588**	0.3715
				(0.0391)	(0.0426)
500	2.4653	11.3300	0.6679	0.3296	**0.3291**
				(0.0297)	(0.0295)
1,000	3.5976	11.3300	0.6689	**0.3162**	0.3256
				(0.0223)	(0.0223)
3,000	3.5111	11.3300	0.6673	**0.3010**	0.3025
				(0.0125)	(0.0125)
5,000	3.4959	11.3300	0.6724	**0.2974**	0.2980
				(0.0098)	(0.0098)

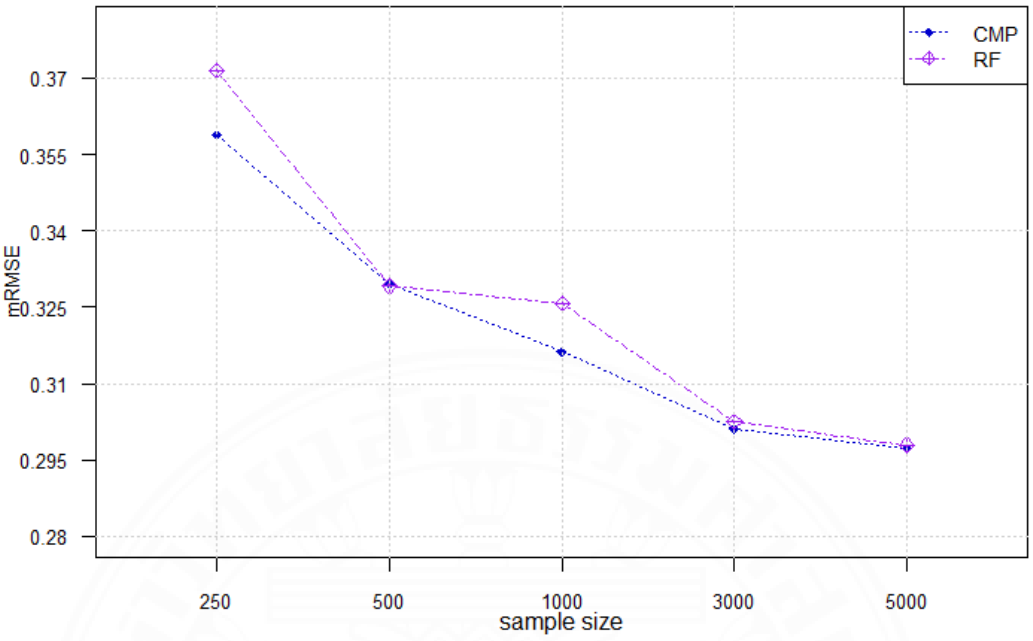
จากตารางที่ 4.3 เมื่อค่าสัมประสิทธิ์การถดถอยจากข้อมูลจริงเพิ่มขึ้น 50 เปอร์เซ็นต์ (ตารางที่ 4.4) พบว่า ตัวแบบการถดถอย CMP และเทคนิค RF มีประสิทธิภาพดีขึ้น

เนื่องจาก mRMSE ลดลงทุกกรณี อีกทั้งพบว่า เมื่อค่าสัมประสิทธิ์การถดถอยเพิ่มขึ้น ค่าเฉลี่ยของพารามิเตอร์ค่าเฉลี่ย ($\overline{\lambda}_i$), ค่าเฉลี่ยของพารามิเตอร์การกระจาย ($\overline{\nu}_i$) และค่าเฉลี่ยของพารามิเตอร์ความน่าจะเป็นที่จะเกิดศูนย์ ($\overline{\pi}_i$) เพิ่มขึ้นที่ขนาดตัวอย่างต่าง ๆ ทุกกรณี



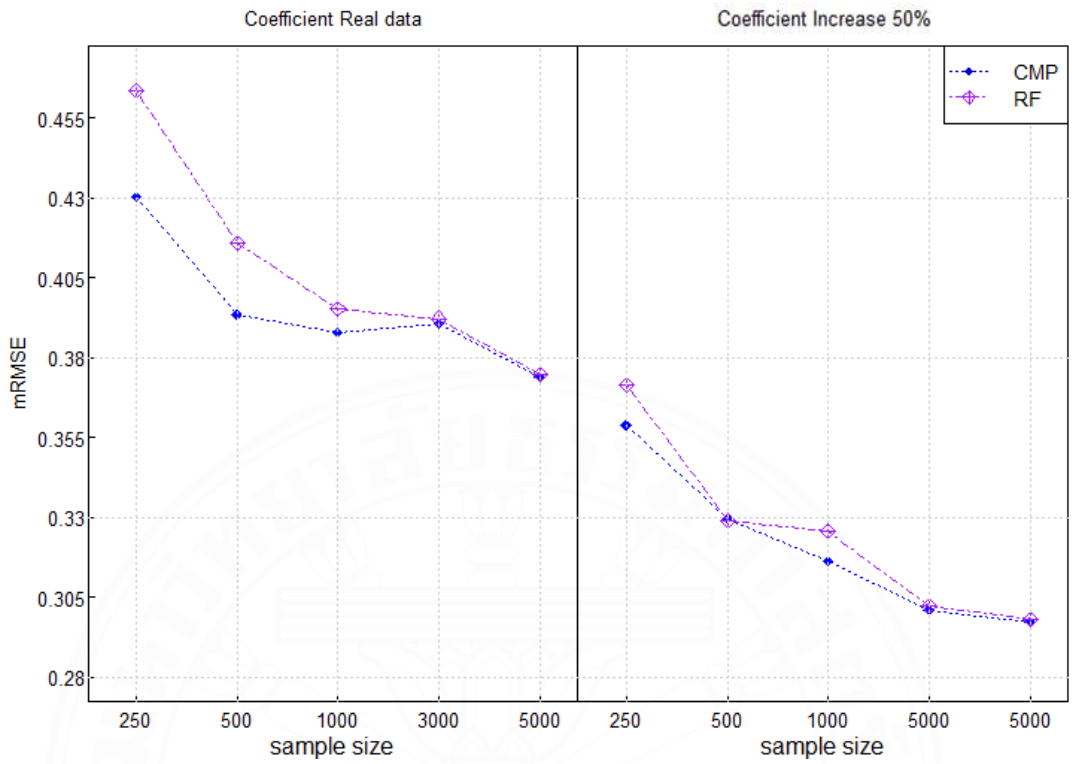
ภาพที่ 4.2 ค่า mRMSE ของตัวแบบการถดถอย CMP และเทคนิค RF เมื่อขนาดตัวอย่าง n = 250 500 1,000 3,000 และ 5,000 จากข้อมูลจำลองชุดที่ 1 (ลักษณะเส้น, สัญลักษณ์และสี คือ ตัวแบบ)

จากภาพที่ 4.2 และตารางที่ 4.3 เป็นข้อมูลจำลองที่ใช้สัมประสิทธิ์การถดถอยจากข้อมูลจริง (ข้อมูลจำลองชุดที่ 1) พบว่า เมื่อขนาดตัวอย่างเพิ่มขึ้นประสิทธิภาพของตัวแบบการถดถอย CMP และเทคนิค RF ดีขึ้นเนื่องจาก mRMSE มีแนวโน้มลดลงและมีค่าใกล้เคียงกันมากขึ้น โดยเฉพาะเมื่อขนาดตัวอย่างเพิ่มขึ้นเป็น 3,000 และ 5,000 เมื่อขนาดตัวอย่างน้อย (250) ตัวแบบมีประสิทธิภาพแตกต่างกัน โดยตัวแบบการถดถอย CMP มีประสิทธิภาพดีกว่าเทคนิค RF เนื่องจากให้ค่า mRMSE ต่ำกว่าทุกกรณี



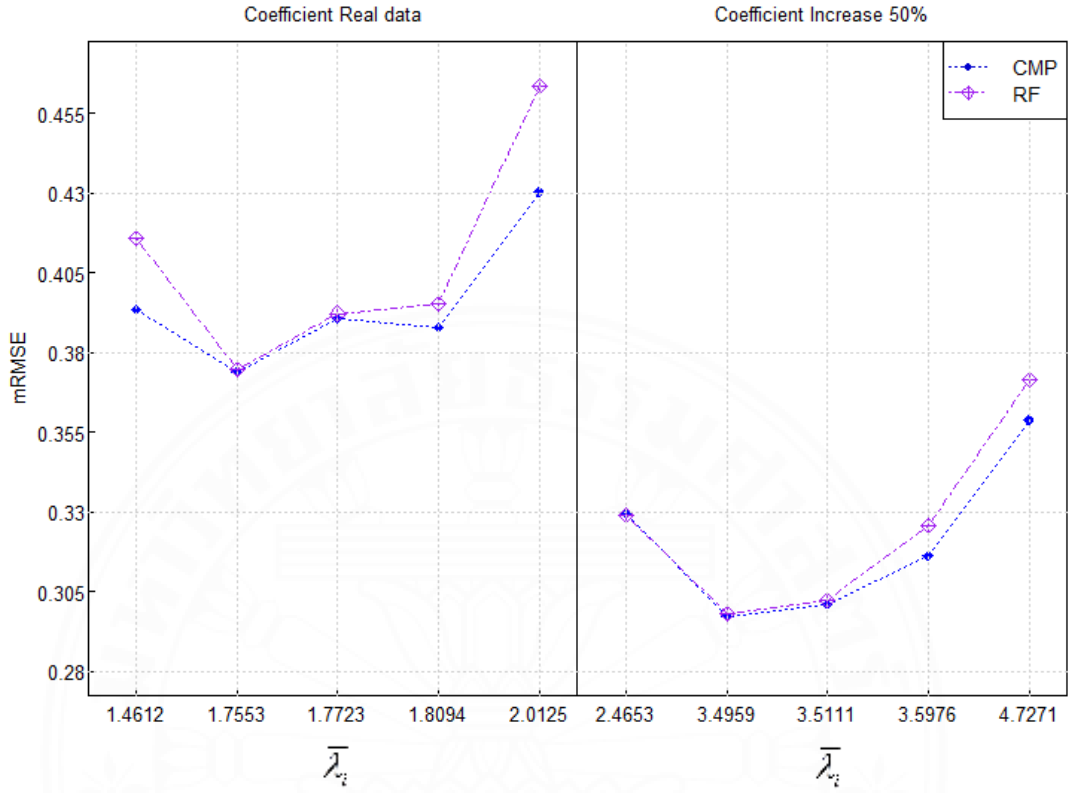
ภาพที่ 4.3 ค่า mRMSE ของตัวแบบการถดถอย CMP และเทคนิค RF เมื่อขนาดตัวอย่าง n = 250 500 1,000 3,000 และ 5,000 จากข้อมูลจำลองชุดที่ 2 (ลักษณะเส้น, สัญลักษณ์และสี คือ ตัวแบบ)

จากภาพที่ 4.3 และตารางที่ 4.4 เป็นข้อมูลจำลองที่ใช้สัมประสิทธิ์การถดถอยจากข้อมูลจริงเพิ่มขึ้น 50 เปอร์เซ็นต์ (ข้อมูลจำลองชุดที่ 2) พบว่า เมื่อขนาดตัวอย่างเพิ่มขึ้น ประสิทธิภาพของตัวแบบการถดถอย CMP และเทคนิค RF ดีขึ้นเนื่องจาก mRMSE มีแนวโน้มลดลง และมีค่าใกล้เคียงกันมากขึ้น โดยเฉพาะที่ขนาดตัวอย่างเท่ากับ 3,000 และ 5,000 และเมื่อขนาดตัวอย่างน้อย (250) ตัวแบบมีประสิทธิภาพแตกต่างกัน โดยตัวแบบการถดถอย CMP มีประสิทธิภาพดีกว่าเทคนิค RF เนื่องจากให้ค่า mRMSE ต่ำกว่าทุกกรณี ยกเว้นขนาดตัวอย่างเท่ากับ 500 พบว่าเทคนิค RF ให้ค่า mRMSE ต่ำกว่าตัวแบบการถดถอย CMP เล็กน้อย



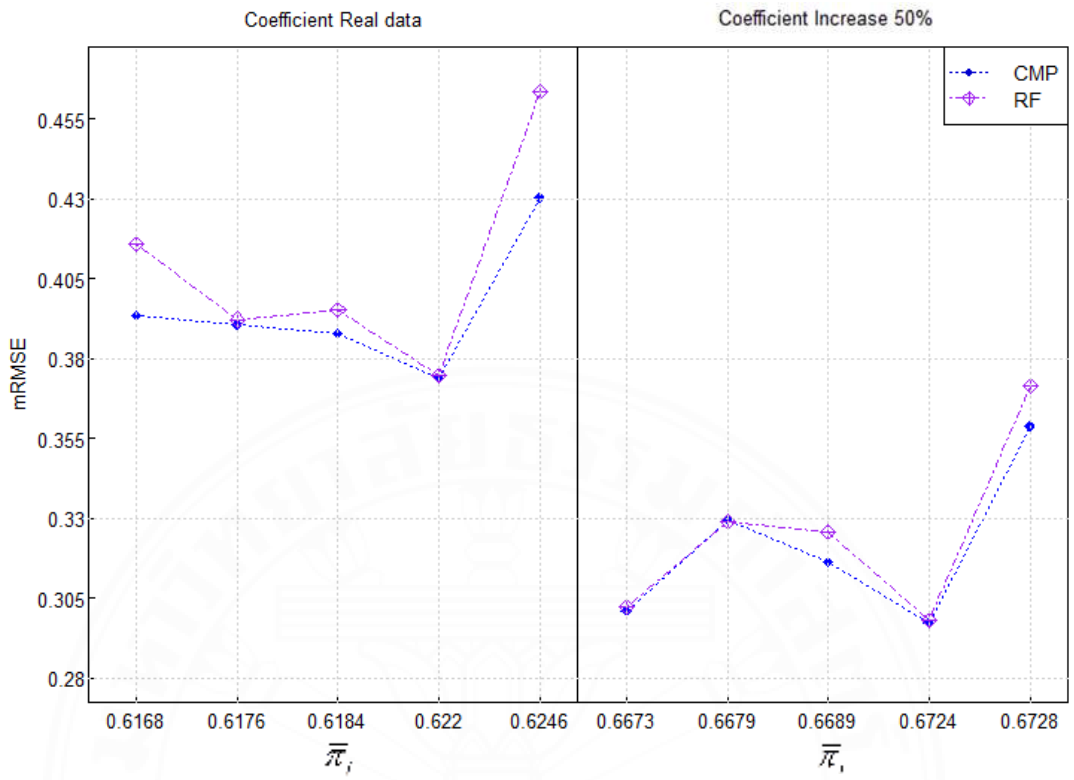
ภาพที่ 4.4 ค่า mRMSE ของตัวแบบการถดถอย CMP และเทคนิค RF เมื่อขนาดตัวอย่าง n = 250 500 1,000 3,000 และ 5,000 ข้อมูลจำลองชุดที่ 1 เปรียบเทียบกับข้อมูลจำลองชุดที่ 2 (ลักษณะเส้น, สัญลักษณ์และสี คือ ตัวแบบ)

จากภาพที่ 4.4 ข้อมูลจำลองชุดที่ 1 (ภาพซ้าย) เมื่อค่าสัมประสิทธิ์การถดถอยเพิ่มขึ้น 50 เปอร์เซ็นต์เป็นข้อมูลจำลองชุดที่ 2 (ภาพขวา) พบว่า ประสิทธิภาพของแต่ละตัวแบบดีขึ้นทุกกรณีและมีพฤติกรรมแบบเดียวกัน นั่นคือ ข้อมูลทั้งสองชุด (ภาพซ้ายและภาพขวา) เมื่อขนาดตัวอย่างน้อยตัวแบบมีประสิทธิภาพแตกต่างกัน โดยตัวแบบการถดถอย CMP มีประสิทธิภาพดีกว่าเทคนิค RF และเมื่อขนาดตัวอย่างเพิ่มขึ้นเป็น 5,000 ประสิทธิภาพของแต่ละตัวแบบดีขึ้นเนื่องจาก mRMSE มีแนวโน้มลดลงและมีประสิทธิภาพใกล้เคียงกันยิ่งขึ้น



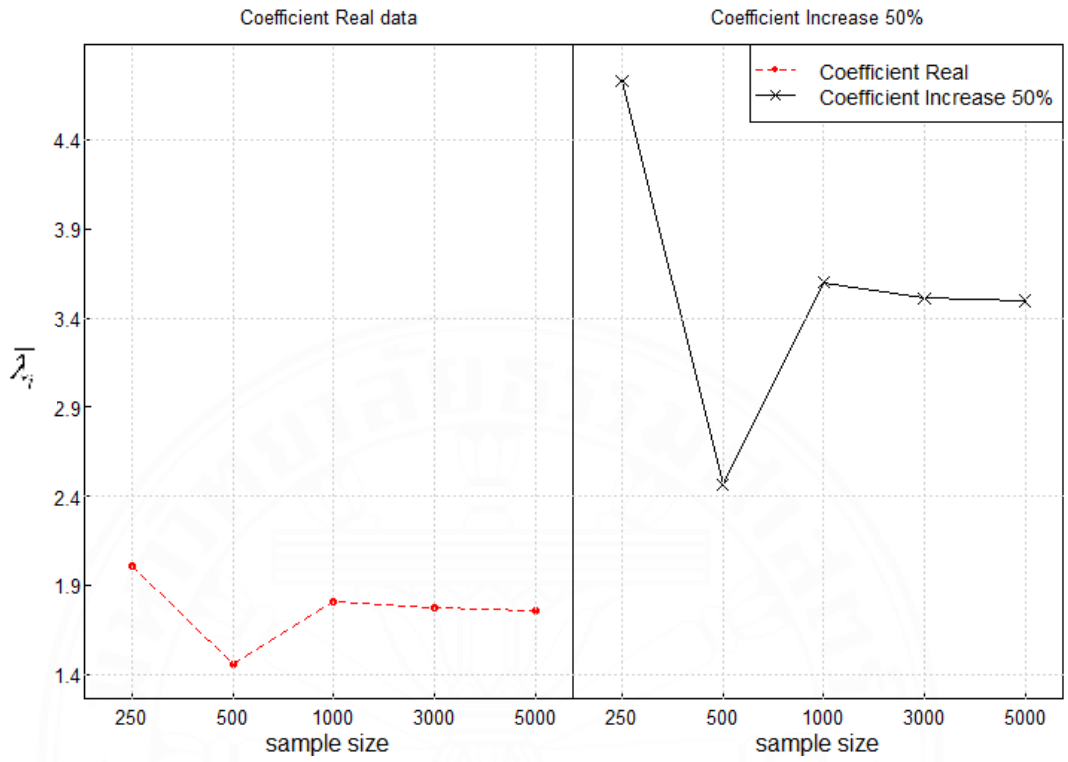
ภาพที่ 4.5 ค่า mRMSE ของตัวแบบการถดถอย CMP และเทคนิค RF ที่ค่า $\overline{\lambda}_i$ ต่าง ๆ ของข้อมูลจำลองชุดที่ 1 เปรียบเทียบกับข้อมูลจำลองชุดที่ 2 (ลักษณะเส้น, สัญลักษณ์และสี คือ ตัวแบบ)

จากภาพที่ 4.5 เมื่อข้อมูลจำลองชุดที่ 1 (ภาพซ้าย) มีค่าสัมประสิทธิ์การถดถอยเพิ่มขึ้น 50 เปอร์เซ็นต์เป็นข้อมูลชุดที่ 2 (ภาพขวา) พบว่า ประสิทธิภาพของแต่ละตัวแบบดีขึ้นทุกกรณี อีกทั้งพบว่า ข้อมูลทั้งสองชุด (ภาพซ้ายและภาพขวา) เมื่อค่า $\overline{\lambda}_i$ เพิ่มขึ้น ประสิทธิภาพของตัวแบบการถดถอย CMP และเทคนิค RF ลดลงเนื่องจาก mRMSE มีแนวโน้มเพิ่มขึ้นและมีพฤติกรรมแบบเดียวกัน



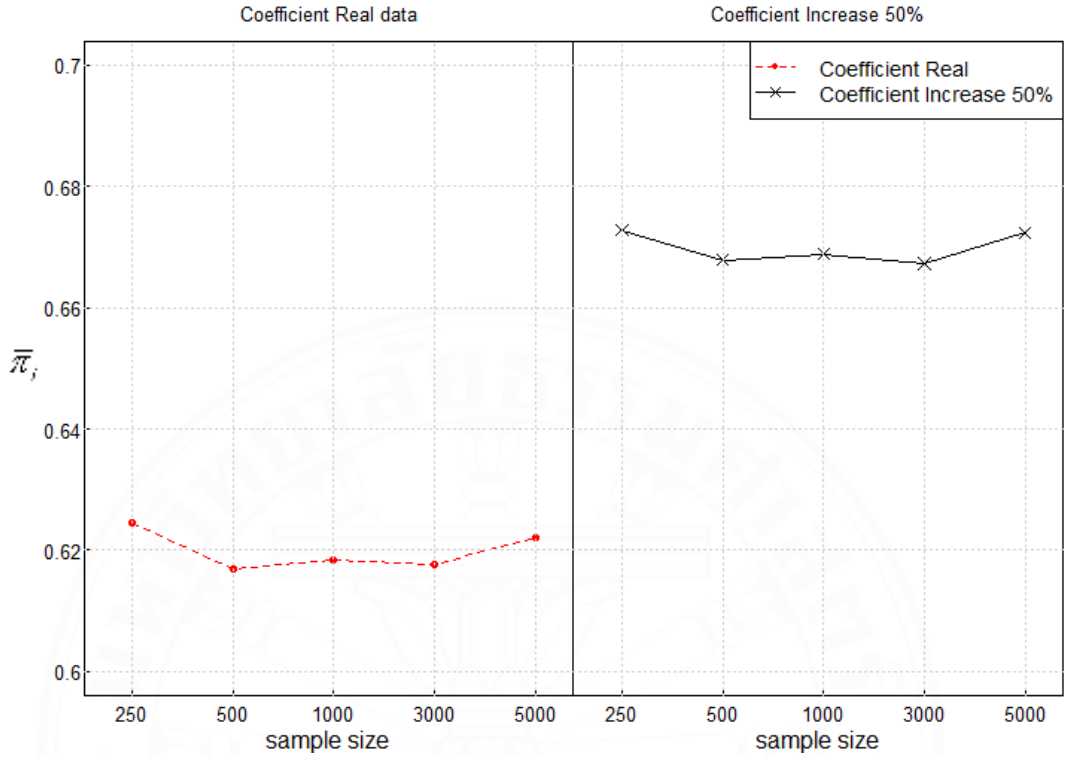
ภาพที่ 4.6 ค่า mRMSE ของตัวแบบการถดถอย CMP และเทคนิค RF ที่ค่า $\overline{\pi}_i$ ต่าง ๆ ของข้อมูลจำลองชุดที่ 1 เปรียบเทียบกับข้อมูลจำลองชุดที่ 2 (ลักษณะเส้น, สัญลักษณ์และสี คือ ตัวแบบ)

จากภาพที่ 4.6 เมื่อข้อมูลจำลองชุดที่ 1 (ภาพซ้าย) มีค่าสัมประสิทธิ์การถดถอยเพิ่มขึ้น 50 เปอร์เซ็นต์เป็นข้อมูลจำลองชุดที่ 2 (ภาพขวา) พบว่า ประสิทธิภาพของแต่ละตัวแบบดีขึ้นทุกกรณี อีกทั้งพบว่า ข้อมูลทั้งสองชุด (ภาพซ้ายและภาพขวา) เมื่อค่า $\overline{\pi}_i$ เพิ่มขึ้น ประสิทธิภาพของตัวแบบการถดถอย CMP และเทคนิค RF ลดลงเล็กน้อยเนื่องจาก mRMSE มีแนวโน้มเพิ่มขึ้นและมีพฤติกรรมแบบเดียวกัน



ภาพที่ 4.7 ค่า $\overline{\lambda_i}$ ที่ขนาดตัวอย่างต่าง ๆ ของข้อมูลจำลองชุดที่ 1 เปรียบเทียบกับข้อมูลจำลองชุดที่ 2 (ลักษณะเส้น, สัญลักษณ์และสี คือ การใช้ค่าสัมประสิทธิ์การถดถอยในการจำลองข้อมูล)

จากภาพที่ 4.7 ในข้อมูลจำลองชุดที่ 1 (รูปซ้าย) พบว่า การเพิ่มขึ้นของขนาดตัวอย่างส่งผลให้ค่าเฉลี่ยของพารามิเตอร์ค่าเฉลี่ยมีแนวโน้มลดลง เมื่อค่าสัมประสิทธิ์การถดถอยเพิ่มขึ้น 50 เปอร์เซ็นต์เป็นข้อมูลชุดที่ 2 (รูปขวา) พบว่า $\overline{\lambda_i}$ มีค่าเพิ่มขึ้นที่ขนาดตัวอย่างต่าง ๆ ทุกกรณีและมีพฤติกรรมเช่นเดียวกับกรณีข้อมูลจำลองชุดที่ 1



ภาพที่ 4.8 ค่า $\bar{\pi}_i$ ที่ขนาดตัวอย่างต่าง ๆ ของข้อมูลจำลองชุดที่ 1 เปรียบเทียบกับข้อมูลจำลองชุดที่ 2 (ลักษณะเส้น, สัญลักษณ์และสี คือ การใช้ค่าสัมประสิทธิ์การถดถอยในการจำลองข้อมูล)

จากภาพที่ 4.8 ในข้อมูลจำลองชุดที่ 1 (รูปซ้าย) พบว่า การเพิ่มขึ้นของขนาดตัวอย่างส่งผลให้ค่าเฉลี่ยของพารามิเตอร์ความน่าจะเป็นที่จะเกิดศูนย์แนวโน้มเพิ่มขึ้น เมื่อค่าสัมประสิทธิ์การถดถอยเพิ่มขึ้น 50 เปอร์เซ็นต์เป็นข้อมูลชุดที่ 2 (รูปขวา) พบว่า $\bar{\pi}_i$ มีค่าเพิ่มขึ้นที่ขนาดตัวอย่างต่าง ๆ ทุกกรณีและมีพฤติกรรมเช่นเดียวกับกรณีข้อมูลจำลองชุดที่ 1

4.2 ผลการวิจัยกรณีข้อมูลมีการกระจายเกินเกณฑ์และค่าศูนย์เฟ้อ

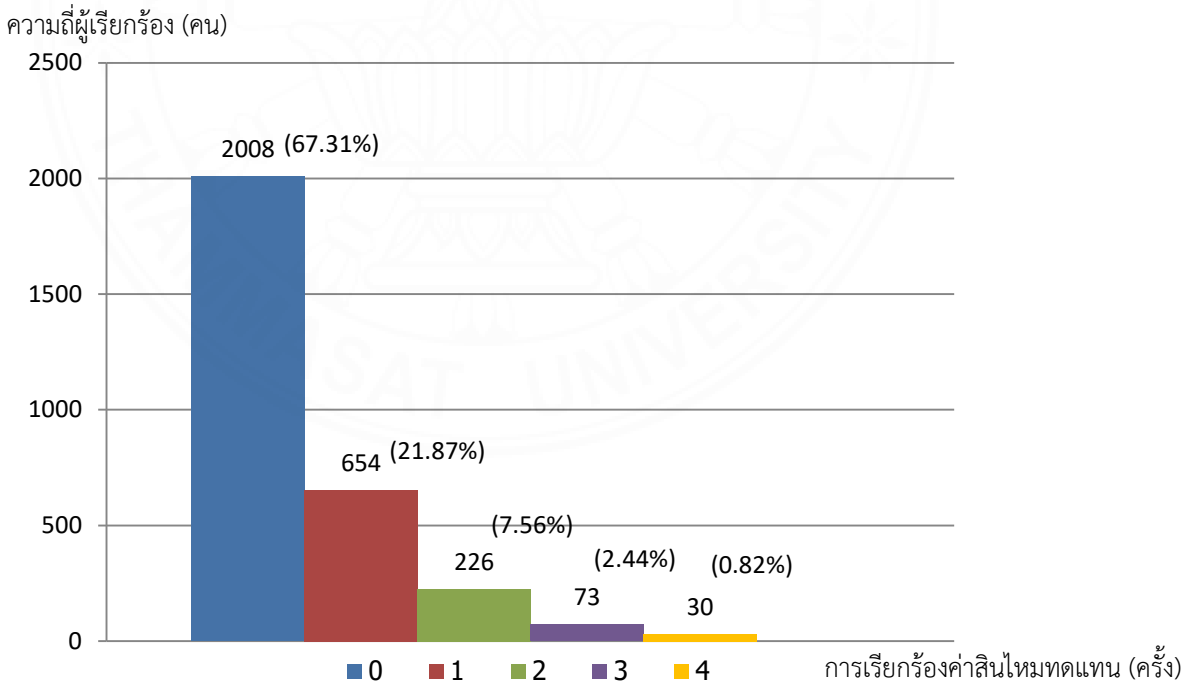
4.2.1 ผลการเปรียบเทียบประสิทธิภาพของตัวแบบสำหรับข้อมูลจริง

กรณีศึกษาเป็นข้อมูลการเรียกร้องค่าสินไหมทดแทนของลูกค้าจากกรมธรรม์ประกันภัยรถยนต์ประเภท 1 จากบริษัทประกันภัยแห่งหนึ่งโดยใช้ข้อมูลจำนวน 2,991 คน ในปี พ.ศ. 2560 มีค่าเฉลี่ยของการเรียกร้องค่าสินไหม (λ) เท่ากับ 0.4983 ครั้ง/คน มีค่าการกระจาย (ν) เท่ากับ 0.7234 เมื่อทำการทดสอบด้วย Dispersion test ได้ ค่าสถิติทดสอบ เท่ากับ 8.6371 และ

p-value น้อยกว่า 0.0001 นั่นคือข้อมูลมีการกระจายเกินเกณฑ์ และจำนวนผู้ที่ไม่มีการเรียกร้องค่าสินไหมทดแทนคิดเป็นร้อยละ (ความน่าจะเป็นที่จะเกิดศูนย์ (π) เท่ากับ 0.6731 เมื่อทำการทดสอบค่าศูนย์เฟ้อด้วย Score test พบว่า ค่าสถิติทดสอบเท่ากับ 169.1142 และ p-value น้อยกว่า 0.0001 นั่นคือ ข้อมูลมีสัดส่วนของค่าศูนย์เฟ้อมีรายละเอียดดังตารางที่ 4.5 และภาพที่ 4.9

ตารางที่ 4.5 สถิติพรรณนาการเรียกร้องค้าสินไหมทดแทนของลูกค้า ปี พ.ศ. 2560 (n = 2,991)

ชื่อตัวแปร	**คำอธิบายตัวแปร**	**ความถี่ (ร้อยละ)**
Y: Claim	จำนวนครั้งของการเรียกร้องค่าสินไหมทดแทน (คน)	
	0 ครั้ง	2,008 (67.31)
	1 ครั้ง	654 (21.87)
	2 ครั้ง	226 (7.56)
	3 ครั้ง	73 (2.44)
	4 ครั้ง	30 (0.82)



ภาพที่ 4.9 แผนภูมิแสดงความถี่การเรียกร้องค่าสินไหมทดแทนจากข้อมูลกรมธรรม์ ปี พ.ศ. 2560

จากตารางที่ 4.5 และภาพที่ 4.9 พบว่า จากข้อมูลกรมธรรม์ประกันภัย จำนวน 2,991 คน มีจำนวนผู้ที่ไม่ได้เรียกร้องค่าสินไหมทดแทนเป็นจำนวนสูงที่สุด จำนวน 2,008 คน คิดเป็นร้อยละ 67.31 ของจำนวนผู้ที่ทำกรมธรรม์ทั้งหมด รองลงมาคือ จำนวนผู้ที่เรียกร้องค่าสินไหมทดแทน 1, 2, 3, 4, 5, 6 และ 8 ครั้ง ได้แก่ 654 คน คิดเป็นร้อยละ 21.87, 226 คน คิดเป็นร้อยละ 7.56, 73 คน คิดเป็นร้อยละ 2.44 และ 30 คน คิดเป็นร้อยละ 0.82 ของจำนวนผู้ที่ทำกรมธรรม์ทั้งหมด ตามลำดับ

ผลการเปรียบเทียบประสิทธิภาพการพยากรณ์ของตัวแบบการถดถอย QP, CMP, ZIP, ZINB และเทคนิค RF ในข้อมูลชุดทดสอบ โดยนำเสนอผลการศึกษาในรูปแบบตาราง ได้แก่ ค่า RMSE, mRMSE และ S.D. (ตารางที่ 4.6)

ตารางที่ 4.6 ผลการเปรียบเทียบประสิทธิภาพการพยากรณ์ของตัวแบบ กรณีศึกษาข้อมูลการเรียกร้องค่าสินไหมจากกรมธรรม์ประกันภัย (n=2,991)

RMSE							
ตัวแบบ	โฟลด์ 1	โฟลด์ 2	โฟลด์ 3	โฟลด์ 4	โฟลด์ 5	mRMSE	S.D.
QP	0.8039	0.8191	0.7969	0.8335	0.8120	0.8131	0.0142
CMP	0.8039	0.8309	0.7965	0.8350	0.8115	0.8156	0.0168
ZIP	0.8022	0.8028	0.7961	**0.8319**	0.8111	0.8089	**0.0134**
ZINB	**0.8017**	**0.8027**	**0.7953**	0.8326	**0.8108**	**0.8086**	0.0145
RF	0.8992	0.9066	0.9020	0.9066	0.9030	0.9035	0.0145

หมายเหตุ **ตัวหนา** แทนวิธีที่ให้ค่าเฉลี่ยของเกณฑ์ประสิทธิภาพของตัวแบบที่ดีที่สุด **(RMSE ต่ำสุด)**

จากตารางที่ 4.6 ค่า mRMSE คำนวณจากชุดทดสอบจำนวน 5 โฟลด์ (n = 599, 20%) ของตัวแบบทั้ง 5 พบว่า มีค่าใกล้เคียงกัน โดยต่างกันที่ทศนิยมตำแหน่งที่สองและสามเท่านั้น และพบว่า ตัวแบบการถดถอย ZINB เป็นตัวแบบที่ให้ค่า mRMSE ต่ำที่สุด ซึ่งมีค่าเท่ากับ 0.8086 บ่งบอกว่า ตัวแบบการถดถอย ZINB สามารถพยากรณ์ข้อมูลใหม่ได้ดีกว่าตัวแบบอื่น ๆ เล็กน้อย นอกจากนี้ตัวแบบที่ให้ค่า mRMSE ต่ำสุดรองลงมาคือ ตัวแบบการถดถอย ZIP, QP, CMP และเทคนิค RF โดยมีค่า mRMSE เท่ากับ 0.8089, 0.8131, 0.8156 และ 0.9035 ตามลำดับ ในขณะที่ตัวแบบการถดถอย ZIP ให้ค่าเบี่ยงเบนมาตรฐานต่ำที่สุด เท่ากับ 0.0134 บ่งบอกว่า ZIP มีประสิทธิภาพในการพยากรณ์ข้อมูลใหม่ใกล้เคียงกันสำหรับสถานการณ์กรณีศึกษาดังกล่าว

4.2.2 ผลการเปรียบเทียบประสิทธิภาพของตัวแบบสำหรับข้อมูลจำลอง

สำหรับการจำลองข้อมูลที่ 1 ใช้ค่าสัมประสิทธิ์การถดถอยจากตัวแบบการถดถอย ZICMP จากข้อมูลจริง (ตารางที่ 3.3) นั่นคือ เวกเตอร์ของพารามิเตอร์ $\delta_1(\boldsymbol{\beta},\boldsymbol{\gamma},\boldsymbol{\varphi})$ และสำหรับข้อมูลจำลองชุดที่ 2 ใช้เวกเตอร์ของพารามิเตอร์ $\delta_2(\boldsymbol{\beta},\boldsymbol{\gamma},\boldsymbol{\varphi}) = 1.5\times\delta_1(\boldsymbol{\beta},\boldsymbol{\gamma},\boldsymbol{\varphi})$ ในการจำลอง (ค่าสัมประสิทธิ์การถดถอยจากข้อมูลจริงเพิ่มขึ้น 50 เปอร์เซ็นต์) จากนั้นทำการเปรียบเทียบประสิทธิภาพการพยากรณ์ของตัวแบบด้วยวิธีเค-โฟลด์ตรวจสอบไขว้ โดยตัวแบบที่ใช้ ได้แก่ ตัวแบบการถดถอย QP, CMP, ZIP, ZINB และเทคนิค RF และคำนวณเกณฑ์การเปรียบเทียบประสิทธิภาพ mRMSE และ mS.D. ในข้อมูลชุดทดสอบที่ขนาดตัวอย่าง (n) = 250, 500, 1000, 3,000 และ 5,000 ทำซ้ำทั้งหมด 500 รอบ ซึ่งแสดงในรูปแบบภาพและตารางต่อไปนี้

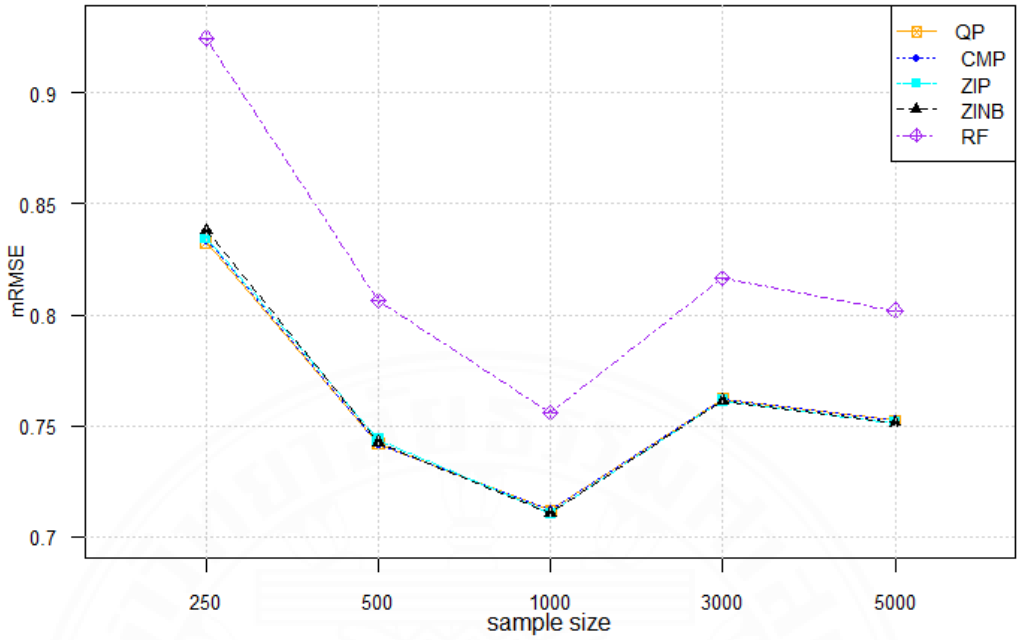
ตารางที่ 4.7 ค่า mRMSE ของตัวแบบการถดถอย QP, CMP, ZIP, ZINB และเทคนิค RF จากข้อมูลจำลองชุดที่ 1

n	$\bar{\lambda}_i$	$\bar{\nu}_i$	$\bar{\pi}_i$	mRMSE (mS.D.)				
				QP	CMP	ZIP	ZINB	RF
250	0.5269	0.6694	0.2942	**0.8327**	0.8341	0.8346	0.8380	0.9248
				(0.0765)	(0.0772)	(0.0774)	(0.0768)	(0.9248)
500	0.5286	0.6694	0.2903	0.7420	**0.7419**	0.7438	0.7424	0.8065
				(0.0629)	(0.0629)	(0.0627)	(0.0627)	(0.0601)
1,000	0.5307	0.6694	0.2937	0.7118	0.7121	**0.7105**	**0.7105**	0.7560
				(0.0287)	(0.0288)	(0.0274)	(0.0274)	(0.0331)
3,000	0.5311	0.6694	0.2968	0.7621	0.7622	**0.7613**	**0.7613**	0.8166
				(0.0324)	(0.0324)	(0.0322)	(0.0322)	(0.0319)
5,000	0.5280	0.6694	0.2953	0.7524	0.7525	0.7513	**0.7512**	0.8021
				(0.0198)	(0.0199)	(0.0198)	(0.0198)	(0.0191)

ตารางที่ 4.8 ค่า mRMSE ของตัวแบบการถดถอย QP, CMP, ZIP, ZINB และเทคนิค RF จากข้อมูลจำลองชุดที่ 2

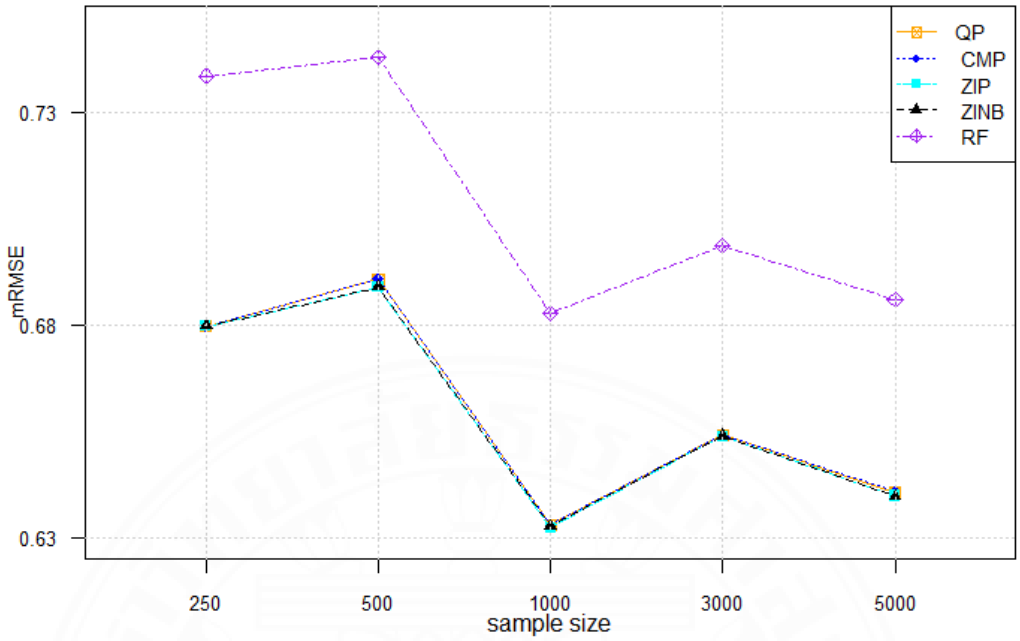
n	$\bar{\lambda}_i$	$\bar{\nu}_i$	$\bar{\pi}_i$	mRMSE (mS.D.)				
				QP	CMP	ZIP	ZINB	RF
250	0.3912	0.5477	0.2282	0.6856	0.6854	**0.6853**	**0.6853**	0.7398
				(0.0525)	(0.0524)	(0.0522)	(0.0522)	(0.0627)
500	0.3865	0.5477	0.2239	0.6965	0.6966	**0.6945**	**0.6945**	0.7479
				(0.0481)	(0.0491)	(0.0432)	(0.0432)	(0.0470)
1,000	0.3841	0.5477	0.2291	0.6331	0.6332	**0.6325**	0.6328	0.6833
				(0.0267)	(0.0267)	(0.0261)	(0.0260)	(0.0279)
3,000	0.3859	0.5477	0.2271	0.6541	0.6543	**0.6538**	0.6539	0.6985
				(0.0171)	(0.0171)	(0.0172)	(0.0171)	(0.0196)
5,000	0.3851	0.5477	0.2255	0.6406	0.6409	0.6398	**0.6397**	0.6860
				(0.0174)	(0.0174)	(0.0171)	(0.0171)	(0.0180)

จากตารางที่ 4.7 เมื่อจำลองชุดที่ 1 มีค่าสัมประสิทธิ์การถดถอยจากข้อมูลจริงเพิ่มขึ้น 50 เปอร์เซ็นต์เป็นข้อมูลจำลองชุดที่ 2 (ตารางที่ 4.8) พบว่า ค่า $\bar{\lambda}_i$, $\bar{\nu}_i$ และ $\bar{\pi}_i$ ลดลงทุกสถานการณ์ที่ขนาดตัวอย่างต่าง ๆ อีกทั้งพบว่า ประสิทธิภาพของแต่ละตัวแบบดีขึ้น เนื่องจากค่า mRMSE ของแต่ละตัวแบบมีค่าลดลงทุกกรณี เช่น ข้อมูลจำลองชุดที่ 1 ที่ขนาดตัวอย่างเท่ากับ 250 พบว่าค่า $\bar{\lambda}_i$, $\bar{\nu}_i$ และ $\bar{\pi}_i$ เท่ากับ 0.5269, 0.6694 และ 0.2942 ตามลำดับ นอกจากนี้ตัวแบบการถดถอย QP, CMP, ZIP, ZINB และเทคนิค RF ให้ค่า mRMSE เท่ากับ 0.8327, 0.8341, 0.8346, 0.8380 และ 0.9248 ตามลำดับและเมื่อสัมประสิทธิ์การถดถอยเพิ่มขึ้น 50 เปอร์เซ็นต์เป็นข้อมูลจำลองชุดที่ 2 พบว่าค่า $\bar{\lambda}_i$, $\bar{\nu}_i$ และ $\bar{\pi}_i$ ลดลงเป็น 0.3912, 0.5477 และ 0.2282 ตามลำดับ ซึ่งแต่ละตัวแบบให้ค่า mRMSE เท่ากับ 0.6856, 0.6854, 0.6853, 0.6853 และ 0.7398 ตามลำดับ



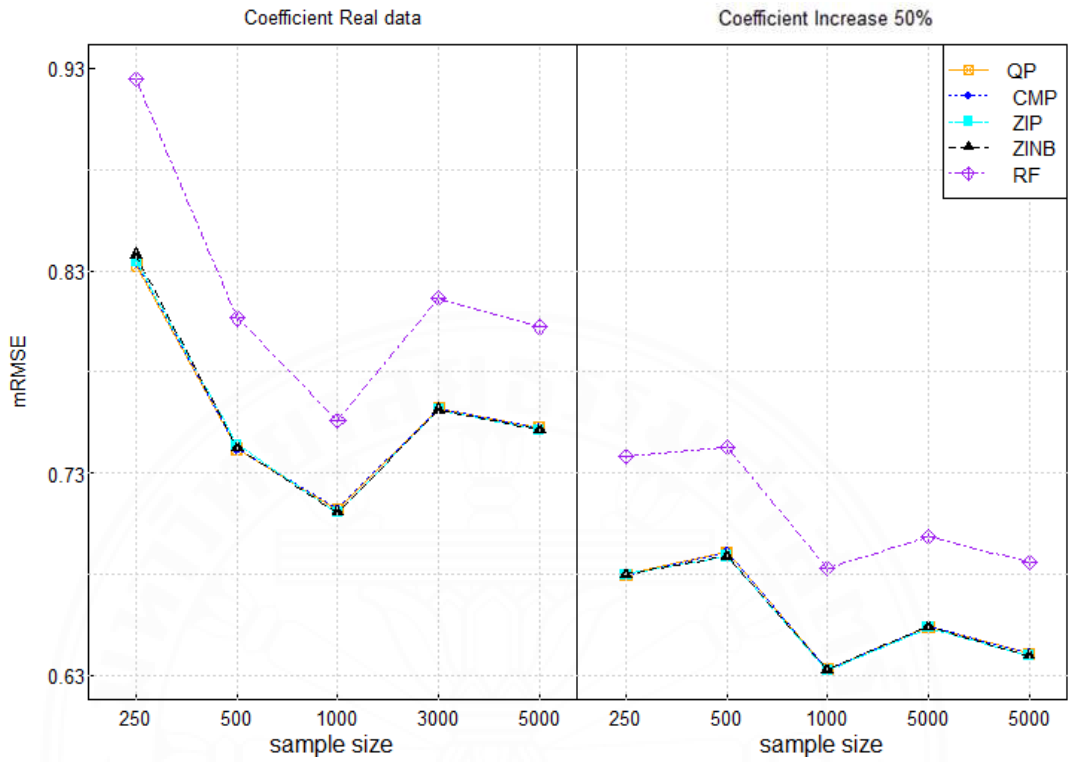
ภาพที่ 4.10 ค่า mRMSE ของตัวแบบทั้ง 5 เมื่อขนาดตัวอย่าง n = 250 500 1,000 3,000 และ 5,000 จากข้อมูลจำลองชุดที่ 1 (ลักษณะเส้น, สัญลักษณ์และสี คือ ตัวแบบ)

จากภาพที่ 4.10 และตารางที่ 4.7 เป็นข้อมูลจำลองที่ใช้สัมประสิทธิ์การถดถอยจากข้อมูลจริง (ข้อมูลจำลองชุดที่ 1) พบว่า เมื่อขนาดตัวอย่างเพิ่มขึ้นตั้งแต่ 500 ขึ้นไป แต่ละตัวแบบมีประสิทธิภาพที่ดีขึ้น เนื่องจาก mRMSE มีแนวโน้มลดลง โดยตัวแบบการถดถอย QP, CMP, ZIP และ ZINB มีประสิทธิภาพใกล้เคียงกัน ยกเว้นเทคนิค RF ที่มีประสิทธิภาพแย่ว่าตัวแบบอื่น ๆ เนื่องจากให้ค่า mRMSE สูงที่สุดทุกกรณี เมื่อขนาดตัวอย่างเป็น 5,000 พบว่า ZINB และ ZIP มีประสิทธิภาพดีกว่าตัวแบบอื่น ๆ เล็กน้อย และเมื่อขนาดตัวอย่างน้อย (250) พบว่า แต่ละตัวแบบมีประสิทธิภาพแตกต่างกัน โดยตัวแบบการถดถอย QP มีประสิทธิภาพดีที่สุด รองลงมาคือ ตัวแบบการถดถอย CMP, ZIP, ZINB และเทคนิค RF ซึ่งให้ค่า mRMSE เท่ากับ 0.8327, 0.8341, 0.8346, 0.8380 และ 0.9248 ตามลำดับ



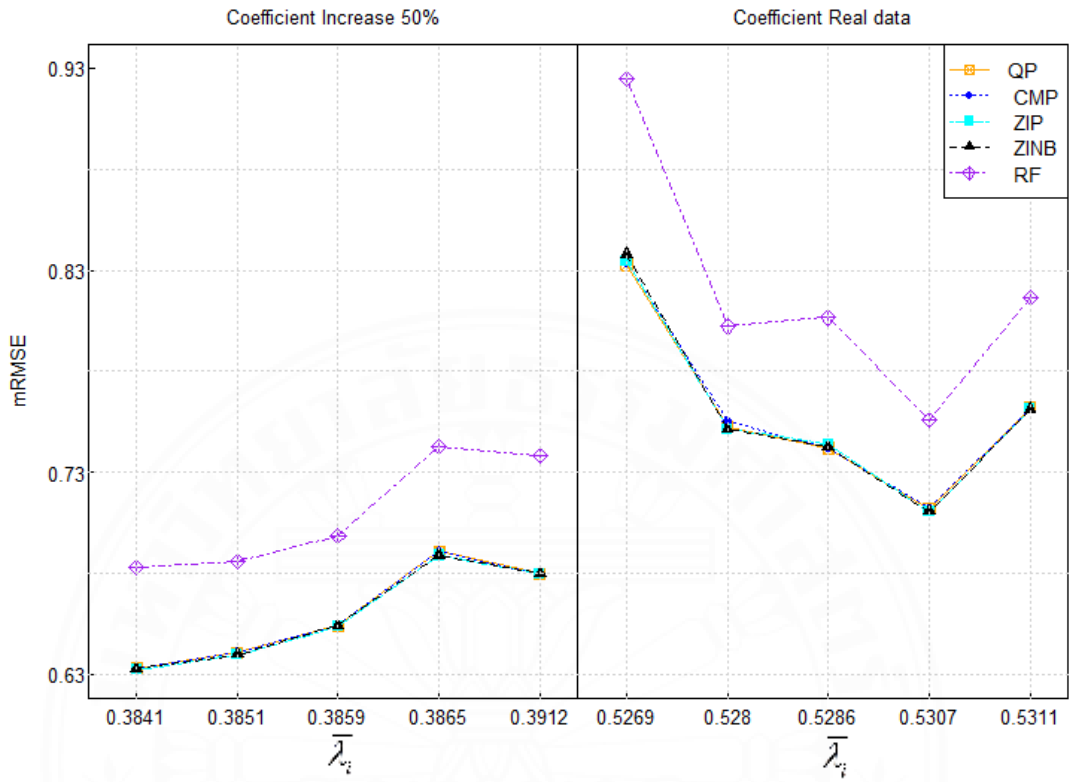
ภาพที่ 4.11 ค่า mRMSE ของตัวแบบทั้ง 5 เมื่อขนาดตัวอย่าง n = 250 500 1,000 3,000 และ 5,000 จากข้อมูลจำลองที่ใช้ค่าสัมประสิทธิ์การถดถอยของข้อมูลจริงเพิ่มขึ้น 50 เปอร์เซ็นต์ (ลักษณะเส้น, สัญลักษณ์และสี คือ ตัวแบบ)

จากภาพที่ 4.11 และตารางที่ 4.8 เป็นข้อมูลจำลองที่ใช้สัมประสิทธิ์การถดถอยจากข้อมูลจริงเพิ่มขึ้น 50 เปอร์เซ็นต์ (ข้อมูลจำลองชุดที่ 2) พบว่า เมื่อขนาดตัวอย่างเพิ่มขึ้นจาก 500 เป็น 5,000 แต่ละตัวแบบมีประสิทธิภาพดีขึ้น เนื่องจาก mRMSE มีแนวโน้มลดลง โดยตัวแบบการถดถอย QP, CMP, ZIP และ ZINB มีประสิทธิภาพใกล้เคียงกัน ยกเว้น RF ที่มีประสิทธิภาพแย่ว่าตัวแบบอื่น ๆ เนื่องจากให้ค่า mRMSE สูงที่สุดทุกกรณี และที่ขนาดตัวอย่างเท่ากับ 5,000 พบว่าตัวแบบ ZINB และ ZIP มีประสิทธิภาพดีกว่าตัวแบบอื่น ๆ เล็กน้อย เมื่อขนาดตัวอย่างน้อย (500) แต่ละตัวแบบมีประสิทธิภาพแตกต่างกัน โดยตัวแบบการถดถอย ZINB และ ZIP มีประสิทธิภาพดีกว่าตัวแบบอื่น ๆ ในขณะที่ขนาดตัวอย่างเท่ากับ 250 ประสิทธิภาพของแต่ละตัวแบบใกล้เคียงกัน



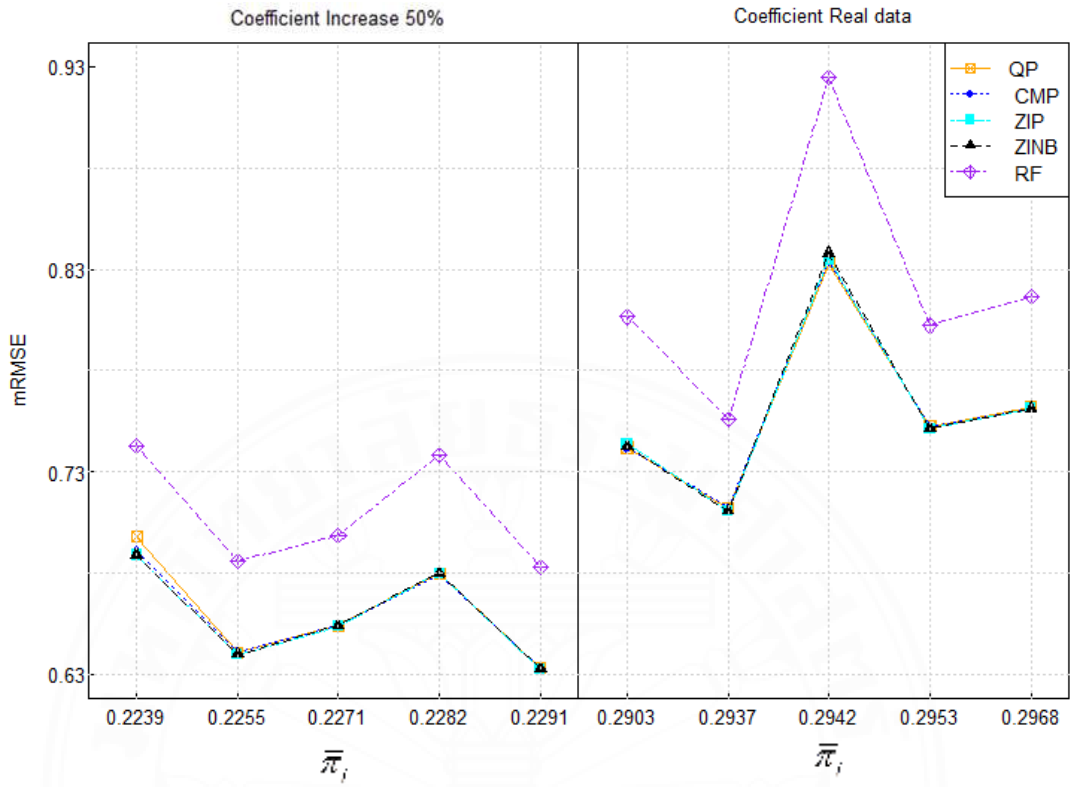
ภาพที่ 4.12 ค่า mRMSE ของตัวแบบทั้ง 5 เมื่อขนาดตัวอย่าง n = 250 500 1,000 3,000 และ 5,000 ของข้อมูลจำลองชุดที่ 1 เปรียบเทียบกับข้อมูลจำลองชุดที่ 2 (ลักษณะเส้น, สัญลักษณ์และสี คือ ตัวแบบ)

จากภาพที่ 4.12 เมื่อข้อมูลจำลองชุดที่ 1 (รูปซ้าย) มีค่าสัมประสิทธิ์เพิ่มขึ้น 50 เปอร์เซ็นต์เป็นข้อมูลจำลองชุดที่ 2 (รูปขวา) พบว่า แต่ละตัวแบบมีประสิทธิภาพดีขึ้นทุกกรณีและเมื่อขนาดตัวอย่างเพิ่มขึ้นจาก 250 เป็น 5,000 พบว่าประสิทธิภาพของแต่ละตัวแบบเพิ่มขึ้นเนื่องจาก mRMSE มีแนวโน้มลดลงทุกกรณีเช่นเดียวกับข้อมูลจำลองชุดที่ 1



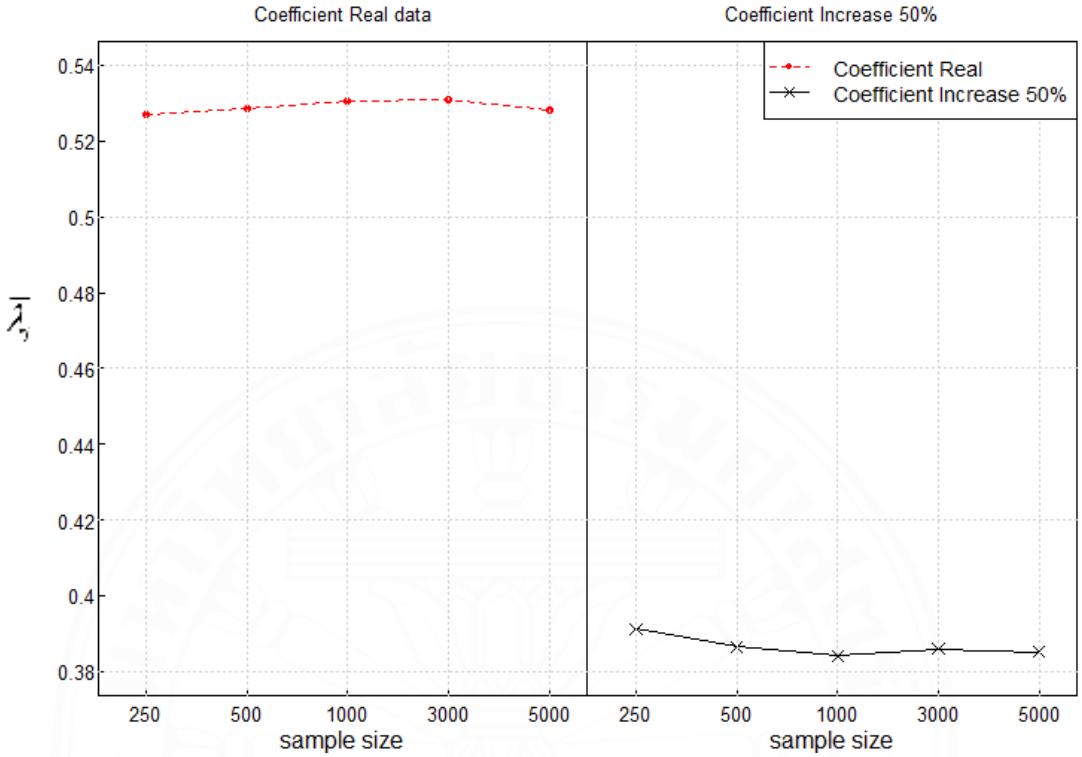
ภาพที่ 4.13 ค่า mRMSE ของตัวแบบทั้ง 5 ที่ค่า $\overline{\lambda_i}$ ต่าง ๆ ของข้อมูลจำลองชุดที่ 1 เปรียบเทียบกับข้อมูลจำลองชุดที่ 2 (ลักษณะเส้น, สัญลักษณ์และสี คือ ตัวแบบ)

จากภาพที่ 4.13 เมื่อข้อมูลจำลองชุดที่ 1 (รูปขวา) มีค่าสัมประสิทธิ์การถดถอยเพิ่มขึ้น 50 เปอร์เซ็นต์เป็นข้อมูลจำลองชุดที่ 2 (รูปซ้าย) พบว่า $\overline{\lambda_i}$ มีค่าลดลง และพบว่าแต่ละตัวแบบมีประสิทธิภาพดีขึ้นเมื่อเทียบกับข้อมูลชุดที่ 1 ทุกกรณี



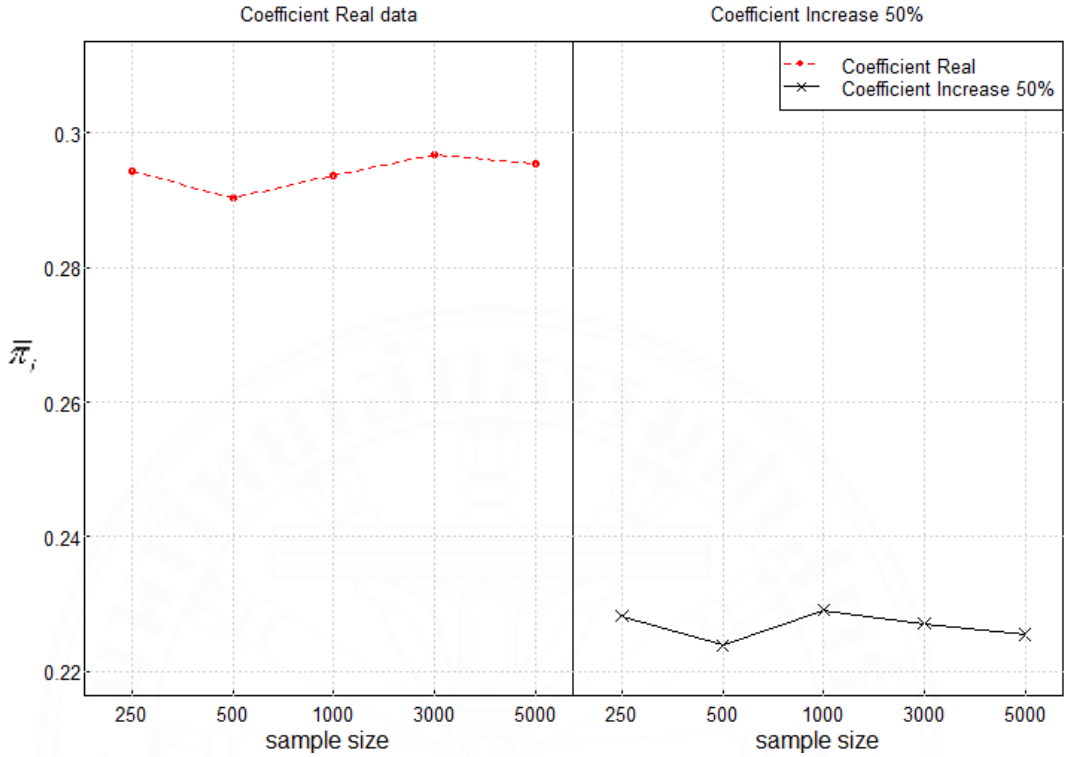
ภาพที่ 4.14 ค่า mRMSE ของตัวแบบทั้ง 5 ที่ค่า $\bar{\pi}_i$ ต่าง ๆ ของข้อมูลจำลองชุดที่ 1 เปรียบเทียบกับข้อมูลจำลองชุดที่ 2 (ลักษณะเส้น, สัญลักษณ์และสี คือ ตัวแบบ)

จากภาพที่ 4.14 เมื่อข้อมูลจำลองชุดที่ 1 (รูปขวา) มีค่าสัมประสิทธิ์การถดถอยเพิ่มขึ้น 50 เปอร์เซ็นต์เป็นข้อมูลจำลองชุดที่ 2 (รูปซ้าย) พบว่า การเพิ่มขึ้นของสัมประสิทธิ์การถดถอย มีผลให้ $\bar{\pi}_i$ มีค่าลดลง และพบว่าแต่ละตัวแบบมีประสิทธิภาพดีขึ้นเมื่อเทียบกับข้อมูลจำลองชุดที่ 1



ภาพที่ 4.15 ค่า $\overline{\lambda_i}$ ที่ขนาดตัวอย่างต่าง ๆ ของข้อมูลจำลองชุดที่ 1 เปรียบเทียบกับข้อมูลจำลองชุดที่ 2 (ลักษณะเส้น, สัญลักษณ์และสี คือ การใช้ค่าสัมประสิทธิ์การถดถอยในการจำลองข้อมูล)

จากภาพที่ 4.15 ในข้อมูลจำลองชุดที่ 1 (รูปซ้าย) พบว่า การเพิ่มขึ้นของขนาดตัวอย่างส่งผลให้ค่าเฉลี่ยของพารามิเตอร์ค่าเฉลี่ยมีแนวโน้มลดลง เมื่อค่าสัมประสิทธิ์การถดถอยเพิ่มขึ้น 50 เปอร์เซ็นต์เป็นข้อมูลจำลองชุดที่ 2 (รูปขวา) พบว่า $\overline{\lambda_i}$ มีค่าลดลงที่ขนาดตัวอย่างต่าง ๆ ทุกกรณีและมีพฤติกรรมเช่นเดียวกับข้อมูลจำลองชุดที่ 1



ภาพที่ 4.16 ค่า $\bar{\pi}_i$ ที่ขนาดตัวอย่างต่าง ๆ ของข้อมูลจำลองชุดที่ 1 เปรียบเทียบกับข้อมูลจำลองที่ชุดที่ 2 (ลักษณะเส้น, สัญลักษณ์และสี คือ การใช้ค่าสัมประสิทธิ์การถดถอยในการจำลองข้อมูล)

จากภาพที่ 4.16 ในข้อมูลจำลองชุดที่ 1 (รูปซ้าย) พบว่า การเพิ่มขึ้นของขนาดตัวอย่างส่งผลให้ค่าเฉลี่ยของพารามิเตอร์ความน่าจะเป็นที่จะเกิดศูนย์มีแนวโน้มลดลง เมื่อค่าสัมประสิทธิ์การถดถอยเพิ่มขึ้น 50 เปอร์เซ็นต์เป็นข้อมูลจำลองชุดที่ 2 (รูปขวา) พบว่า $\bar{\pi}_i$ มีค่าลดลงที่ขนาดตัวอย่างต่าง ๆ ทุกกรณีและมีพฤติกรรมเช่นเดียวกับข้อมูลจำลองชุดที่ 1

บทที่ 5
สรุปผลการวิจัยและข้อเสนอแนะ

จากที่มาและความสำคัญของปัญหา ในทางปฏิบัติข้อมูลจำนวนนับมักจะมีปัญหาการกระจายและค่าศูนย์เฟ้อ จึงทำให้ผู้วิจัยสนใจที่จะเปรียบเทียบประสิทธิภาพของตัวแบบต่าง ๆ ในการพยากรณ์ของข้อมูลดังกล่าว ด้วยเหตุผลข้างต้นงานวิจัยนี้จึงมีวัตถุประสงค์เพื่อเปรียบเทียบตัวแบบพยากรณ์ที่ใช้วิเคราะห์ข้อมูลจำนวนนับที่มีปัญหาการกระจายและค่าศูนย์เฟ้อ โดยแบ่งเป็นสองกรณี คือ กรณีจำนวนนับที่มีลักษณะการกระจายต่ำกว่าเกณฑ์ ซึ่งจะพิจารณา 2 ตัวแบบ ได้แก่ ตัวแบบการถดถอย CMP และเทคนิค RF และกรณีจำนวนนับที่มีลักษณะการกระจายเกินเกณฑ์ ซึ่งจะพิจารณา 5 ตัวแบบ ได้แก่ ตัวแบบการถดถอย QP, CMP, ZIP, ZINB และเทคนิค RF และเพื่อศึกษาอิทธิพลของการเพิ่มขึ้นของสัมประสิทธิ์การถดถอยและขนาดตัวอย่าง (n) ในสถานการณ์ต่าง ๆ การพิจารณาประสิทธิภาพของตัวแบบพยากรณ์ใช้ค่าเฉลี่ยของรากที่สองของค่าคลาดเคลื่อนกำลังสองเฉลี่ย (mRMSE) ถ้าตัวแบบใดที่ให้ค่าเฉลี่ยของรากที่สองของค่าคลาดเคลื่อนกำลังสองเฉลี่ยต่ำกว่า ตัวแบบนั้นจะมีประสิทธิภาพในการพยากรณ์ดีกว่าตัวแบบอื่น ๆ จากผลการวิจัยสามารถสรุปผลและให้ข้อเสนอแนะได้ดังนี้

5.1 สรุปผลการวิจัย

5.1.1 กรณีข้อมูลที่มีการกระจายต่ำกว่าเกณฑ์และค่าศูนย์เฟ้อ

การศึกษาข้อมูลสถิติอุบัติเหตุทางถนนที่ส่งผลให้มีผู้เสียชีวิตทั่วประเทศไทย เดือนเมษายน ปี พ.ศ. 2558 มีจำนวนทั้งหมด 4,666 เหตุการณ์ ความน่าจะเป็นที่จะเกิดศูนย์เท่ากับ 0.8221 จากการคัดเลือกตัวแปรอิสระด้วยตัวแบบการถดถอย ZICMP ที่เหมาะสมที่สุด พบว่ามีตัวแปรอิสระที่มีนัยสำคัญทางสถิติและเป็นตัวแปรเชิงคุณภาพทั้งหมด ได้แก่ ประเภทสายทาง, ลักษณะสายทาง, พื้นผิวถนนและช่วงเวลา และทดสอบประสิทธิภาพการพยากรณ์ด้วยวิธีเค-โฟลด์ตรวจสอบไขว้ที่ K = 5 พบว่า ตัวแบบการถดถอย CMP และเทคนิค RF มีประสิทธิภาพใกล้เคียงกัน โดยตัวแบบให้ค่าเฉลี่ยของรากของค่าคลาดเคลื่อนกำลังสองเฉลี่ย เมื่อพิจารณาทั้ง 5 โฟลด์ใกล้เคียงกันทั้งหมดจากสถานการณ์จำลอง สรุปได้ว่า เมื่อจำลองข้อมูลโดยใช้ค่าสัมประสิทธิ์การถดถอยจากข้อมูลจริงเพิ่มขึ้นเป็น 50 เปอร์เซ็นต์ (ข้อมูลจำลองชุดที่ 2) พบว่า การเพิ่มขึ้นของสัมประสิทธิ์การ

ถดถอยส่งผลให้ค่าเฉลี่ยพารามิเตอร์ค่าเฉลี่ย, พารามิเตอร์การกระจายและพารามิเตอร์ความน่าจะเป็นที่จะเกิดศูนย์เพิ่มขึ้น อีกทั้งส่งผลให้ทุกตัวแบบมีประสิทธิภาพที่ดีขึ้นทุกกรณี และในกรณีข้อมูลจำลองโดยใช้สัมประสิทธิ์การถดถอยจากข้อมูลจริง (ข้อมูลจำลองชุดที่ 1) พบว่าการเพิ่มขึ้นของขนาดตัวอย่างส่งผลให้ค่าเฉลี่ยของพารามิเตอร์ค่าเฉลี่ยมีแนวโน้มลดลง ในขณะที่การเพิ่มขึ้นของขนาดตัวอย่างส่งผลต่อค่าเฉลี่ยของพารามิเตอร์ความน่าจะเป็นที่จะเกิดศูนย์มีแนวโน้มเพิ่มขึ้น นอกจากนี้ พบว่า เมื่อขนาดตัวอย่างน้อย ตัวแบบมีประสิทธิภาพแตกต่างกัน โดยเฉพาะขนาดตัวอย่างเท่ากับ 250 ตัวแบบการถดถอย CMP มีประสิทธิภาพดีกว่าเทคนิค RF ทุกกรณี และเมื่อขนาดตัวอย่างเพิ่มขึ้นประสิทธิภาพของตัวแบบการถดถอย CMP และเทคนิค RF ดีขึ้นและมีประสิทธิภาพใกล้เคียงกัน โดยเฉพาะเมื่อขนาดตัวอย่างเพิ่มขึ้นเป็น 3,000 และ 5,000 สำหรับข้อมูลจำลองชุดที่ 2 พบว่า มีพฤติกรรมแบบเดียวกัน ซึ่งผลลัพธ์การจำลองจากข้อมูลทั้งสองชุดสอดคล้องกับผลสรุปของข้อมูลอุบัติเหตุทางถนนที่ส่งผลให้มีผู้เสียชีวิตข้างต้น เนื่องจากเป็นข้อมูลที่มีขนาดตัวอย่างมากเช่นเดียวกัน

5.1.2 กรณีข้อมูลที่มีการกระจายเกินเกณฑ์และค่าศูนย์เฟ้อ

การศึกษาข้อมูลการเรียกร้องค่าสินไหมทดแทนของลูกค้าจากกรมธรรม์ประกันภัยรถยนต์จากบริษัทประกันภัยแห่งหนึ่งโดยใช้ข้อมูลจำนวน 2,991 กรมธรรม์ในปี พ.ศ. 2560 ความน่าจะเป็นที่จะเกิดศูนย์ เท่ากับ 0.6731 จากการคัดเลือกตัวแปรอิสระด้วยตัวแบบการถดถอย ZICMP ที่เหมาะสมที่สุด พบว่ามีตัวแปรอิสระที่มีนัยสำคัญทางสถิติซึ่งเป็นตัวแปรเชิงปริมาณทั้งหมด ได้แก่ เบี้ยประกันภัยและส่วนลดประวัติดี และการทดสอบประสิทธิภาพในการพยากรณ์ด้วยวิธีเค-โฟลด์ตรวจสอบไขว้ พบว่า ตัวแบบการถดถอย QP, CMP, ZIP และ ZINB มีประสิทธิภาพใกล้เคียงกัน โดยตัวแบบดังกล่าวให้ค่าเฉลี่ยของรากของค่าคลาดเคลื่อนกำลังสองเฉลี่ยเมื่อพิจารณาทั้ง 5 โฟลด์ ใกล้เคียงกัน ยกเว้นเทคนิค RF มีประสิทธิภาพต่ำกว่าตัวแบบอื่น ๆ

จากสถานการณ์จำลอง สรุปได้ว่า เมื่อจำลองข้อมูลโดยใช้ค่าสัมประสิทธิ์การถดถอยจากข้อมูลจริงเพิ่มขึ้นเป็น 50 เปอร์เซ็นต์ (ข้อมูลจำลองชุดที่ 2) พบว่า การเพิ่มขึ้นของสัมประสิทธิ์การถดถอยส่งผลให้ค่าเฉลี่ยพารามิเตอร์ค่าเฉลี่ย, พารามิเตอร์การกระจายและพารามิเตอร์ความน่าจะเป็นที่จะเกิดศูนย์ลดลง อีกทั้งส่งผลให้ทุกตัวแบบมีประสิทธิภาพที่ดีขึ้นทุกกรณี และในกรณีข้อมูลจำลองที่ใช้สัมประสิทธิ์การถดถอยจากข้อมูลจริง (ข้อมูลจำลองชุดที่ 1) พบว่า การเพิ่มขึ้นของขนาดตัวอย่างส่งผลให้ค่าเฉลี่ยของพารามิเตอร์ค่าเฉลี่ยและความน่าจะเป็นที่จะเกิดศูนย์มีแนวโน้มลดลง และเมื่อขนาดตัวอย่างน้อย ตัวแบบมีประสิทธิภาพแตกต่างกัน โดยเฉพาะขนาดตัวอย่างเท่ากับ 250 ตัวแบบการถดถอย QP มีประสิทธิภาพดีที่สุด รองลงมาคือ ตัวแบบการถดถอย CMP, ZIP, ZINB และเทคนิค RF ตามลำดับ จากการทบทวนวรรณกรรม พบว่า สอดคล้องกับ

งานวิจัยของ นวพรรณ เชื้ออ่ำ, บุญอ้อม โฉมที และ อภิญญา หิรัญวงษ์ (2561) เป็นการเปรียบเทียบความเหมาะสมของตัวแบบการถดถอย QP และ ZINB โดยใช้ข้อมูลการเกิดอุบัติเหตุบนท้องถนนที่เป็นถนนทางหลวง เดือนมกราคมถึงเดือนธันวาคม ปี พ.ศ. 2559 ซึ่งเป็นขนาดตัวอย่างเล็ก พบว่า ตัวแบบการถดถอย QP มีความเหมาะสมมากกว่า ZINB เกือบทุกกรณี และ Ma และ Yuan (2018) เปรียบเทียบประสิทธิภาพของตัวแบบในการพยากรณ์การเกิดอุบัติเหตุบนท้องถนน มีขนาดตัวอย่างเท่ากับ 200 พบว่า ตัวแบบการถดถอย ZINB มีประสิทธิภาพดีกว่า NB, RF และ P ตามลำดับ นอกจากนี้เมื่อขนาดตัวอย่างเพิ่มขึ้นตั้งแต่ 1,000 ขึ้นไป ประสิทธิภาพของทุกตัวแบบดีขึ้นและมีประสิทธิภาพใกล้เคียงกัน โดยเทคนิค RF มีประสิทธิภาพแย่กว่าตัวแบบอื่น ๆ ทุกกรณี สำหรับข้อมูลจำลองชุดที่ 2 พบว่ามีพฤติกรรมแบบเดียวกัน ซึ่งผลลัพธ์จากข้อมูลจำลองทั้งสองชุดสอดคล้องกับผลการวิเคราะห์ข้อมูลการเรียกร้องค่าสินไหมทดแทนของลูกค้าจากกรมธรรม์ประกันภัยรถยนต์ ปี พ.ศ. 2560 เนื่องจากเป็นข้อมูลที่มีขนาดตัวอย่างมากเช่นเดียวกัน และสอดคล้องกับงานวิจัยของ Lord, Guikema และ Geedipally (2008) เปรียบเทียบประสิทธิภาพการพยากรณ์โดยใช้ข้อมูลการชนของรถยนต์บนท้องถนน โดยข้อมูลชุดแรก มี 868 เหตุการณ์และข้อมูลชุดที่สอง มี 3,220 เหตุการณ์ พบว่า ตัวแบบการถดถอย CMP และ NB มีประสิทธิภาพใกล้เคียงกัน

5.1.3 การพิจารณาเลือกตัวแบบ

กรณีการกระจายต่ำกว่าเกณฑ์และค่าศูนย์เฟ้อ เมื่อขนาดตัวอย่างน้อย โดยเฉพาะขนาดตัวอย่างเท่ากับ 250 ตัวแบบการถดถอย CMP มีประสิทธิภาพดีกว่าเทคนิค RF ทุกกรณี และพบว่าเมื่อข้อมูลมีขนาดตัวอย่างมาก ตัวแบบการถดถอย CMP และเทคนิค RF จะมีประสิทธิภาพใกล้เคียงกัน

กรณีการกระจายเกินเกณฑ์และค่าศูนย์เฟ้อ โดยตัวแบบการถดถอย QP และ CMP จะมีประสิทธิภาพดีกว่าตัวแบบอื่น ๆ ที่ขนาดตัวอย่างน้อย

ตัวแบบ ZIP และ ZINB จะมีประสิทธิภาพดีกว่าตัวแบบอื่น ๆ เล็กน้อย เมื่อข้อมูลมีขนาดตัวอย่างเพิ่มขึ้น โดยเฉพาะที่ขนาดตัวอย่างเท่ากับ 5,000

ตัวแบบการถดถอย QP, CMP, ZIP และ ZINB จะมีประสิทธิภาพใกล้เคียงกัน เมื่อข้อมูลมีขนาดตัวอย่างมาก ยกเว้นเทคนิค RF มีประสิทธิภาพต่ำกว่าตัวแบบอื่น ๆ ทุกกรณี

จากการศึกษาพบว่า การที่ข้อมูลมีขนาดตัวอย่างน้อยและมีจำนวนตัวแปรอิสระน้อยจะส่งผลให้เทคนิค RF มีประสิทธิภาพในการพยากรณ์ต่ำกว่าตัวแบบอื่น ๆ (ดังกรณีการกระจายเกินเกณฑ์และค่าศูนย์เฟ้อภายใต้ตัวแปรอิสระที่มีนัยสำคัญทางสถิติ 2 ตัว) นอกจากนี้พบว่า เมื่อขนาดตัวอย่างเพิ่มขึ้นส่งผลให้ประสิทธิภาพของเทคนิค RF ดีขึ้นและเมื่อมีจำนวนตัวแปรอิสระมากขึ้น

จะทำให้เทคนิค RF มีประสิทธิภาพดียิ่งขึ้น (ดังกรณีการกระจายต่ำกว่าเกณฑ์และค่าศูนย์เฟ้อภายใต้ตัวแปรอิสระที่มีนัยสำคัญทางสถิติ 3 ตัวข้างต้น)

5.2 ข้อจำกัดและข้อเสนอแนะ

1. การใช้วิธีเค-โฟลด์ตรวจสอบไขว้ เพื่อเปรียบเทียบประสิทธิภาพการพยากรณ์โดยใช้ตัวแบบที่แตกต่างกัน การเลือกค่า K จะต้องคำนึงถึงขนาดตัวอย่างในแต่ละโฟลด์ ในงานวิจัยนี้ เลือก K = 5 เท่านั้น การจำลองข้อมูลจึงเลือกขนาดตัวอย่างเริ่มต้นที่ 250 เพื่อรองรับเทคนิค RF เนื่องจากเทคนิค RF ไม่สามารถวิเคราะห์ได้หากข้อมูลมีขนาดตัวอย่างน้อยเกินไป ดังนั้นในการศึกษาครั้งต่อไปควรพิจารณาเลือกค่า K อื่น ๆ ให้มีความเหมาะสมสำหรับตัวแบบ

2. การจำลองข้อมูลภายใต้แจกแจงคอนเวย์แม็กซ์เวลล์ปัวซงค่าศูนย์เฟ้อร่วมกับวิธีเค-โฟลด์ตรวจสอบไขว้ต้องตรวจสอบเงื่อนไขเกี่ยวกับค่าการกระจาย หากข้อมูลในโฟลด์ใดไม่เป็นไปตามเงื่อนไข ไม่ควรนำมาพิจารณาในการจำลองข้อมูล โดยเฉพาะกรณีการกระจายต่ำกว่าเกณฑ์หรือการกระจายเกินเกณฑ์ที่มีค่าการกระจายเข้าใกล้หนึ่ง การใช้วิธีเคโฟลด์ตรวจสอบไขว้อาจทำให้ชุดข้อมูลในโฟลด์นั้นไม่มีคุณสมบัติตามที่ต้องการ

3. สำหรับงานวิจัยต่อไป อาจเพิ่มจำนวนตัวแปรอิสระ มีการขยายขอบเขตของพารามิเตอร์ให้ครอบคลุมมากยิ่งขึ้น นำตัวแบบหรือวิธีการอื่น ๆ มาร่วมในการวิเคราะห์ เช่น ตัวแบบการเรียนรู้ของเครื่องอื่น ๆ ตัวแบบการถดถอย ZICMP และตัวแบบการถดถอยปัวซงนัยทั่วไปค่าศูนย์เฟ้อ (Zero-inflated Generalized Poisson regression: ZIGP) เป็นต้น

รายการอ้างอิง

หนังสือและบทความในหนังสือ

บุญเสริม กิจศิริกุล. (2548). ปัญญาประดิษฐ์ Artificial Intelligence. กรุงเทพมหานคร: จุฬาลงกรณ์มหาวิทยาลัย.ศูนย์วิจัยอุบัติเหตุแห่งประเทศไทย. (2552). การพัฒนาแบบจำลองการเกิดอุบัติเหตุโครงการต่อเนื่องศูนย์วิจัยอุบัติเหตุแห่งประเทศไทยเพื่อพัฒนาและเผยแพร่องค์ความรู้ด้านความปลอดภัยทางถนน. จุฬาลงกรณ์มหาวิทยาลัย. สืบค้นจาก https://www.cp.eng.chula.ac.th/books/ai/

James, G., Witten, D., Hastie, T. & Tibshirani, R. (2013). An Introduction to Statistical Learning: with Applications in R, Springer Texts in Statistics". Springer Science Business Media New York 2013.

บทความวารสาร

สุภัทรา โนนคล้อ, ชนพงษ์ โรจนวรฤทธิ์, สุทธิ เจริญพิทักษ์, ชูเกียรติ วิวัฒน์วงศ์เกษม และ ธีรวัจน์ ปัญญานะ. (2561). ปัจจัยเสี่ยงทางคลินิกที่มีผลต่อการเกิดปอดบวมของผู้ป่วยโรคหลอดเลือดสมองที่เข้ารับการรักษาในโรงพยาบาลน่าน ประเทศไทย. *การประชุมวิชาการสาธารณสุขแห่งชาติ ครั้งที่ 16*, 339-348.

อดิเทพ ไชยวรรณ, วสันต์ บุญโฮ้ และ พิษณุ ทองขาว. (2555). การวิเคราะห์ปัจจัยเสี่ยงของการผลิตสินค้าบกพร่องในโรงงานผลิตชิ้นส่วนรถยนต์ โดยใช้ตัวแบบเชิงเส้นนัยทั่วไป (GLM). *วารสารการประชุมวิชาการแห่งชาติ มหาวิทยาลัยเกษตรศาสตร์ วิทยาเขตกำแพงแสน ครั้งที่ 9.*

Alqawba, M. & Diawara, N. (2020). Copula-based Markov zero-inflated count time series models with application. *Journal of Applied Statistics.* https://www.tandfonline.com/doi/abs/10.1080/02664763.2020.1748581

Anan, O., Böhning, D., & Maruotti, A. (2016). Population size estimation and heterogeneity in capture-recapture data: a linear regression estimator based on the Conway-Maxwell-Poisson distribution. *Statistical Methods and Applications*, doi:10.1007

Annafari, M. T. (2010). An Empirical Analysis of the Factors Determining Multiple Subscriptions in the Swedish Mobile Phone Market. 2010 Ninth International

Conference on Mobile Business and 2010 Ninth Global Mobility Roundtable (ICMB-GMR). doi:10.1109/icmb-gmr.2010.43

Ayati, E., & Abbasi, E. (2014). Modeling Accidents on Mashhad Urban Highways. *Journal of Safety Science and Technology*, 4, 22-35.

Breiman, L. (2001). Random forests. *Machine Learning*, Vol. 45 (1). 5–32.

Boatwright, P., Borle, S. & Kadane J.B. (2003). A Model of the Joint Distribution of Purchase Quantity and Timing. *Journal of the American Statistical Association*, 98(463), 564-572.

Cameron, A.C., & Trivedi P.K. (1986). Econometric Models Based On Count Data: Comparisons of Some Estimators and Tests. *Journal of Applied Econometrics*, 1, 29-54.

Cameron., A.C.& Trivedi., P.K. (1990). Regression-based Tests for Over-dispersion in the Poisson Model", *Journal of Econometrics*, 46, 347–364.

Chang, L.Y. & Chen, W.C. (2005). Data mining of tree-based models to analyze freeway accident frequency. *Journal of Safety Research*, 36, 365-375. doi:10.1016/j.jsr.2005.06.013

Conway, R.W. & Maxwell, W.L. (1962). A queuing model with state dependent service rates. *Journal of Industrial Engineering*, 12:132–136.

Czado, C. & Santner, T. J. (1992). The Effect of Link Misspecification on Binary Regression Inference. *Journal of Statistics Planning and Inference*, 33, 213-231.

Grömping, U., (2009). Variable Importance Assessment in Regression: Linear Regression versus Random Forest. *The American Statistician*, 63:4, 308-319.

Guikema, S.D. & Coffelt J. (2008). A flexible count data regression model for risk analysis. *Risk Analysis*, 28(1), 213-223.

Huang, A. (2017). Mean-parametrized Conway–Maxwell–Poisson regression models for dispersed counts. Statistical Modelling: *An International Journal*, 17(6), 359–380. doi:10.1177/1471082x17697749

Hauer, E. (2001). Overdispersion in modelling accidents on road sections and in empirical Bayes estimation. *Accident Analysis & Prevention*, 33(6), 799–808.

Krishnaveni, S., & Hemalatha, M. (2011). A Perspective Analysis of Traffic Accident using Data Mining Techniques. *International Journal of Computer Applications*, 0975 – 8887.

Lambert, D. (1992). Zero-infated Poisson regression with application to defects in manufacturing. *Technometrics*, 34(1), 1.

Liaw, A., & Wiene, M. (2002). Classification and Regression by Random Forest. *Machine Learning, 2(3)*.

Lord, D., Guikema, S.D., & Geedipally S.R. (2008). Application of the Conway–Maxwell–Poisson generalized linear model for analyzing motor vehicle crashes. *Accident Analysis and Prevention*, 40(3), 1123–1134. doi:10.1016/j.aap.2007.12.003

Ma, W. & Yuan, Z. (2018). Analysis and Comparison of Traffic Accident Regression Prediction Model. *3rd International Conference on Electromechanical Control Technology and Transportation (ICECTT 2018)*, 364-369

Malekipirbazari, M. & Aksakalli, V. (2008). Risk assessment in social lending via random forests. *Expert Syst. Appl. Vol. 42 (10)*, 4621–4631. 12.

Melkersson, M. & Rooth, D. (2000). Modeling Female Fertility Using Inflated Count Data Models. *Populatio Economics*. 13 189–203. 13.

Miaou, S.P., Lord, D. (2003). Modeling traffic crash flow relationships for intersections: Dispersion parameter, functional form, and full Bayes versus empirical Bayes. *Transportation Research Record 1840*, 31–40.

Momeni, F. (2011).The Generalized Power Series Distributions and their Application. *Mathematics and Computer Science 2*, No.4, 691-697.

Mwalili, S. M., Lesaffre, E., & Declerck, D. (2007). The zero-inflated negative binomial regression model with correction for misclassification: An example in caries research. *Statistical Methods in Medical Research*, 17(2), 123-139.

Payne, E. H., Gebregziabher, M., Hardin, J. W., Ramakrishnan, V., & Egede, L. E. (2017). An empirical approach to determine a threshold for assessing over dispersion in Poisson and negative binomial models for count data. *Communications in Statistics - Simulation and Computation*, 47(6), 1722–1738. doi:10.1080/03610918.2017.1323223

Potts, J.M., & Elith, J. (2006). Comparing species abundance models. *Ecological Modeling*, 199, 153-163.

Prasetijo, J., Musa, W. Z., Mohd Jawi, Z., Zainal, Z. F., Hamid, N. B., Subramaniyan, A., & Mohd Hafzi Md, I. (2020). Vehicle Road Accident Prediction Model along Federal Road FT050 Kluang-A/Hitam-B/Pahat Route Using Excess Zero Data. IOP Conference Series: Materials Science and Engineering, 852, 012144. doi:10.1088/1757-899x/852/1/012144

Ridout, M. S., Demétrio, C. G. B., & Hinde, J. P. (1998). Models for count data with many zeros. *The XIXth International Biometric Conference*, 179-192.

Rodriguez-Galiano, V.F., Chica-Olmo, M. & Chica-Rivas, M. (2014). Predictive modeling of gold potential with the integration of multisource information based on random forest: a case study on the Rodalquilar area, Southern Spain. *International Journal of Geographical Information Science*, Vol. 28, No. 7, 1336–1354.

Sadler, J. M., Goodall, J. L., Morsy, M. M., & Spencer, K. (2018). Modeling urban coastal flood severity from crowd-sourced flood reports using Poisson regression and Random Forest. *Journal of Hydrology*, 559, 43–55. doi:10.1016/j.jhydrol.2018.01.044

Sellers KF & Raim A. (2016) A flexible zero-inflated model to address data dispersion. *Compute Stat Data* Anal; 99:68–80.

Seyoum, A., Ndlovu, P., & Zewotir, T. (2016). Quasi-Poisson versus negative binomial regression models in identifying factors affecting initial CD4 cell count change due to antiretroviral therapy administered to HIV-positive adults in North–West Ethiopia (Amhara region). *AIDS Research and Therapy*, 13, 36.

Sim, S. Z., Gupta, R. C., & Ong, S. H. (2018). Zero-inflated Conway-Maxwell Poisson Distribution to Analyze Discrete Data. *The International Journal of Biostatistics*, 14(1).

Thakali, L. (2008). Development of accident prediction models for the highway of Thailand. Asian Institute of Technology.

Wedderburn, R.W.M. (1974). Quasi-likelihood functions, generalize linear model, and the Guss-Newton method. *Biometrika*, 61, 3, 439.

Xie, M., He, B., & Goh, T. (2001). Zero-inflated Poisson model in statistical process control. *Computational Statistics & Data Analysis*, 38(2), 191-201.

Yang, S., Harlow, L. L., Puggioni, G., & Redding, C. A. (2017). "A comparison of different methods of zero-inflated data analysis and an application in health surveys. *Journal of Modern Applied Statistical Methods*, 16(1), 518-543.

Yang, Z., Hardin, J. W., & Addy C. L. (2010). "Score Tests for Zero-Inflation in Overdispersed Count Data. *Communications in Statistics - Theory and Methods*, 39 (11): 2008–30.

วิทยานิพนธ์

กษมะ นิจจันทร์พันศรี. (2554). การศึกษาเปรียบเทียบความเหมาะสมของตัวแบบเชิงเส้นวางนัยทั่วไป:การแจกแจงซีโร-อินเฟรตเต็ด. (วิทยานิพนธ์ปริญญามหาบัณฑิต). มหาวิทยาลัยธรรมศาสตร์.

จันทิรา แย้มสรวล. (2559). การประมาณขนาดประชากรภายใต้การแจกแจงคอนเวย์แม็กซ์เวลล์ปัวซง. ภาควิชาคณิตศาสตร์และสถิติ สาขาวิชาสถิติประยุกต์ มหาวิทยาลัยธรรมศาสตร์.

นวพรรณ เชื้ออ่ำ, บุญอ้อม โฉมที และ อภิญญา หิรัญวงษ์. (2561). การเปรียบเทียบตัวแบบการถดถอยควอไซปัวซงและตัวแบบการถดถอยทวินามเชิงลบที่มีศูนย์มากสำหรับข้อมูลนับที่มีค่าความแปรปรวนมากกว่าค่าเฉลี่ย. ภาควิชาสถิติ คณะวิทยาศาสตร มหาวิทยาลัยเกษตรศาสตร์.

รุ่งนภา ศรีประโค. (2557). การลดปริมาณการขาดแคลนสินค้าโดยใช้เทคนิคพยากรณ์ กรณีศึกษาบริษัท ไอเซโล (ประเทศไทย) จำกัด. วิศวกรรมศาสตรมหาบัณฑิต สาขาวิชาการจัดการทางวิศวกรรม คณะวิศวกรรมศาสตร์ มหาวิทยาลัยธุรกิจบัณฑิตย์.

ธนวรรณ พรายดสถ์. (2559). การประเมินระดับความรุนแรงของผู้ป่วยที่มีภาวะหยุดหายใจขณะหลับด้วยวิธีการวิเคราะห์จากสัญญาณเสียงกรน. วิศวกรรมศาสตรมหาบัณฑิต สาขาวิศวกรรมไฟฟ้า มหาวิทยาลัยสงขลานครินทร์.

เมษา ทิพเวท. (2555). แบบจำลองคาดการณ์อุบัติเหตุสำหรับทางหลวงในเขตภูเขา. วิศวกรรมศาสตรมหาบัณฑิต สาขาวิศวกรรมขนส่ง มหาวิทยาลัยเทคโนโลยีสุรนารี. 20-21.

สุนิสา จันทร์น้ำท่วม. (2561). ผลกระทบของฟังก์ชั่นเชื่อมโยงและตัวแบบที่มีผลต่อช่วงความเชื่อมั่นสำหรับพารามิเตอร์ส่วนประกอบแบร์นูลลี. วิทยาศาสตรมหาบัณฑิต (สถิติประยุกต์) สาขาวิชาคณิตศาสตร์และสถิติ คณะวิทยาศาสตร์และเทคโนโลยี มหาวิทยาลัยธรรมศาสตร์.

สุทธิชัย งามจันทร์. (2553). แบบจำลองทำนายอุบัติเหตุบนทางด่วน กรณีศึกษาทางพิเศษเฉลิมมหานคร (ระบบทางด่วนขั้นที่ 1) และทางพิเศษศรีรัช (ระบบทางด่วนขั้นที่ 2). วิศวกรรมศาสตรมหาบัณฑิต สาขาวิชาวิศวกรรมขนส่ง มหาวิทยาลัยเทคโนโลยีสุรนารี.

ศิรินทิพย์ เสริมสุข และ กัลยา วานิชย์บัญชา. (2553). การเปรียบเทียบตัวแบบ Generalized Linear Model และตัวแบบ Generalized Estimating Equations ด้วยวิธีการประมาณพารามิเตอร์แบบ Quasi-Likelihood สำหรับข้อมูลระยะยาว. วิทยาศาสตร์กายภาพและคณิตศาสตรมหาบัณฑิต สาขาวิทยาศาสตร์กายภาพและคณิตศาสตร์ จุฬาลงกรณ์มหาวิทยาลัยกรุงเทพ.

Choo-Wosoba, H. (2016). Inference for a zero-inflated Conway-Maxwell. University of Louisville สืบค้นจาก https://ir.library.louisville.edu/cgi/viewcontent.cgi?article=3392&context=etd

Miller, J.M. (2007). Comparing Poisson, Hurdle and ZIP model fit under varying degrees of skew and Zero-inflation [Ph.D. dissertation]. University of Florida. สืบค้นจาก https://ufdcimages.uflib.ufl.edu/UF/E0/01/85/00/00001/miller_j.pdf

ภาคผนวก

ภาคผนวก

โปรแกรมสำหรับจำลองข้อมูลในงานวิจัย

กำหนดตัวแปรในโปรแกรมดังนี้

n คือ จำนวนค่าสังเกตในชุดข้อมูล
lambda คือ ค่าเฉลี่ยของการแจกแจง Poisson (λ)
nu คือ ค่าการกระจาย (ν)
p0 คือ ความน่าจะเป็นที่จะเกิดศูนย์ (π)
M คือ จำนวนรอบของการทำซ้ำ
trainData คือ ข้อมูลชุดฝึกสอน
testData คือ ข้อมูลชุดทดสอบ

กรณีข้อมูลมีปัญหาการกระจาย Under-dispersion

โปรแกรม

```
######################### Call Package ########################
library(MASS)
library(pscl)        #package for Zero-inflated Poisson / Negative Binomial
library(munsell)   #package for Cross Validation
library(fansi)       #package for Cross Validation
library(utf8)        #package for Cross Validation
library(car)         #package for Cross Validation
library(AER)         #package Dispersion test
library(ggplot2)
library(tidyverse)   #data manipulation and visualization
library(lava)        #package for Cross Validation
library(caret)       #package for Cross Validation
library(e1071)       #package for Cross Validation
library(lsr)         #package for Cross Validation
library(randomForest)  #package for Randomforest
```

```
library(mpcmp)        #package for Mean Parametrized CMP
library(COMPoissonReg) #package for CMP/ Zero-inflated CMP
library(rmutil)              #package for burr distribution
library(extraDistr)          # package for Fatigue life distribution

RMSE.test.cmp <- vector()
sd.test.cmp <- vector()
RMSE.test.RF <- vector()
sd.test.RF <- vector()
```

#simmulation fucntion

n is sample size #b is Coefficient of formula Lambda #gamma is Coefficient of formula nu #phi is Coefficient of formula pi form ZICMP real data #M is count of simulation

#STEP 1 : Generate y by Coefficient form ZICMP Real data

```
sim.data <-
function(n,b0,b1,b2,b3,b4,b5,b6,b7,b8,b9,gamma0,phi0,phi1,phi2,phi3,phi4,phi5,phi6,phi7)
{
  #Category Variable
  gen.X0 <- matrix(1, n, 1)
  gen.X1 <- t(rmultinom(n = n, size = 1, prob =
c(0.3995,0.1149,0.1232,0.3624)))
  gen.X2 <- t(rmultinom(n = n, size = 1, prob = c(0.8341,0.1659)))
  gen.X3 <- t(rmultinom(n = n, size = 1, prob =
c(0.6089,0.1789,0.1522,0.0600)))
  gen.X5 <- t(rmultinom(n = n, size = 1, prob =
c(0.5473,0.2261,0.1820,0.0446)))
```

```
#gen y form x used coefficient form real data
S = matrix(1, n, 1)
beta.true = b0
beta.true.X1 = c(b1, b2, b3)
beta.true.X3 = c(b4, b5, b6)
beta.true.X5 = c(b7, b8, b9)
gamma.true = gamma0
phi.true = phi0
phi.true.X2 = c(1,phi1)
phi.true.X3 = c(phi2, phi3, phi4)
phi.true.X5 = c(phi5, phi6, phi7)

lambda.true = exp(gen.X0 %*% beta.true + gen.X1[,-1] %*%
beta.true.X1 + gen.X3[,-1] %*% beta.true.X3
                  + gen.X5[,-1] %*% beta.true.X5) #formula.lambda (X) in
ZICMP

nu.true = exp(S %*% gamma.true) #formula.nu (S) in ZICMP

p.true = plogis(gen.X0 %*% phi.true + gen.X2[,-1] * phi.true.X2 +
gen.X3[,-1] %*% phi.true.X3
                  + gen.X5[,-1] %*% phi.true.X5 ) #fomula.p (W) in ZICMP

y = rzicmp(n, lambda = lambda.true, nu = nu.true, p = p.true)
#Combination y and X

dummy.X1 <- c(1,2,3,4)
dummy.X2 <- c(1,2)
dummy.X3 <- c(1,2,3,4)
dummy.X5 <- c(1,2,3,4)
X1 <- gen.X1 %*% dummy.X1
```

```
  X2 <- gen.X2 %*% dummy.X2
  X3 <- gen.X3 %*% dummy.X3
  X5 <- gen.X5 %*% dummy.X5
  sim <- data.frame(y,X1,X3,X5,lambda.true,nu.true,p.true)
}

# Step 2 : Model Efficiency Method by Cross Validation K=5
gen.data <- function(sim.data,M)
{
  for (i in 1:M) {
    # Create Training and Test data
    folds <- createFolds(sim.data$y, k=5, list=T)
    ## look at data in each fold
    dataFolds <- lapply(folds, FUN = function(x) sim.data[x, ])
    for(fold in folds) {
      trainData <- sim.data[-fold, ]
      testData <- sim.data[fold, ]
      trainData <- data.frame(trainData)
      testData <- data.frame(testData)

      #Check  Over-dispersion
      rd <- glm(y ~ ., data =  trainData , family = poisson)
      ## Quadratic specification (in terms of alpha:NB2)
      dis.test <- dispersiontest(rd, trafo = 2, alternative = c( "two.sided"))

      if( dis.test$p.value < 0.10) {

#STEP 3 :Fit Model

        #CMP regression
```

```
                cmp <- glm.cmp(y ~ X1 + X3  + X5  , data = trainData)
#accuracy for test
                p.pred <- predict(cmp,newdata = testData,type = "response")
                pre <- round(p.pred)
                #RMSE
                RMSE.cmp.t<-RMSE(p.pred,testData$y)
                RMSE.test.cmp <- append(RMSE.test.cmp,RMSE.cmp.t)

                #Random forest
                RF_model <- randomForest(y ~ X1 +  X3 + X5  , mtry = 3, data =
trainData,importance = TRUE)
                #accuracy for test
                p.pred.RF <- predict( RF_model,newdata = testData,type =
"response")
                pre.RF <- round(p.pred.RF)
                #RMSE
                RMSE.RF.t<-RMSE(p.pred.RF,testData$y)
                RMSE.test.RF <- append(RMSE.test.RF,RMSE.RF.t)
              }
            }

              sd.test.cmp[i] <- sqrt(var(RMSE.test.cmp))

              sd.test.RF[i] <- sqrt(var(RMSE.test.RF))

            }
            cat("\nCMP Model:")
            cat("\n Average RMSE of test:", round(mean(RMSE.test.cmp),5) )
            cat("\n SD RMSE of test:", round(mean(sd.test.cmp),5) )
            cat("\nRF Model:")
            cat("\n Average RMSE of test:", round(mean(RMSE.test.RF),5) )
```

```
  cat("\n SD RMSE of test:", round(mean(sd.test.RF),5) )
 }
```

<u>ตัวอย่างการใช้ฟังก์ชัน</u>

#1. Coefficient Original from ZICMP real data

```
library(COMPoissonReg)
b0= 0.9345
b1=-0.6681
b2= -2.0443
b3=-2.0889
b4=0.9000
b5=-0.1771
b6=-1.3811
b7=0.1440
b8=-0.3740
b9=1.4457
gamma0 = 1.6183
phi0 =0.7221
phi1 = 0.4091
phi2 =0.5717
phi3 =-0.3107
phi4 = -7.1086
phi5 =0.0051
phi6 =-1.0289
phi7 = -0.0535

#simulate data
set.seed(111)

#n = 250
```

```
sim.data1 <-
sim.data(n=250,b0,b1,b2,b3,b4,b5,b6,b7,b8,b9,gamma0,phi0,phi1,phi2,phi3,phi4,phi5,p
hi6,phi7)
mean(sim.data1$lambda.true)
mean(sim.data1$nu.true)
mean(sim.data1$p.true)

#n = 500
sim.data2 <-
sim.data(n=500,b0,b1,b2,b3,b4,b5,b6,b7,b8,b9,gamma0,phi0,phi1,phi2,phi3,phi4,phi5,p
hi6,phi7)
mean(sim.data2$lambda.true)
mean(sim.data2$nu.true)
mean(sim.data2$p.true)

#n = 1,000
sim.data3 <-
sim.data(n=1000,b0,b1,b2,b3,b4,b5,b6,b7,b8,b9,gamma0,phi0,phi1,phi2,phi3,phi4,phi5,
phi6,phi7)
mean(sim.data3$lambda.true)
mean(sim.data3$nu.true)
mean(sim.data3$p.true)

#n = 3,000
sim.data4 <-
sim.data(n=3000,b0,b1,b2,b3,b4,b5,b6,b7,b8,b9,gamma0,phi0,phi1,phi2,phi3,phi4,phi5,
phi6,phi7)
mean(sim.data4$lambda.true)
mean(sim.data4$nu.true)
mean(sim.data4$p.true)
```

```
#n = 5,000
sim.data5 <-
sim.data(n=5000,b0,b1,b2,b3,b4,b5,b6,b7,b8,b9,gamma0,phi0,phi1,phi2,phi3,phi4,phi5,
phi6,phi7)
mean(sim.data5$lambda.true)
mean(sim.data5$nu.true)
mean(sim.data5$p.true)
detach("package:COMPoissonReg", unload = TRUE)
detach("package:mpcmp", unload = TRUE)
# Model Efficiency Function
library(mpcmp)
gen.data(sim.data1,M=500)
gen.data(sim.data2,M=500)
gen.data(sim.data3,M=500)
gen.data(sim.data4,M=250)
gen.data(sim.data5,M=250)

#2. Coefficient Original Increase 50%
library(COMPoissonReg)
#coefficient form real data (Under-dispersion)
b0= 0.9345*1.5
b1=-0.6681*1.5
b2= -2.0443*1.5
b3=-2.0889*1.5
b4=0.9000*1.5
b5=-0.1771*1.5
b6=-1.3811*1.5
b7=0.1440*1.5
b8=-0.3740*1.5
b9=1.4457*1.5
gamma0 = 1.6183*1.5
```

```
phi0 =0.7221*1.5
phi1 = 0.4091*1.5
phi2 =0.5717*1.5
phi3 =-0.3107*1.5
phi4 = -7.1086*1.5
phi5 =0.0051*1.5
phi6 =-1.0289*1.5
phi7 = -0.0535*1.5

#simulate data
set.seed(100)
sim.data6 <-
sim.data(n=250,b0,b1,b2,b3,b4,b5,b6,b7,b8,b9,gamma0,phi0,phi1,phi2,phi3,phi4,phi5,p
hi6,phi7)
mean(sim.data6$lambda.true)
mean(sim.data6$nu.true)
mean(sim.data6$p.true)

sim.data7 <-
sim.data(n=500,b0,b1,b2,b3,b4,b5,b6,b7,b8,b9,gamma0,phi0,phi1,phi2,phi3,phi4,phi5,p
hi6,phi7)
mean(sim.data7$lambda.true)
mean(sim.data7$nu.true)
mean(sim.data7$p.true)

sim.data8 <-
sim.data(n=1000,b0,b1,b2,b3,b4,b5,b6,b7,b8,b9,gamma0,phi0,phi1,phi2,phi3,phi4,phi5,
phi6,phi7)
mean(sim.data8$lambda.true)
mean(sim.data8$nu.true)
mean(sim.data8$p.true)
```

```
sim.data9 <-
sim.data(n=3000,b0,b1,b2,b3,b4,b5,b6,b7,b8,b9,gamma0,phi0,phi1,phi2,phi3,phi4,phi5,
phi6,phi7)
mean(sim.data9$lambda.true)
mean(sim.data9$nu.true)
mean(sim.data9$p.true)

sim.data10 <-
sim.data(n=5000,b0,b1,b2,b3,b4,b5,b6,b7,b8,b9,gamma0,phi0,phi1,phi2,phi3,phi4,phi5,
phi6,phi7)
mean(sim.data10$lambda.true)
mean(sim.data10$nu.true)
mean(sim.data10$p.true)
detach("package:COMPoissonReg", unload = TRUE)
detach("package:mpcmp", unload = TRUE)

# Model Efficiency Function
library(mpcmp)
gen.data(sim.data6,M=500)
gen.data(sim.data7,M=500)
gen.data(sim.data8,M=500)
gen.data(sim.data9,M=500)
gen.data(sim.data10,M=500)
```

กรณีข้อมูลมีปัญหาการกระจาย Over-dispersion

โปรแกรม

```
RMSE.test.qp <- vector()
sd.test.qp <- vector()
RMSE.test.cmp <- vector()
sd.test.cmp <- vector()
RMSE.test.zip <- vector()
sd.test.zip <- vector()
RMSE.test.zinb <- vector()
sd.test.zinb <- vector()
RMSE.test.RF <- vector()
sd.test.RF <- vector()
```

```
#simmulation fucntion
# n is sample size #b is Coefficient of formula Lambda #gamma is Coefficient of formula nu #phi is Coefficient of formula pi form ZICMP real data #M is count of simulation

#STEP 1 : Generate y by Coefficient form ZICMP Real data
sim.data <- function(n,b0, b1, b2, gamma0, phi0,phi1)
{
  #gen X from real X n= 2,991
  X1 <- rburr(n, m=105.08, s=6.1361, f=0.63709 ) #Burr distribution m (beta) = location s, (alpha) = dispersion and f, (k) = family  parameters #beta, alpha, k is parameter from Easy Fit version 5.5

  X4 <- rfatigue(n, alpha=0.4852, beta = 72.352, mu = 0)  #Fatigue distribution  alpha=shape, beta=scale and mu=location parameters #beta, alpha, mu is parameter from Easy Fit version 5.5

  #gen y form x used coefficient form real data
```

```
    X = model.matrix(~ X1 + X4) #formula.lambda (beta)
    S = matrix(1, n, 1) # formula.nu intercept only (gamma)
    W = model.matrix(~ X1) #formula.p (phi)
    beta.true = c(b0, b1, b2)
    gamma.true = gamma0
    phi.true = c(phi0,phi1)

    lambda.true = exp(X %*% beta.true)
    nu.true = exp(S %*% gamma.true)
    p.true = plogis(W %*% phi.true)
    y = rzicmp(n, lambda = lambda.true, nu = nu.true, p = p.true)
    sim <- data.frame(y,X1,X4,lambda.true,nu.true,p.true) #Data
    }

# Step 2 : Model Efficiency Method by Cross Validation K=5
    gen.data <- function(data,M)
    {
    for (i in 1:M) {
    # Create Training and Test data
    folds <- createFolds(data$y, k=5, list=T)
    ## look at data in each fold
    dataFolds <- lapply(folds, FUN = function(x) data[x, ])
    for(fold in folds) {
      trainData <- data[-fold, ]
      testData <- data[fold, ]
      trainData <- data.frame(trainData)
      testData <- data.frame(testData)

      #Check dispersion
      rd <- glm(y ~ ., data =  trainData , family = poisson)
      ## Quadratic specification (in terms of alpha: NB2)
```

```
                dis.test <- dispersiontest(rd, trafo = 2, alternative = c( "two.sided"))

                if( dis.test$p.value < 0.10) {
#STEP 3 : Fit Model
                #Quasi poisson regression Model
                qp <- glm(y ~ X1 + X4, data = trainData, family = quasipoisson)
                #accuracy for test
                testData <- data.frame(testData)
                p.pred.qp <- predict( qp,newdata = testData,type = "response")
                #RMSE
                RMSE.qp.t<-RMSE(p.pred.qp,testData$y)
                RMSE.test.qp <- append(RMSE.test.qp,RMSE.qp.t)

                #CMP regression Model
                cmp <- glm.cmp(y ~ X1 + X4, data = trainData)
                #accuracy for test
                testData <- data.frame(testData)
                p.pred <- predict( cmp,newdata = testData,type = "response")

                #RMSE
                RMSE.cmp.t<-RMSE(p.pred,testData$y)
                RMSE.test.cmp <- append(RMSE.test.cmp,RMSE.cmp.t)

                #Zero-inflated poisson regression Model
                zip <- zeroinfl(y ~ X1 + X4 , data = trainData)
                #accuracy for test
                p.pred.zip <- predict( zip,newdata = testData,type = "response")
                #RMSE
                RMSE.zip.t<-RMSE(p.pred.zip,testData$y)
                RMSE.test.zip <- append(RMSE.test.zip,RMSE.zip.t)
```

```
                #Zero-inflated negative binomial regression Model
                zinb <- zeroinfl(y ~ X1 + X4, data = trainData,dist = "negbin")
                #accuracy for test
                p.pred.zinb <- predict( zinb,newdata = testData,type = "response")
                #RMSE
                RMSE.zinb.t<-RMSE(p.pred.zinb,testData$y)
                RMSE.test.zinb <- append(RMSE.test.zinb,RMSE.zinb.t)

                #Random forest
                RF_model <- randomForest(y ~ X1 + X4 ,mtry = 2, data =
trainData,importance = TRUE)
                #accuracy for test
                p.pred.RF <- predict( RF_model,newdata = testData,type =
"response")
                #RMSE each K
                RMSE.RF.t<-RMSE(p.pred.RF,testData$y)
                RMSE.test.RF <- append(RMSE.test.RF,RMSE.RF.t)
                }
              }
             #S.D. each i
              sd.test.qp[i] <- sqrt(var(RMSE.test.qp))

              sd.test.cmp[i] <- sqrt(var(RMSE.test.cmp))

              sd.test.zip[i] <- sqrt(var(RMSE.test.zip))

              sd.test.zinb[i] <- sqrt(var(RMSE.test.zinb))

              sd.test.RF[i] <- sqrt(var(RMSE.test.RF))

              }
```

```
  cat("QP Model:")
  cat("\n Average RMSE of test:", round(mean(RMSE.test.qp),5) )
  cat("\n SD RMSE of test:", round(mean(sd.test.qp),5) )
  cat("\nCMP Model:")
  cat("\n Average RMSE of test:", round(mean(RMSE.test.cmp),5) )
  cat("\n SD RMSE of test:", round(mean(sd.test.cmp),5) )
  cat("\nZIP Model:")
  cat("\n Average RMSE of test:", round(mean(RMSE.test.zip),5) )
  cat("\n SD RMSE of test:", round(mean(sd.test.zip),5) )
  cat("\nZINB Model:")
  cat("\n Average RMSE of test:", round(mean(RMSE.test.zinb),5) )
  cat("\n SD RMSE of test:", round(mean(sd.test.zinb),5) )
  cat("\nRF Model:")
  cat("\n Average RMSE of test:", round(mean(RMSE.test.RF),5) )
  cat("\n SD RMSE of test:", round(mean(sd.test.RF),5) )
}
```

<u>ตัวอย่างการใช้ฟังก์ชัน</u>

```
#1. Coefficient Original from ZICMP real data
        library(COMPoissonReg)
        b0= -0.3997
        b1=0.00009
        b2= -0.0032
        gamma0 = -0.4013
        phi0 =0.6330
        phi1 = -0.0123
        #simulate data
        set.seed(100)
        #n = 250
        sim.data1 <-sim.data(n=250,b0,b1,b2,gamma0,phi0,phi1)
```

```
mean(sim.data1$lambda.true) #Mean Lambda parameter
mean(sim.data1$nu.true) #Mean nu parameter
mean(sim.data1$p.true) #Mean pi parameter

#n = 500
sim.data2 <-sim.data(n=500,b0,b1,b2,gamma0,phi0,phi1)
mean(sim.data2$lambda.true)
mean(sim.data2$nu.true)
mean(sim.data2$p.true)

#n = 1,000
sim.data3 <-sim.data(n=1000,b0,b1,b2,gamma0,phi0,phi1)
mean(sim.data3$lambda.true)
mean(sim.data3$nu.true)
mean(sim.data3$p.true)

#n = 3,000
sim.data4 <-sim.data(n=3000,b0,b1,b2,gamma0,phi0,phi1)
mean(sim.data4$lambda.true)
mean(sim.data4$nu.true)
mean(sim.data4$p.true)

#n = 5,000
sim.data5 <-sim.data(n=5000,b0,b1,b2,gamma0,phi0,phi1)
mean(sim.data5$lambda.true)
mean(sim.data5$nu.true)
mean(sim.data5$p.true)
detach("package:COMPoissonReg", unload = TRUE)
detach("package:mpcmp", unload = TRUE)

# Model Efficiency Function
```

```
library(mpcmp)
gen.data(sim.data1,M=500)
gen.data(sim.data2,M=500)
gen.data(sim.data3,M=500)
gen.data(sim.data4,M=500)
gen.data(sim.data5,M=500)
```

#2. Coefficient Increase 50%

```
library(COMPoissonReg)
b0= -0.3997*1.5
b1=0.00009*1.5
b2= -0.0032*1.5
gamma0 = -0.4013*1.5
phi0 =0.6330*1.5
phi1 = -0.0123*1.5

#simulate data
set.seed(102)

#n = 250
sim.data6 <-sim.data(n=250,b0,b1,b2,gamma0,phi0,phi1)
mean(sim.data6$lambda.true)
mean(sim.data6$nu.true)
mean(sim.data6$p.true)

#n = 500
sim.data7 <-sim.data(n=500,b0,b1,b2,gamma0,phi0,phi1)
mean(sim.data7$lambda.true)
mean(sim.data7$nu.true)
mean(sim.data7$p.true)
```

```
#n = 1,000
sim.data8 <-sim.data(n=1000,b0,b1,b2,gamma0,phi0,phi1)
mean(sim.data8$lambda.true)
mean(sim.data8$nu.true)
mean(sim.data8$p.true)

#n = 3,000
sim.data9 <-sim.data(n=3000,b0,b1,b2,gamma0,phi0,phi1)
mean(sim.data9$lambda.true)
mean(sim.data9$nu.true)
mean(sim.data9$p.true)

#n = 5,000
sim.data10 <-sim.data(n=5000,b0,b1,b2,gamma0,phi0,phi1)
mean(sim.data10$lambda.true)
mean(sim.data10$nu.true)
mean(sim.data10$p.true)
detach("package:COMPoissonReg", unload = TRUE)
detach("package:mpcmp", unload = TRUE)

# Model Efficiency Function
library(mpcmp)
gen.data(sim.data6,M=500)
gen.data(sim.data7,M=500)
gen.data(sim.data8,M=500)
gen.data(sim.data9,M=500)
gen.data(sim.data10,M=500)
```

ตารางที่ 1ก ค่าพยากรณ์จำนวนผู้เสียชีวิตจากอุบัติเหตุทั่วประเทศไทยจากข้อมูลชุดทดสอบ (n=933) ของตัวแบบ CMP และ RF กรณีข้อมูลจริงที่มีการกระจายตัวต่ำกว่าเกณฑ์และค่าศูนย์เฟ้อ

อุบัติเหตุ (ครั้งที่)	ค่าพยากรณ์จำนวนผู้เสียชีวิต	
	CMP	RF
14	0.0939	0.0434
23	0.2422	0.2322
30	0.0743	0.1230
32	0.3012	0.3028
34	0.2726	0.2925
35	0.3656	0.4656
48	0.0794	0.0719
50	0.2676	0.2690
51	0.2676	0.2690
52	0.0743	0.1230
53	0.0893	0.0960
60	0.1786	0.1912
65	0.2268	0.1725
67	0.2415	0.3357
81	0.3656	0.4656
82	0.2415	0.3357
85	0.2726	0.2925
88	0.1198	0.1084
94	0.3422	0.2726
98	0.1617	0.2725
100	0.0794	0.0719
107	0.1617	0.2725
110	0.1121	0.1595
111	0.0821	0.1045
113	0.2726	0.2925
117	0.2676	0.2690
119	0.0743	0.1230

อุบัติเหตุ (ครั้งที่)	ค่าพยากรณ์จำนวนผู้เสียชีวิต	
	CMP	RF
137	0.2506	0.2649
149	0.2676	0.2690
157	0.0922	0.1006
166	0.0743	0.1230
171	0.1198	0.1084
174	0.2676	0.2690
179	0.2422	0.2322
192	0.1033	0.0469
200	0.2268	0.1725
202	0.3422	0.2726
210	0.1198	0.1084
211	0.0834	0.1263
216	0.3656	0.4656
229	0.3656	0.4656
231	0.0794	0.0719
237	0.2268	0.1725
240	0.2676	0.2690
242	0.2422	0.2322
251	0.3656	0.4656
254	0.3656	0.4656
255	0.0834	0.1263
256	0.0781	0.0727
260	0.0877	0.0672
266	0.0877	0.0672
268	0.0987	0.0503
270	0.0834	0.1263
284	0.0794	0.0719
288	0.3656	0.4656
293	0.0922	0.1006

อุบัติเหตุ (ครั้งที่)	ค่าพยากรณ์จำนวนผู้เสียชีวิต	
	CMP	RF
298	0.3012	0.3028
305	0.0877	0.0672
307	0.2248	0.4993
308	0.3656	0.4656
309	0.2422	0.2322
310	0.0743	0.1230
314	0.3656	0.4656
318	0.0794	0.0719
320	0.0794	0.0719
324	0.3656	0.4656
340	0.2726	0.2925
344	0.2268	0.1725
345	0.1727	0.1294
358	0.0794	0.0719
370	0.2726	0.2925
379	0.2676	0.2690
382	0.2726	0.2925
387	0.0939	0.0434
392	0.0794	0.0719
395	0.2676	0.2690
400	0.1037	0.1197
401	0.1033	0.0469
406	0.2726	0.2925
421	0.2268	0.1725
422	0.0794	0.0719
429	0.1259	0.0932
435	0.2676	0.2690
436	0.1142	0.0000
443	0.2726	0.2925

อุบัติเหตุ (ครั้งที่)	ค่าพยากรณ์จำนวนผู้เสียชีวิต	
	CMP	RF
447	0.0794	0.0719
448	0.0794	0.0719
450	0.1033	0.0469
457	0.1617	0.2725
461	0.0794	0.0719
463	0.2676	0.2690
469	0.2268	0.1725
477	0.1617	0.2725
478	0.2676	0.2690
479	0.3422	0.2726
487	0.5662	0.3934
488	0.2422	0.2322
489	0.3012	0.3028
492	0.0893	0.0960
495	0.0834	0.1263
501	0.2726	0.2925
517	0.3656	0.4656
519	0.0794	0.0719
525	0.1121	0.1595
526	0.2676	0.2690
528	0.0834	0.1263
530	0.2422	0.2322
540	0.2726	0.2925
545	0.2676	0.2690
547	0.1033	0.0469
558	0.0743	0.1230
581	0.0794	0.0719
582	0.2676	0.2690
588	0.2147	0.2258

อุบัติเหตุ (ครั้งที่)	ค่าพยากรณ์จำนวนผู้เสียชีวิต	
	CMP	RF
591	0.1727	0.1294
592	0.2726	0.2925
594	0.0893	0.0960
596	0.1727	0.1294
613	0.2484	0.0504
618	0.1617	0.2725
620	0.1617	0.2725
622	0.3012	0.3028
631	0.0821	0.1045
660	0.1727	0.1294
666	0.2676	0.2690
669	0.0877	0.0672
673	0.1259	0.0932
674	0.3012	0.3028
681	0.5662	0.3934
699	0.0794	0.0719
704	0.2268	0.1725
717	0.0821	0.1045
730	0.0877	0.0672
737	0.2726	0.2925
746	0.0877	0.0672
748	0.0877	0.0672
750	0.0893	0.0960
751	0.1198	0.1084
752	0.0794	0.0719
757	0.2422	0.2322
768	0.3656	0.4656
769	0.0794	0.0719
771	0.0893	0.0960

อุบัติเหตุ (ครั้งที่)	ค่าพยากรณ์จำนวนผู้เสียชีวิต	
	CMP	RF
775	0.1727	0.1294
788	0.0834	0.1263
789	0.3012	0.3028
790	0.0922	0.1006
793	0.2726	0.2925
804	0.2933	0.3223
805	0.0834	0.1263
808	0.1037	0.1197
814	0.2676	0.2690
833	0.0987	0.0503
839	0.2268	0.1725
847	0.2676	0.2690
851	0.2422	0.2322
858	0.0834	0.1263
861	0.2676	0.2690
864	0.2676	0.2690
873	0.1198	0.1084
885	0.0743	0.1230
886	0.0794	0.0719
887	0.0939	0.0434
893	0.2676	0.2690
904	0.4759	0.0004
911	0.1348	0.0000
914	0.2422	0.2322
917	0.0794	0.0719
918	0.1121	0.1595
923	0.0863	0.0876
936	0.2676	0.2690
939	0.1198	0.1084

อุบัติเหตุ (ครั้งที่)	ค่าพยากรณ์จำนวนผู้เสียชีวิต	
	CMP	RF
940	0.0893	0.0960
952	0.1727	0.1294
956	0.1855	0.1539
958	0.2422	0.2322
967	0.7370	0.8937
971	0.2268	0.1725
972	0.0893	0.0960
976	0.2676	0.2690
979	0.2506	0.2649
987	0.1727	0.1294
993	0.0794	0.0719
1007	0.3778	0.0016
1009	0.0781	0.0727
1017	0.0743	0.1230
1018	0.2606	0.2608
1020	0.2422	0.2322
1025	0.0893	0.0960
1030	0.3656	0.4656
1038	0.2422	0.2322
1039	0.0877	0.0672
1046	0.2422	0.2322
1050	0.1033	0.0469
1058	0.2726	0.2925
1060	0.1121	0.1595
1061	0.0794	0.0719
1069	0.0794	0.0719
1075	0.0834	0.1263
1083	0.0794	0.0719
1088	0.0987	0.0503

อุบัติเหตุ (ครั้งที่)	ค่าพยากรณ์จำนวนผู้เสียชีวิต	
	CMP	RF
1094	0.1200	0.0006
1095	0.1198	0.1084
1097	0.1855	0.1539
1101	0.2422	0.2322
1102	0.1727	0.1294
1104	0.0939	0.0434
1112	0.3656	0.4656
1114	0.0939	0.0434
1116	0.1417	0.0000
1122	0.1198	0.1084
1124	0.1198	0.1084
1132	0.2726	0.2925
1133	0.2415	0.3357
1139	0.0794	0.0719
1142	0.0794	0.0719
1143	0.2726	0.2925
1149	0.0877	0.0672
1155	0.2726	0.2925
1157	0.1943	0.3119
1159	0.1943	0.3119
1160	0.1121	0.1595
1161	0.3656	0.4656
1162	0.0834	0.1263
1164	0.2422	0.2322
1169	0.2676	0.2690
1183	0.2268	0.1725
1186	0.7370	0.8937
1192	0.0794	0.0719
1193	0.0893	0.0960

อุบัติเหตุ (ครั้งที่)	ค่าพยากรณ์จำนวนผู้เสียชีวิต	
	CMP	RF
1200	0.1727	0.1294
1211	0.2422	0.2322
1213	0.0922	0.1006
1220	0.3656	0.4656
1223	0.2268	0.1725
1227	0.0893	0.0960
1238	0.2439	0.1471
1244	0.0794	0.0719
1247	0.3656	0.4656
1251	0.2415	0.3357
1261	0.2422	0.2322
1265	0.1727	0.1294
1266	0.0794	0.0719
1275	0.0743	0.1230
1283	0.2422	0.2322
1286	0.1727	0.1294
1292	0.2422	0.2322
1293	0.2422	0.2322
1298	0.1417	0.0000
1299	0.0781	0.0727
1301	0.2726	0.2925
1312	0.2439	0.1471
1313	0.2676	0.2690
1318	0.0987	0.0503
1330	0.1121	0.1595
1333	0.2439	0.1471
1336	0.2726	0.2925
1343	0.2422	0.2322
1344	0.2676	0.2690

อุบัติเหตุ (ครั้งที่)	ค่าพยากรณ์จำนวนผู้เสียชีวิต	
	CMP	RF
1352	0.2439	0.1471
1359	0.3656	0.4656
1360	0.0794	0.0719
1361	0.1259	0.0932
1365	0.0794	0.0719
1367	0.1121	0.1595
1370	0.0794	0.0719
1374	0.2676	0.2690
1382	0.0794	0.0719
1385	0.2676	0.2690
1387	0.1617	0.2725
1389	0.0794	0.0719
1390	0.0794	0.0719
1414	0.1121	0.1595
1422	0.2676	0.2690
1427	0.2676	0.2690
1434	0.3012	0.3028
1439	0.2676	0.2690
1440	0.2676	0.2690
1443	0.1198	0.1084
1447	0.1198	0.1084
1465	0.2422	0.2322
1472	0.2676	0.2690
1487	0.1198	0.1084
1490	0.0794	0.0719
1492	0.2676	0.2690
1493	0.3656	0.4656
1505	0.1348	0.0000
1511	0.0794	0.0719

อุบัติเหตุ (ครั้งที่)	ค่าพยากรณ์จำนวนผู้เสียชีวิต	
	CMP	RF
1515	0.1121	0.1595
1528	0.5662	0.3934
1539	0.2422	0.2322
1540	0.1198	0.1084
1552	0.0794	0.0719
1555	0.2268	0.1725
1558	0.2726	0.2925
1559	0.0794	0.0719
1569	0.0893	0.0960
1572	0.2506	0.2649
1573	0.3012	0.3028
1576	0.1348	0.0000
1578	0.1121	0.1595
1583	0.2422	0.2322
1598	0.0834	0.1263
1600	0.2606	0.2608
1601	0.2422	0.2322
1602	0.0834	0.1263
1606	0.2422	0.2322
1616	0.2422	0.2322
1619	0.2422	0.2322
1629	0.2422	0.2322
1630	0.3484	0.5263
1635	0.0922	0.1006
1637	0.0834	0.1263
1640	0.0834	0.1263
1643	0.2676	0.2690
1650	0.1617	0.2725
1652	0.5662	0.3934

อุบัติเหตุ (ครั้งที่)	ค่าพยากรณ์จำนวนผู้เสียชีวิต	
	CMP	RF
1664	0.2268	0.1725
1672	0.3484	0.5263
1677	0.2606	0.2608
1678	0.2268	0.1725
1683	0.0794	0.0719
1687	0.2268	0.1725
1693	0.0794	0.0719
1695	0.2415	0.3357
1707	0.0821	0.1045
1708	0.0794	0.0719
1710	0.1617	0.2725
1713	0.2415	0.3357
1716	0.2422	0.2322
1722	0.0922	0.1006
1728	0.2726	0.2925
1734	0.2415	0.3357
1736	0.6371	0.3994
1738	0.3656	0.4656
1748	0.2147	0.2258
1750	0.4759	0.0004
1754	0.0922	0.1006
1756	0.7370	0.8937
1764	0.1121	0.1595
1770	0.2268	0.1725
1776	0.2506	0.2649
1784	0.0893	0.0960
1788	0.2422	0.2322
1791	0.3656	0.4656
1792	0.3656	0.4656

อุบัติเหตุ (ครั้งที่)	ค่าพยากรณ์จำนวนผู้เสียชีวิต	
	CMP	RF
1793	0.0794	0.0719
1812	0.1198	0.1084
1815	0.0794	0.0719
1823	0.1037	0.1197
1842	0.0922	0.1006
1845	0.2268	0.1725
1846	0.1617	0.2725
1848	0.2422	0.2322
1859	0.1259	0.0932
1864	0.0939	0.0434
1868	0.2415	0.3357
1875	0.3656	0.4656
1877	0.3656	0.4656
1881	0.2676	0.2690
1882	0.0821	0.1045
1887	0.1198	0.1084
1888	0.0794	0.0719
1896	0.0794	0.0719
1905	0.2422	0.2322
1924	0.0743	0.1230
1925	0.2422	0.2322
1931	0.0794	0.0719
1936	0.3656	0.4656
1938	0.0743	0.1230
1940	0.0794	0.0719
1946	0.2422	0.2322
1947	0.3656	0.4656
1950	0.0877	0.0672
1951	0.0794	0.0719

อุบัติเหตุ (ครั้งที่)	ค่าพยากรณ์จำนวนผู้เสียชีวิต	
	CMP	RF
1952	0.1033	0.0469
1955	0.2606	0.2608
1957	0.0794	0.0719
1967	0.0863	0.0876
1980	0.2676	0.2690
1981	0.2676	0.2690
1984	0.4114	0.3354
1987	0.3656	0.4656
1989	0.1736	0.1651
1994	0.2268	0.1725
2000	0.0877	0.0672
2010	0.1559	0.0465
2012	0.0743	0.1230
2017	0.2268	0.1725
2028	0.0893	0.0960
2032	0.2268	0.1725
2033	0.2676	0.2690
2035	0.0834	0.1263
2044	0.2676	0.2690
2049	0.2268	0.1725
2054	0.3422	0.2726
2061	0.2676	0.2690
2063	0.2268	0.1725
2068	0.3422	0.2726
2078	0.0834	0.1263
2083	0.0743	0.1230
2095	0.2422	0.2322
2099	0.2422	0.2322
2107	0.2606	0.2608

อุบัติเหตุ (ครั้งที่)	ค่าพยากรณ์จำนวนผู้เสียชีวิต	
	CMP	RF
2108	0.2676	0.2690
2110	0.0834	0.1263
2115	0.0877	0.0672
2120	0.0794	0.0719
2123	0.2422	0.2322
2124	0.0834	0.1263
2132	0.3656	0.4656
2133	0.1198	0.1084
2138	0.1727	0.1294
2145	0.2422	0.2322
2146	0.0743	0.1230
2147	0.2506	0.2649
2148	0.0939	0.0434
2149	0.2676	0.2690
2150	0.0987	0.0503
2151	0.1198	0.1084
2155	0.1727	0.1294
2156	0.1259	0.0932
2163	0.1033	0.0469
2164	0.2268	0.1725
2168	0.2726	0.2925
2177	0.0794	0.0719
2186	0.1121	0.1595
2187	0.0987	0.0503
2193	0.2726	0.2925
2195	0.1121	0.1595
2196	0.1727	0.1294
2202	0.2422	0.2322
2216	0.0794	0.0719

อุบัติเหตุ (ครั้งที่)	ค่าพยากรณ์จำนวนผู้เสียชีวิต	
	CMP	RF
2226	0.2268	0.1725
2228	0.1855	0.1539
2231	0.7370	0.8937
2235	0.2422	0.2322
2239	0.2726	0.2925
2247	0.2422	0.2322
2250	0.2268	0.1725
2263	0.1855	0.1539
2269	0.2726	0.2925
2271	0.1727	0.1294
2276	0.0794	0.0719
2278	0.2422	0.2322
2279	0.0794	0.0719
2286	0.0939	0.0434
2290	0.1727	0.1294
2292	0.1855	0.1539
2294	0.2726	0.2925
2297	0.0794	0.0719
2302	0.1121	0.1595
2305	0.4114	0.3354
2314	0.2422	0.2322
2317	0.0922	0.1006
2321	0.2676	0.2690
2330	0.0794	0.0719
2333	0.2676	0.2690
2338	0.2422	0.2322
2340	0.1121	0.1595
2343	0.0743	0.1230
2344	0.2422	0.2322

อุบัติเหตุ (ครั้งที่)	ค่าพยากรณ์จำนวนผู้เสียชีวิต	
	CMP	RF
2345	0.2422	0.2322
2346	0.1348	0.0000
2359	0.2422	0.2322
2360	0.0794	0.0719
2365	0.2422	0.2322
2367	0.0794	0.0719
2369	0.0743	0.1230
2376	0.1727	0.1294
2380	0.0794	0.0719
2381	0.0834	0.1263
2388	0.0877	0.0672
2389	0.1142	0.0000
2399	0.1727	0.1294
2405	0.2676	0.2690
2407	0.2676	0.2690
2408	0.1037	0.1197
2410	0.1559	0.0465
2419	0.2422	0.2322
2421	0.1855	0.1539
2424	0.2676	0.2690
2425	0.2422	0.2322
2427	0.2415	0.3357
2441	0.1348	0.0000
2449	0.0743	0.1230
2454	0.2422	0.2322
2460	0.3656	0.4656
2461	0.1121	0.1595
2466	0.5300	0.1729
2467	0.2422	0.2322

อุบัติเหตุ (ครั้งที่)	ค่าพยากรณ์จำนวนผู้เสียชีวิต	
	CMP	RF
2473	0.2676	0.2690
2476	0.2422	0.2322
2485	0.0794	0.0719
2487	0.0794	0.0719
2496	0.2422	0.2322
2502	0.0794	0.0719
2513	0.2422	0.2322
2516	0.1943	0.3119
2521	0.2676	0.2690
2522	0.0743	0.1230
2524	0.2726	0.2925
2532	0.2606	0.2608
2540	0.0794	0.0719
2549	0.2676	0.2690
2557	0.2422	0.2322
2573	0.0794	0.0719
2575	0.2422	0.2322
2576	0.0743	0.1230
2578	0.2676	0.2690
2589	0.7370	0.8937
2591	0.0794	0.0719
2594	0.2422	0.2322
2597	0.2676	0.2690
2598	0.2422	0.2322
2599	0.0821	0.1045
2600	0.2726	0.2925
2606	0.1943	0.3119
2608	0.0939	0.0434
2609	0.1037	0.1197

อุบัติเหตุ (ครั้งที่)	ค่าพยากรณ์จำนวนผู้เสียชีวิต	
	CMP	RF
2611	0.1908	0.1766
2613	0.2676	0.2690
2614	0.3154	0.2005
2623	0.0877	0.0672
2625	0.0834	0.1263
2629	0.1855	0.1539
2630	0.2422	0.2322
2640	0.2933	0.3223
2641	0.0834	0.1263
2657	0.2726	0.2925
2659	0.4114	0.3354
2660	0.2676	0.2690
2662	0.2422	0.2322
2667	0.1142	0.0000
2669	0.1559	0.0465
2671	0.2422	0.2322
2673	0.0794	0.0719
2677	0.1033	0.0469
2679	0.1121	0.1595
2686	0.1033	0.0469
2687	0.0794	0.0719
2691	0.0834	0.1263
2695	0.0922	0.1006
2699	0.2422	0.2322
2705	0.7370	0.8937
2710	0.1348	0.0000
2712	0.2415	0.3357
2718	0.3012	0.3028
2734	0.2268	0.1725

อุบัติเหตุ (ครั้งที่)	ค่าพยากรณ์จำนวนผู้เสียชีวิต	
	CMP	RF
2737	0.2422	0.2322
2739	0.0877	0.0672
2745	0.0877	0.0672
2747	0.2422	0.2322
2750	0.2726	0.2925
2760	0.3656	0.4656
2767	0.0794	0.0719
2773	0.3012	0.3028
2778	0.2676	0.2690
2781	0.2726	0.2925
2782	0.2676	0.2690
2790	0.3656	0.4656
2803	0.2422	0.2322
2808	0.2676	0.2690
2810	0.1727	0.1294
2822	0.2676	0.2690
2823	0.0794	0.0719
2827	0.2676	0.2690
2829	0.0794	0.0719
2841	0.1198	0.1084
2849	0.2506	0.2649
2862	0.2676	0.2690
2864	0.2606	0.2608
2869	0.2422	0.2322
2875	0.0743	0.1230
2890	0.2726	0.2925
2902	0.3656	0.4656
2908	0.2676	0.2690
2910	0.1198	0.1084

อุบัติเหตุ (ครั้งที่)	ค่าพยากรณ์จำนวนผู้เสียชีวิต	
	CMP	RF
2913	0.0794	0.0719
2920	0.2268	0.1725
2928	0.3012	0.3028
2935	0.3656	0.4656
2941	0.2676	0.2690
2945	0.2676	0.2690
2949	0.7370	0.8937
2952	0.1727	0.1294
2961	0.2415	0.3357
2967	0.0939	0.0434
2968	0.3778	0.0016
2970	0.2676	0.2690
2976	0.3656	0.4656
2979	0.2422	0.2322
2981	0.0893	0.0960
2983	0.0743	0.1230
2987	0.1943	0.3119
2989	0.2268	0.1725
2991	0.2726	0.2925
2997	0.0794	0.0719
3003	0.2726	0.2925
3005	0.2422	0.2322
3010	0.6371	0.3994
3012	0.2676	0.2690
3017	0.7370	0.8937
3044	0.0922	0.1006
3048	0.2676	0.2690
3057	0.2422	0.2322
3061	0.0877	0.0672

อุบัติเหตุ (ครั้งที่)	ค่าพยากรณ์จำนวนผู้เสียชีวิต	
	CMP	RF
3062	0.2933	0.3223
3063	0.2676	0.2690
3064	0.2422	0.2322
3071	0.0922	0.1006
3072	0.0834	0.1263
3074	0.0794	0.0719
3082	0.2726	0.2925
3091	0.2676	0.2690
3093	0.0794	0.0719
3095	0.0922	0.1006
3097	0.0922	0.1006
3100	0.2268	0.1725
3113	0.2422	0.2322
3117	0.0794	0.0719
3122	0.2676	0.2690
3128	0.2268	0.1725
3143	0.1198	0.1084
3144	0.3656	0.4656
3147	0.1198	0.1084
3149	0.2422	0.2322
3152	0.1727	0.1294
3156	0.2422	0.2322
3159	0.1727	0.1294
3160	0.3656	0.4656
3175	0.0877	0.0672
3177	0.0794	0.0719
3178	0.2268	0.1725
3179	0.0821	0.1045
3186	0.2422	0.2322

อุบัติเหตุ (ครั้งที่)	ค่าพยากรณ์จำนวนผู้เสียชีวิต	
	CMP	RF
3190	0.3154	0.2005
3192	0.0834	0.1263
3194	0.2422	0.2322
3197	0.2676	0.2690
3198	0.2422	0.2322
3201	0.0821	0.1045
3206	0.0922	0.1006
3218	0.1121	0.1595
3220	0.0877	0.0672
3222	0.1033	0.0469
3229	0.1198	0.1084
3232	0.1198	0.1084
3233	0.5662	0.3934
3234	0.0834	0.1263
3240	0.2422	0.2322
3242	0.0877	0.0672
3244	0.0877	0.0672
3247	0.2676	0.2690
3250	0.2606	0.2608
3251	0.1198	0.1084
3257	0.0743	0.1230
3260	0.2422	0.2322
3261	0.2606	0.2608
3263	0.2422	0.2322
3267	0.0794	0.0719
3273	0.0877	0.0672
3275	0.0877	0.0672
3279	0.0821	0.1045
3289	0.2726	0.2925

อุบัติเหตุ (ครั้งที่)	ค่าพยากรณ์จำนวนผู้เสียชีวิต	
	CMP	RF
3290	0.1033	0.0469
3296	0.2422	0.2322
3299	0.0794	0.0719
3313	0.2506	0.2649
3316	0.0743	0.1230
3328	0.2484	0.0504
3333	0.0794	0.0719
3334	0.1037	0.1197
3356	0.1727	0.1294
3357	0.1727	0.1294
3360	0.0922	0.1006
3374	0.1198	0.1084
3381	0.2422	0.2322
3383	0.2268	0.1725
3385	0.1259	0.0932
3388	0.1121	0.1595
3389	0.1121	0.1595
3390	0.2268	0.1725
3398	0.0877	0.0672
3400	0.2676	0.2690
3401	0.2422	0.2322
3415	0.0794	0.0719
3419	0.3656	0.4656
3420	0.2676	0.2690
3424	0.0877	0.0672
3439	0.1617	0.2725
3443	0.1617	0.2725
3454	0.2422	0.2322
3456	0.2933	0.3223

อุบัติเหตุ (ครั้งที่)	ค่าพยากรณ์จำนวนผู้เสียชีวิต	
	CMP	RF
3461	0.0834	0.1263
3466	0.2422	0.2322
3480	0.2676	0.2690
3485	0.1198	0.1084
3492	0.0794	0.0719
3494	0.2506	0.2649
3495	0.0781	0.0727
3507	0.3012	0.3028
3509	0.3656	0.4656
3520	0.2506	0.2649
3527	0.0893	0.0960
3536	0.0781	0.0727
3539	0.0877	0.0672
3544	0.0922	0.1006
3548	0.0794	0.0719
3556	0.2422	0.2322
3565	0.2422	0.2322
3569	0.0939	0.0434
3570	0.0794	0.0719
3572	0.2422	0.2322
3573	0.1198	0.1084
3600	0.4114	0.3354
3604	0.2422	0.2322
3607	0.0834	0.1263
3611	0.2422	0.2322
3615	0.3012	0.3028
3616	0.1198	0.1084
3620	0.0939	0.0434
3621	0.3422	0.2726

อุบัติเหตุ (ครั้งที่)	ค่าพยากรณ์จำนวนผู้เสียชีวิต	
	CMP	RF
3636	0.1037	0.1197
3644	0.2676	0.2690
3645	0.2676	0.2690
3647	0.0794	0.0719
3649	0.2268	0.1725
3660	0.1259	0.0932
3662	0.2676	0.2690
3669	0.1198	0.1084
3670	0.1559	0.0465
3676	0.2422	0.2322
3681	0.0794	0.0719
3684	0.0877	0.0672
3685	0.0743	0.1230
3689	0.2268	0.1725
3700	0.0794	0.0719
3702	0.1033	0.0469
3703	0.2422	0.2322
3713	0.2726	0.2925
3719	0.2676	0.2690
3731	0.1786	0.1912
3732	0.2676	0.2690
3745	0.2422	0.2322
3758	0.0939	0.0434
3760	0.5662	0.3934
3762	0.2268	0.1725
3764	0.1033	0.0469
3777	0.1786	0.1912
3778	0.0743	0.1230
3780	0.2676	0.2690

อุบัติเหตุ (ครั้งที่)	ค่าพยากรณ์จำนวนผู้เสียชีวิต	
	CMP	RF
3785	0.2676	0.2690
3789	0.1855	0.1539
3798	0.1121	0.1595
3801	0.1033	0.0469
3809	0.5300	0.1729
3816	0.2726	0.2925
3817	0.2422	0.2322
3818	0.1727	0.1294
3821	0.2422	0.2322
3822	0.2422	0.2322
3830	0.0939	0.0434
3832	0.2676	0.2690
3838	0.3422	0.2726
3841	0.0794	0.0719
3842	0.1179	0.2030
3850	0.0794	0.0719
3851	0.1198	0.1084
3862	0.0794	0.0719
3867	0.0834	0.1263
3869	0.1198	0.1084
3870	0.0794	0.0719
3882	0.2506	0.2649
3883	0.2676	0.2690
3889	0.1037	0.1197
3894	0.2268	0.1725
3897	0.1727	0.1294
3899	0.3656	0.4656
3900	0.2606	0.2608
3901	0.0794	0.0719

อุบัติเหตุ (ครั้งที่)	ค่าพยากรณ์จำนวนผู้เสียชีวิต	
	CMP	RF
3922	0.2726	0.2925
3924	0.1037	0.1197
3927	0.0743	0.1230
3930	0.0794	0.0719
3934	0.1559	0.0465
3935	0.2676	0.2690
3937	0.2676	0.2690
3944	0.2506	0.2649
3947	0.1198	0.1084
3951	0.0893	0.0960
3953	0.0743	0.1230
3958	0.0743	0.1230
3962	0.3422	0.2726
3963	0.1908	0.1766
3964	0.0834	0.1263
3971	0.3656	0.4656
3977	0.1198	0.1084
3980	0.0922	0.1006
3986	0.2268	0.1725
3990	0.5662	0.3934
3991	0.0893	0.0960
3996	0.2422	0.2322
3998	0.1037	0.1197
4007	0.2422	0.2322
4013	0.0877	0.0672
4014	0.3656	0.4656
4029	0.2676	0.2690
4030	0.4036	0.2356
4036	0.2422	0.2322

อุบัติเหตุ (ครั้งที่)	ค่าพยากรณ์จำนวนผู้เสียชีวิต	
	CMP	RF
4046	0.0877	0.0672
4047	0.2676	0.2690
4048	0.2676	0.2690
4054	0.2439	0.1471
4056	0.0922	0.1006
4062	0.0893	0.0960
4068	0.0743	0.1230
4108	0.2676	0.2690
4112	0.3422	0.2726
4120	0.0922	0.1006
4123	0.1033	0.0469
4124	0.0922	0.1006
4126	0.1037	0.1197
4129	0.0922	0.1006
4133	0.0939	0.0434
4134	0.1179	0.2030
4139	0.2268	0.1725
4141	0.2422	0.2322
4148	0.0877	0.0672
4167	0.3656	0.4656
4168	0.3656	0.4656
4173	0.1617	0.2725
4175	0.1727	0.1294
4177	0.2147	0.2258
4189	0.5662	0.3934
4195	0.0877	0.0672
4204	0.0794	0.0719
4212	0.0794	0.0719
4214	0.2422	0.2322

อุบัติเหตุ (ครั้งที่)	ค่าพยากรณ์จำนวนผู้เสียชีวิต	
	CMP	RF
4216	0.0794	0.0719
4220	0.0893	0.0960
4222	0.3656	0.4656
4226	0.1198	0.1084
4236	0.0743	0.1230
4248	0.0781	0.0727
4253	0.2268	0.1725
4259	0.1908	0.1766
4263	0.1348	0.0000
4273	0.0877	0.0672
4281	0.2422	0.2322
4282	0.2676	0.2690
4287	0.2606	0.2608
4288	0.0922	0.1006
4291	0.0743	0.1230
4294	0.0877	0.0672
4296	0.3656	0.4656
4299	0.1559	0.0465
4310	0.0794	0.0719
4314	0.2726	0.2925
4319	0.3656	0.4656
4327	0.0743	0.1230
4335	0.2422	0.2322
4342	0.2422	0.2322
4343	0.2422	0.2322
4356	0.3656	0.4656
4358	0.0743	0.1230
4361	0.1727	0.1294
4367	0.0794	0.0719

อุบัติเหตุ (ครั้งที่)	ค่าพยากรณ์จำนวนผู้เสียชีวิต	
	CMP	RF
4377	0.0877	0.0672
4378	0.1617	0.2725
4386	0.2422	0.2322
4387	0.0794	0.0719
4396	0.1727	0.1294
4404	0.1736	0.1651
4408	0.0794	0.0719
4413	0.2676	0.2690
4414	0.2422	0.2322
4416	0.2676	0.2690
4418	0.3656	0.4656
4423	0.2248	0.4993
4424	0.2422	0.2322
4431	0.0794	0.0719
4435	0.2422	0.2322
4439	0.1198	0.1084
4443	0.0939	0.0434
4444	0.2268	0.1725
4446	0.0877	0.0672
4450	0.0794	0.0719
4453	0.0743	0.1230
4457	0.0743	0.1230
4459	0.0922	0.1006
4463	0.1727	0.1294
4466	0.0794	0.0719
4473	0.0743	0.1230
4474	0.1908	0.1766
4475	0.0794	0.0719
4478	0.0893	0.0960

อุบัติเหตุ (ครั้งที่)	ค่าพยากรณ์จำนวนผู้เสียชีวิต	
	CMP	RF
4480	0.1943	0.3119
4481	0.2422	0.2322
4484	0.3154	0.2005
4493	0.2676	0.2690
4496	0.2676	0.2690
4504	0.0987	0.0503
4510	0.2422	0.2322
4514	0.1121	0.1595
4516	0.2726	0.2925
4524	0.1617	0.2725
4527	0.2415	0.3357
4531	0.2422	0.2322
4553	0.1259	0.0932
4562	0.0893	0.0960
4582	0.0863	0.0876
4591	0.0834	0.1263
4592	0.0834	0.1263
4595	0.0834	0.1263
4602	0.3656	0.4656
4603	0.1727	0.1294
4606	0.2726	0.2925
4609	0.2268	0.1725
4616	0.2422	0.2322
4620	0.3422	0.2726
4630	0.1908	0.1766
4632	0.2422	0.2322
4636	0.0743	0.1230
4641	0.2422	0.2322
4648	0.1198	0.1084

อุบัติเหตุ (ครั้งที่)	ค่าพยากรณ์จำนวนผู้เสียชีวิต	
	CMP	RF
4651	0.2676	0.2690
4652	0.1727	0.1294
4654	0.0893	0.0960
4657	0.3656	0.4656
4659	0.0743	0.1230
4662	0.2726	0.2925
4666	0.2726	0.2925

ตารางที่ 2ก ค่าพยากรณ์จำนวนครั้งการเรียกร้องค่าสินไหมทดแทนของลูกค้าของบริษัทประกันภัยแห่งหนึ่งจากข้อมูลชุดทดสอบ (n=597) กรณีข้อมูลจริงที่มีการกระจายตัวเกินเกณฑ์และค่าศูนย์เฟ้อ

ลูกค้า (คนที่)	ค่าพยากรณ์จำนวนครั้งการเรียกร้องค่าสินไหมทดแทน				
	QP	CMP	ZIP	ZINB	RF
8	0.4364	0.4329	0.4329	0.4331	0.0327
11	0.4620	0.4693	0.4966	0.5217	0.3860
19	0.5202	0.5230	0.5336	0.5427	0.8889
21	0.3943	0.3898	0.3858	0.3817	0.0433
24	0.3794	0.3758	0.3735	0.3712	0.5108
34	0.4579	0.4540	0.4531	0.4524	0.4759
38	0.5759	0.5810	0.5886	0.5943	0.9519
41	0.4103	0.4047	0.3987	0.3926	0.5451
42	0.4124	0.4066	0.4003	0.3940	0.1266
62	0.5166	0.5198	0.5320	0.5424	0.5853
65	0.5546	0.5570	0.5622	0.5664	0.2969
66	0.7130	0.7290	0.7226	0.7164	0.8574
70	0.4591	0.4587	0.4664	0.4737	0.8220
74	0.4582	0.4552	0.4565	0.4579	1.0880
75	0.4174	0.4113	0.4043	0.3973	0.1597
77	0.5993	0.6075	0.6172	0.6239	0.3877
79	0.5155	0.5138	0.5146	0.5156	0.2255

ลูกค้า (คนที่)	ค่าพยากรณ์จำนวนครั้งการเรียกร้องค่าสินไหมทดแทน				
	QP	CMP	ZIP	ZINB	RF
99	0.5833	0.5893	0.5977	0.6038	1.0850
101	0.5149	0.5158	0.5228	0.5289	0.3018
102	0.3785	0.3750	0.3727	0.3705	0.1046
105	0.4583	0.4558	0.4581	0.4605	0.8559
108	0.5356	0.5403	0.5535	0.5643	0.0725
109	0.5885	0.5953	0.6041	0.6105	0.2707
111	0.3952	0.3890	0.3803	0.3715	0.3273
117	0.5950	0.6026	0.6120	0.6185	0.6695
119	0.3691	0.3662	0.3648	0.3636	0.4987
134	0.3746	0.3695	0.3624	0.3549	0.1742
145	0.3981	0.3911	0.3808	0.3703	0.1039
146	0.4004	0.3955	0.3907	0.3859	0.1024
148	0.4014	0.3964	0.3916	0.3866	0.3924
154	0.5326	0.5367	0.5490	0.5592	0.7149
160	0.5051	0.5072	0.5179	0.5273	0.2296
162	0.6180	0.6288	0.6396	0.6466	0.1839
168	0.5992	0.6075	0.6171	0.6238	0.3877
176	0.4381	0.4341	0.4330	0.4321	0.6532
178	0.5245	0.5271	0.5370	0.5454	0.5287
181	0.4133	0.4059	0.3953	0.3848	0.0204
182	0.4136	0.4078	0.4013	0.3948	0.1136
186	0.4580	0.4545	0.4544	0.4545	0.6371
189	0.4375	0.4356	0.4397	0.4437	0.3791
190	0.4895	0.4874	0.4894	0.4914	0.7768
193	0.4025	0.3974	0.3924	0.3873	0.0009
208	0.4586	0.4569	0.4610	0.4651	0.3033
211	0.5307	0.5345	0.5463	0.5560	0.8407
229	0.4098	0.4043	0.3983	0.3923	0.5310
230	0.4587	0.4571	0.4617	0.4663	0.1619

ลูกค้า (คนที่)	ค่าพยากรณ์จำนวนครั้งการเรียกร้องค่าสินไหมทดแทน				
	QP	CMP	ZIP	ZINB	RF
236	0.4142	0.4083	0.4017	0.3952	0.1501
237	0.5051	0.5042	0.5080	0.5116	0.3635
239	0.4099	0.4044	0.3984	0.3924	0.4848
248	0.4083	0.4028	0.3970	0.3913	1.1231
252	0.4176	0.4115	0.4044	0.3975	0.7633
259	0.3938	0.3893	0.3853	0.3813	0.6489
261	0.5326	0.5368	0.5491	0.5593	0.6618
269	0.4102	0.4040	0.3964	0.3887	0.1259
272	0.3819	0.3773	0.3720	0.3666	0.4321
273	0.4170	0.4096	0.3992	0.3889	0.0303
280	0.5676	0.5716	0.5783	0.5836	0.1732
287	0.4592	0.4590	0.4670	0.4747	0.4388
289	0.5286	0.5320	0.5432	0.5525	1.4122
293	0.3658	0.3631	0.3620	0.3611	0.5804
298	0.4038	0.3986	0.3935	0.3882	0.8855
299	0.4166	0.4105	0.4036	0.3968	1.0426
307	0.5513	0.5532	0.5580	0.5619	0.2035
308	0.4576	0.4531	0.4505	0.4483	0.5720
316	0.5553	0.5577	0.5629	0.5672	1.0260
317	0.5713	0.5828	0.6054	0.6221	0.3271
318	0.5958	0.6036	0.6130	0.6195	0.1416
325	0.5124	0.5128	0.5190	0.5245	0.7170
328	0.4999	0.4982	0.5003	0.5025	0.1980
330	0.3829	0.3765	0.3667	0.3564	0.3168
340	0.6036	0.6133	0.6252	0.6332	0.5786
342	0.5938	0.6016	0.6117	0.6187	0.3372
356	0.5512	0.5541	0.5608	0.5663	0.1526
357	0.4843	0.4816	0.4824	0.4834	0.8494
365	0.5180	0.5195	0.5274	0.5343	0.2409

ลูกค้า (คนที่)	ค่าพยากรณ์จำนวนครั้งการเรียกร้องค่าสินไหมทดแทน				
	QP	CMP	ZIP	ZINB	RF
369	0.9587	1.0291	0.9921	0.9702	0.9901
370	0.4200	0.4137	0.4063	0.3990	0.5323
378	0.4964	0.4965	0.5031	0.5092	0.4618
389	0.3904	0.3861	0.3826	0.3790	0.5082
390	0.3833	0.3778	0.3704	0.3627	0.1547
398	0.4139	0.4081	0.4015	0.3950	0.0639
402	0.4586	0.4569	0.4612	0.4653	0.7154
407	0.4030	0.3979	0.3928	0.3877	0.0130
408	0.4578	0.4540	0.4529	0.4521	1.3469
414	0.6090	0.6185	0.6288	0.6357	1.5034
415	0.4035	0.3983	0.3932	0.3880	0.2615
417	0.7342	0.7529	0.7436	0.7354	0.7133
431	0.6719	0.6832	0.6808	0.6779	1.3503
444	0.4007	0.3957	0.3909	0.3861	0.1066
447	0.4029	0.3978	0.3928	0.3877	0.0130
448	0.4062	0.4009	0.3954	0.3899	1.3282
479	0.3917	0.3873	0.3836	0.3799	0.5545
483	0.5132	0.5138	0.5202	0.5259	0.1685
493	0.3886	0.3844	0.3811	0.3777	0.1615
497	0.5574	0.5601	0.5657	0.5701	0.2457
498	0.6044	0.6133	0.6233	0.6301	0.3957
505	0.4062	0.4009	0.3954	0.3899	1.5417
507	0.4578	0.4540	0.4530	0.4523	0.3171
519	0.5191	0.5208	0.5290	0.5362	0.7184
523	0.7630	0.7993	0.8091	0.8109	0.3673
528	0.4968	0.4951	0.4975	0.4999	0.3358
530	0.4516	0.4488	0.4504	0.4520	0.0724
531	0.3864	0.3801	0.3709	0.3613	0.4483
535	0.4583	0.4556	0.4575	0.4595	0.0806

ลูกค้า (คนที่)	ค่าพยากรณ์จำนวนครั้งการเรียกร้องค่าสินไหมทดแทน				
	QP	CMP	ZIP	ZINB	RF
536	0.3965	0.3918	0.3876	0.3832	1.7526
539	0.4583	0.4558	0.4580	0.4602	0.0314
540	0.3978	0.3930	0.3886	0.3841	0.0933
542	0.6282	0.6348	0.6346	0.6339	0.8168
545	0.3927	0.3868	0.3789	0.3708	0.0399
546	0.5166	0.5178	0.5253	0.5319	0.1919
552	0.5823	0.5882	0.5964	0.6025	1.0476
553	0.4506	0.4491	0.4542	0.4591	0.2550
556	0.4041	0.3989	0.3937	0.3884	0.0211
557	1.1333	1.2442	1.1460	1.1072	0.6679
558	0.5063	0.5057	0.5099	0.5139	0.9249
561	0.4596	0.4604	0.4711	0.4813	0.8465
563	0.4579	0.4542	0.4537	0.4534	0.0135
564	0.4010	0.3960	0.3912	0.3863	0.0442
565	0.3999	0.3950	0.3903	0.3856	0.3993
568	0.5585	0.5613	0.5670	0.5716	0.6891
576	0.4124	0.4066	0.4003	0.3940	0.1161
580	0.5351	0.5397	0.5528	0.5635	0.1030
584	0.3915	0.3853	0.3769	0.3682	0.4573
592	0.3941	0.3895	0.3856	0.3815	0.0091
600	0.5136	0.5143	0.5208	0.5266	0.0868
609	0.4580	0.4544	0.4542	0.4542	0.7258
612	0.4465	0.4422	0.4405	0.4389	0.1945
615	0.5154	0.5164	0.5235	0.5297	0.2957
616	0.7742	0.7978	0.7821	0.7696	1.7802
631	0.6122	0.6223	0.6327	0.6396	0.7165
638	0.5121	0.5125	0.5186	0.5240	0.9623
646	0.3856	0.3816	0.3786	0.3756	0.1970
647	0.5149	0.5158	0.5227	0.5289	0.3049

ลูกค้า (คนที่)	ค่าพยากรณ์จำนวนครั้งการเรียกร้องค่าสินไหมทดแทน				
	QP	CMP	ZIP	ZINB	RF
655	0.3944	0.3899	0.3859	0.3818	1.1300
656	0.5526	0.5547	0.5596	0.5637	0.0105
660	0.4586	0.4566	0.4603	0.4640	0.3524
663	0.4097	0.4041	0.3981	0.3922	0.5547
668	0.3906	0.3864	0.3828	0.3792	0.5184
670	0.5143	0.5151	0.5219	0.5279	0.1940
672	0.4188	0.4126	0.4054	0.3983	0.0214
673	0.4192	0.4130	0.4057	0.3985	0.4116
677	1.0571	1.1483	1.0778	1.0451	0.6022
680	0.4589	0.4580	0.4643	0.4704	0.0161
681	0.3758	0.3725	0.3705	0.3685	0.0470
682	0.4067	0.4013	0.3958	0.3902	2.2547
685	0.7800	0.8044	0.7876	0.7745	1.8557
700	0.3194	0.3194	0.3211	0.3241	0.5459
703	0.3932	0.3888	0.3849	0.3810	0.0061
707	0.5115	0.5119	0.5178	0.5231	0.7029
711	0.3998	0.3949	0.3902	0.3855	1.1624
718	0.4908	0.4896	0.4935	0.4973	0.4259
728	0.5960	0.6038	0.6132	0.6198	0.0887
733	0.4105	0.4049	0.3988	0.3927	0.4369
736	0.4584	0.4560	0.4587	0.4614	1.0155
741	0.5921	0.5993	0.6084	0.6149	0.6331
744	0.3085	0.3092	0.3111	0.3149	0.1243
752	0.5195	0.5212	0.5296	0.5369	0.3426
763	0.4612	0.4666	0.4888	0.5093	0.1467
765	0.5772	0.5828	0.5913	0.5976	0.6158
778	0.3965	0.3918	0.3876	0.3832	1.8753
780	0.5500	0.5521	0.5572	0.5614	0.3867
792	0.5429	0.5437	0.5473	0.5505	0.0856

ลูกค้า (คนที่)	ค่าพยากรณ์จำนวนครั้งการเรียกร้องค่าสินไหมทดแทน				
	QP	CMP	ZIP	ZINB	RF
795	0.4199	0.4136	0.4062	0.3989	0.0971
799	0.4584	0.4561	0.4589	0.4617	0.1164
809	0.5939	0.6121	0.6443	0.6666	0.4682
818	0.3833	0.3795	0.3767	0.3740	1.2536
825	0.4131	0.4073	0.4009	0.3945	0.1819
826	0.4182	0.4120	0.4049	0.3978	0.0586
827	0.5855	0.5918	0.6004	0.6066	1.6420
832	0.3817	0.3770	0.3718	0.3664	0.7390
838	0.5348	0.5393	0.5522	0.5629	0.2623
839	0.6204	0.6321	0.6439	0.6515	0.9722
842	0.2958	0.2971	0.2992	0.3036	0.1773
853	0.4072	0.4018	0.3961	0.3905	0.0582
854	0.3993	0.3944	0.3898	0.3851	0.1394
860	0.5935	0.6009	0.6102	0.6167	0.2516
868	0.5029	0.5017	0.5049	0.5079	0.0292
869	0.5075	0.5093	0.5190	0.5276	0.8837
870	0.5375	0.5379	0.5412	0.5441	1.1536
876	0.4588	0.4577	0.4633	0.4687	0.5724
880	0.5277	0.5268	0.5280	0.5294	0.1395
892	0.6167	0.6273	0.6380	0.6450	1.5037
900	0.4668	0.4660	0.4725	0.4786	0.5211
904	0.5100	0.5108	0.5179	0.5243	0.4204
912	0.4353	0.4321	0.4328	0.4337	0.0372
921	0.3814	0.3777	0.3751	0.3726	0.5684
931	0.5157	0.5168	0.5240	0.5303	0.2311
933	0.4154	0.4094	0.4027	0.3960	0.0182
937	0.4591	0.4586	0.4659	0.4729	0.3954
949	0.5577	0.5605	0.5660	0.5706	0.5123
954	0.3575	0.3553	0.3549	0.3548	0.4852

ลูกค้า (คนที่)	ค่าพยากรณ์จำนวนครั้งการเรียกร้องค่าสินไหมทดแทน				
	QP	CMP	ZIP	ZINB	RF
959	0.4583	0.4556	0.4574	0.4594	0.0685
960	0.4210	0.4146	0.4071	0.3996	0.2619
963	0.5582	0.5610	0.5666	0.5711	0.4172
965	0.4523	0.4484	0.4474	0.4467	0.3301
966	0.3929	0.3867	0.3782	0.3695	0.0732
967	0.3991	0.3943	0.3897	0.3850	0.0380
978	0.7422	0.7722	0.7798	0.7809	0.4922
981	0.5877	0.5944	0.6031	0.6094	0.1251
983	0.4059	0.4006	0.3951	0.3896	0.0122
994	0.3555	0.3490	0.3364	0.3229	0.0984
1003	0.4013	0.3963	0.3914	0.3865	0.0204
1008	0.4582	0.4552	0.4564	0.4577	0.7731
1011	0.4038	0.3986	0.3935	0.3882	0.8755
1020	0.5124	0.5128	0.5190	0.5245	0.5557
1025	0.4040	0.3988	0.3936	0.3883	0.0277
1028	0.5499	0.5519	0.5570	0.5612	0.1471
1033	0.3861	0.3821	0.3791	0.3760	0.5461
1034	0.3869	0.3829	0.3797	0.3765	0.8838
1039	0.5924	0.6001	0.6100	0.6170	0.1561
1040	0.4077	0.4117	0.4315	0.4516	0.3289
1049	0.4095	0.4024	0.3924	0.3823	0.2053
1052	0.4588	0.4576	0.4633	0.4688	0.5682
1055	0.5582	0.5617	0.5690	0.5748	1.0825
1058	0.4817	0.4805	0.4851	0.4894	0.2427
1059	0.4062	0.4009	0.3954	0.3899	1.5417
1060	0.5704	0.5748	0.5818	0.5872	0.5505
1062	0.5624	0.5657	0.5719	0.5767	1.6532
1067	0.3901	0.3859	0.3824	0.3788	0.5532
1069	0.4595	0.4600	0.4701	0.4796	1.7733

ลูกค้า (คนที่)	ค่าพยากรณ์จำนวนครั้งการเรียกร้องค่าสินไหมทดแทน				
	QP	CMP	ZIP	ZINB	RF
1081	0.3988	0.3940	0.3894	0.3848	0.0295
1082	0.5860	0.5937	0.6051	0.6133	0.4395
1090	0.5255	0.5284	0.5386	0.5472	0.0535
1100	0.4139	0.4080	0.4015	0.3950	0.0462
1106	0.4037	0.3985	0.3934	0.3882	0.6209
1115	0.4342	0.4261	0.4146	0.4036	0.1137
1118	0.4002	0.3939	0.3856	0.3772	0.3229
1120	0.4865	0.4847	0.4876	0.4905	0.8997
1121	0.5145	0.5153	0.5222	0.5282	0.2018
1126	0.5096	0.5096	0.5149	0.5197	0.1854
1127	0.5311	0.5349	0.5467	0.5566	0.9205
1128	0.3997	0.3948	0.3902	0.3855	0.4852
1132	0.2951	0.2964	0.2985	0.3029	0.1616
1135	0.4083	0.4028	0.3971	0.3913	1.1382
1139	0.3945	0.3900	0.3859	0.3819	0.1475
1142	0.5311	0.5349	0.5467	0.5566	0.9177
1150	0.4591	0.4585	0.4656	0.4724	0.2146
1151	0.3948	0.3869	0.3738	0.3604	0.2143
1152	0.4580	0.4544	0.4542	0.4543	0.7298
1170	0.4255	0.4188	0.4106	0.4025	0.4492
1172	0.5571	0.5597	0.5652	0.5697	0.6952
1181	0.6169	0.6276	0.6383	0.6453	1.5277
1184	0.4586	0.4569	0.4612	0.4654	1.1585
1186	0.4182	0.4121	0.4049	0.3978	0.5220
1198	0.4589	0.4577	0.4634	0.4688	0.5982
1207	0.3753	0.3695	0.3604	0.3508	0.2607
1208	0.4857	0.4845	0.4890	0.4932	0.3048
1210	0.5536	0.5561	0.5617	0.5663	0.4981
1212	0.4895	0.4861	0.4847	0.4838	0.6505

ลูกค้า (คนที่)	ค่าพยากรณ์จำนวนครั้งการเรียกร้องค่าสินไหมทดแทน				
	QP	CMP	ZIP	ZINB	RF
1214	0.8535	0.8876	0.8545	0.8324	0.0139
1218	0.5935	0.6013	0.6113	0.6184	0.5558
1223	0.3659	0.3632	0.3622	0.3613	0.5705
1225	0.5228	0.5252	0.5346	0.5426	0.3976
1226	0.5043	0.5039	0.5089	0.5134	0.3918
1240	0.6089	0.6196	0.6324	0.6409	0.3714
1244	0.3912	0.3868	0.3832	0.3795	0.3189
1249	0.5039	0.5028	0.5063	0.5096	0.2599
1251	0.4583	0.4558	0.4582	0.4605	0.8742
1253	0.4589	0.4578	0.4637	0.4694	0.1520
1254	0.4043	0.3991	0.3939	0.3886	0.1215
1258	0.4077	0.4022	0.3965	0.3908	0.4794
1268	0.5212	0.5233	0.5322	0.5399	0.6239
1270	0.5138	0.5154	0.5238	0.5313	0.4053
1274	0.4313	0.4242	0.4151	0.4062	0.0483
1275	0.4590	0.4581	0.4647	0.4710	2.0396
1280	0.3994	0.3945	0.3899	0.3852	0.6008
1285	0.5888	0.5964	0.6069	0.6145	0.1945
1287	0.6230	0.6454	0.6785	0.6999	0.1364
1300	0.4481	0.4431	0.4393	0.4359	0.6918
1302	0.4020	0.3969	0.3920	0.3870	0.0278
1305	0.5724	0.5770	0.5843	0.5898	0.4905
1308	0.3890	0.3848	0.3814	0.3780	0.4067
1310	0.4589	0.4579	0.4641	0.4700	0.9604
1313	0.4121	0.4063	0.4001	0.3938	0.1181
1316	0.4008	0.3958	0.3910	0.3862	0.3633
1320	0.5008	0.4992	0.5016	0.5040	0.9272
1321	0.4586	0.4569	0.4613	0.4656	0.3663
1322	0.5628	0.5662	0.5724	0.5773	0.7458

ลูกค้า (คนที่)	ค่าพยากรณ์จำนวนครั้งการเรียกร้องค่าสินไหมทดแทน				
	QP	CMP	ZIP	ZINB	RF
1326	0.3902	0.3851	0.3793	0.3733	0.2638
1330	0.4079	0.4018	0.3944	0.3869	0.1096
1339	0.4102	0.4046	0.3986	0.3926	0.6195
1340	0.4585	0.4565	0.4600	0.4635	1.6729
1341	0.4233	0.4168	0.4089	0.4011	0.1307
1346	0.4318	0.4247	0.4155	0.4065	0.0502
1359	0.3095	0.3101	0.3121	0.3157	0.1423
1370	0.4104	0.4048	0.3987	0.3927	0.1101
1372	0.5069	0.5064	0.5108	0.5149	0.5826
1373	0.4148	0.4089	0.4022	0.3956	0.3896
1374	0.3898	0.3834	0.3740	0.3644	0.1245
1377	0.4134	0.4076	0.4011	0.3947	0.2021
1386	0.4581	0.4548	0.4554	0.4561	0.1628
1390	0.4017	0.3966	0.3917	0.3868	0.0494
1393	0.3884	0.3842	0.3809	0.3776	0.8755
1399	0.5800	0.5857	0.5937	0.5997	1.0706
1400	0.6088	0.6183	0.6286	0.6355	1.3325
1406	0.4090	0.4034	0.3976	0.3917	0.0426
1411	0.5258	0.5287	0.5390	0.5477	0.1026
1417	0.6282	0.6410	0.6533	0.6608	0.1811
1420	0.3965	0.3910	0.3847	0.3782	0.6331
1429	0.4153	0.4093	0.4026	0.3959	0.0514
1447	0.5072	0.5067	0.5112	0.5154	0.3905
1453	0.4122	0.4065	0.4002	0.3939	0.1153
1456	0.4447	0.4413	0.4417	0.4423	0.2531
1459	0.5242	0.5235	0.5256	0.5277	0.2394
1460	0.5111	0.5121	0.5196	0.5262	0.1443
1472	0.6989	0.7226	0.7347	0.7398	0.5648
1475	0.6201	0.6312	0.6421	0.6490	0.6760

ลูกค้า (คนที่)	ค่าพยากรณ์จำนวนครั้งการเรียกร้องค่าสินไหมทดแทน				
	QP	CMP	ZIP	ZINB	RF
1480	0.4108	0.4052	0.3991	0.3930	0.4874
1507	0.9821	1.0573	1.0130	0.9885	0.8413
1508	0.4453	0.4553	0.4907	0.5238	0.1941
1515	0.4589	0.4579	0.4640	0.4699	0.5584
1523	0.5545	0.5568	0.5620	0.5662	0.1127
1524	0.4234	0.4234	0.4323	0.4412	0.7636
1528	0.4006	0.3957	0.3909	0.3861	0.1078
1533	0.3997	0.3934	0.3852	0.3767	0.5711
1544	0.6035	0.6124	0.6223	0.6291	0.4441
1548	0.3380	0.3355	0.3322	0.3291	1.0630
1551	0.7414	0.7609	0.7507	0.7417	0.5766
1552	0.5717	0.5763	0.5835	0.5890	0.0946
1555	0.4590	0.4583	0.4651	0.4717	0.0749
1556	0.4513	0.4488	0.4514	0.4539	0.0540
1557	0.4405	0.4359	0.4331	0.4306	0.5677
1583	0.4304	0.4285	0.4326	0.4368	0.7385
1585	0.4129	0.4071	0.4007	0.3944	0.6281
1593	0.4581	0.4551	0.4560	0.4571	0.1425
1603	0.4122	0.4065	0.4002	0.3939	0.2520
1609	0.3989	0.3941	0.3895	0.3849	0.0110
1610	0.4398	0.4371	0.4389	0.4409	0.2603
1613	0.4603	0.4631	0.4788	0.4935	1.3773
1619	0.4626	0.4718	0.5037	0.5327	0.5179
1624	0.5028	0.5016	0.5046	0.5076	0.0175
1626	0.4411	0.4353	0.4297	0.4245	0.5857
1630	0.3993	0.3944	0.3898	0.3852	0.1414
1638	0.4587	0.4571	0.4618	0.4663	1.1942
1643	0.4089	0.4019	0.3919	0.3818	0.2795
1646	1.4153	1.5434	1.2635	1.1734	0.8865

ลูกค้า (คนที่)	ค่าพยากรณ์จำนวนครั้งการเรียกร้องค่าสินไหมทดแทน				
	QP	CMP	|ZIP	ZINB	RF
1649	0.4017	0.3967	0.3918	0.3868	0.5218
1652	0.4872	0.4851	0.4873	0.4895	0.7635
1658	0.5628	0.5662	0.5724	0.5773	0.7264
1659	0.4040	0.3989	0.3936	0.3884	0.0191
1663	0.5166	0.5204	0.5339	0.5454	0.1748
1665	0.5712	0.5760	0.5838	0.5898	0.3019
1668	0.3923	0.3879	0.3841	0.3803	0.0950
1672	0.6630	0.6734	0.6716	0.6692	0.4809
1678	0.7974	0.8427	0.8523	0.8531	1.3070
1681	0.6454	0.6602	0.6718	0.6785	1.0054
1684	0.5419	0.5435	0.5488	0.5532	0.4366
1687	0.5219	0.5249	0.5362	0.5457	0.2236
1689	0.3988	0.3940	0.3894	0.3848	0.0236
1699	0.4619	0.4689	0.4955	0.5198	0.6189
1700	0.5395	0.5448	0.5591	0.5707	0.4650
1701	0.7223	0.7680	0.8113	0.8331	0.5446
1702	0.5251	0.5279	0.5380	0.5465	0.5668
1703	0.5751	0.5813	0.5914	0.5989	0.3789
1704	0.4081	0.4026	0.3969	0.3911	0.2647
1705	0.4222	0.4158	0.4080	0.4004	1.3840
1716	0.4061	0.3994	0.3902	0.3809	0.5499
1722	0.6572	0.6669	0.6655	0.6634	0.7459
1738	0.5885	0.5952	0.6040	0.6104	0.2181
1745	0.6431	0.6575	0.6692	0.6759	0.3985
1748	0.4858	0.4834	0.4849	0.4865	0.4734
1753	0.5033	0.5028	0.5076	0.5120	0.7136
1754	0.3838	0.3800	0.3772	0.3743	0.3584
1758	0.5055	0.5077	0.5187	0.5283	0.5002
1763	0.4048	0.3996	0.3943	0.3889	0.1839

ลูกค้า (คนที่)	ค่าพยากรณ์จำนวนครั้งการเรียกร้องค่าสินไหมทดแทน				
	QP	CMP	|ZIP	ZINB	RF
1774	0.5554	0.5579	0.5631	0.5674	1.5468
1777	0.4086	0.4031	0.3973	0.3915	0.0112
1778	0.4067	0.4014	0.3958	0.3902	1.5144
1783	0.4503	0.4492	0.4554	0.4614	0.0530
1787	0.5155	0.5166	0.5237	0.5300	0.2172
1790	0.3870	0.3830	0.3798	0.3766	0.7327
1792	0.6235	0.6356	0.6476	0.6552	0.2562
1794	0.5050	0.5042	0.5080	0.5116	0.3833
1805	0.4057	0.4004	0.3950	0.3895	0.0068
1806	0.4444	0.4412	0.4420	0.4429	0.0637
1817	0.5754	0.5812	0.5903	0.5971	0.4243
1822	0.4050	0.3978	0.3873	0.3767	0.3334
1823	0.3967	0.3906	0.3825	0.3742	0.0440
1829	0.5180	0.5195	0.5274	0.5343	0.3320
1838	0.5562	0.5588	0.5642	0.5686	0.1043
1839	0.4513	0.4488	0.4515	0.4541	0.2170
1843	0.3895	0.3845	0.3787	0.3727	0.0439
1844	0.5661	0.5705	0.5785	0.5847	0.7045
1849	0.4100	0.4044	0.3984	0.3924	1.5649
1858	0.3968	0.3921	0.3878	0.3834	0.4452
1863	0.5160	0.5180	0.5272	0.5352	0.2028
1865	0.4587	0.4569	0.4613	0.4656	0.6186
1871	0.5721	0.5767	0.5840	0.5895	0.1297
1876	0.4524	0.4484	0.4472	0.4463	0.5397
1879	0.4585	0.4564	0.4598	0.4632	0.8179
1887	0.4875	0.4882	0.4968	0.5047	0.5517
1893	0.3693	0.3641	0.3563	0.3480	0.2258
1895	0.4588	0.4575	0.4628	0.4679	0.5687
1898	0.4584	0.4558	0.4582	0.4605	0.8825

ลูกค้า (คนที่)	ค่าพยากรณ์จำนวนครั้งการเรียกร้องค่าสินไหมทดแทน				
	QP	CMP	|ZIP	ZINB	RF
1900	0.4037	0.3979	0.3909	0.3838	0.5641
1904	0.5950	0.6027	0.6120	0.6185	0.6695
1911	0.4086	0.4031	0.3973	0.3915	0.0112
1912	0.4068	0.4014	0.3959	0.3903	0.7068
1922	0.3536	0.3517	0.3516	0.3518	0.3913
1924	0.6044	0.6133	0.6233	0.6301	0.4161
1932	0.5651	0.5688	0.5752	0.5803	0.3267
1935	0.4592	0.4589	0.4669	0.4745	0.3447
1939	0.4517	0.4487	0.4500	0.4514	0.5305
1940	0.3984	0.3936	0.3891	0.3845	0.0100
1957	0.4583	0.4557	0.4579	0.4601	0.1247
1961	0.7460	0.7660	0.7551	0.7457	0.6515
1964	0.4041	0.3989	0.3937	0.3885	0.0416
1966	0.4581	0.4549	0.4555	0.4563	1.0627
1986	0.4972	0.4956	0.4981	0.5006	0.4344
1987	0.4007	0.3957	0.3909	0.3861	0.1066
1991	0.5091	0.5090	0.5141	0.5188	0.6476
1998	0.4193	0.4131	0.4058	0.3986	0.3871
2004	0.6664	0.6772	0.6751	0.6725	2.5243
2005	0.3972	0.3925	0.3881	0.3837	0.8186
2008	0.4973	0.4951	0.4964	0.4978	0.7068
2021	0.6611	0.6782	0.6900	0.6962	1.4470
2028	0.5471	0.5487	0.5534	0.5573	0.5452
2034	0.4520	0.4486	0.4488	0.4492	0.5494
2035	0.5526	0.5549	0.5604	0.5649	0.1278
2039	0.3948	0.3902	0.3862	0.3820	0.4619
2048	0.5544	0.5705	0.6064	0.6336	0.3192
2053	0.5765	0.5817	0.5894	0.5952	0.9603
2058	0.3942	0.3897	0.3857	0.3816	0.0000

ลูกค้า (คนที่)	ค่าพยากรณ์จำนวนครั้งการเรียกร้องค่าสินไหมทดแทน				
	QP	CMP	|ZIP	ZINB	RF
2063	0.4086	0.4031	0.3973	0.3915	0.0112
2069	0.4018	0.3967	0.3918	0.3868	0.2735
2073	0.4138	0.4080	0.4014	0.3949	0.5888
2075	0.4454	0.4371	0.4253	0.4141	0.0023
2078	0.5073	0.5069	0.5114	0.5156	0.3777
2081	0.5831	0.5891	0.5975	0.6036	0.2852
2085	0.4580	0.4545	0.4544	0.4545	0.6371
2087	0.3886	0.3844	0.3811	0.3777	0.1769
2090	0.4050	0.3991	0.3920	0.3848	0.3170
2091	0.4179	0.4117	0.4046	0.3976	0.0983
2094	0.5950	0.6030	0.6132	0.6203	0.5984
2098	0.4167	0.4106	0.4037	0.3968	0.2293
2099	0.3475	0.3459	0.3463	0.3471	0.0931
2101	0.3949	0.3903	0.3862	0.3821	0.1869
2103	0.5319	0.5358	0.5479	0.5579	0.1686
2108	0.5388	0.5440	0.5581	0.5695	0.9203
2112	0.5789	0.5844	0.5924	0.5983	1.0344
2114	0.3974	0.3927	0.3883	0.3839	0.0905
2130	0.4581	0.4550	0.4557	0.4566	0.5028
2131	0.5467	0.5483	0.5530	0.5569	0.4324
2132	0.5305	0.5342	0.5459	0.5557	0.7237
2140	0.4050	0.3998	0.3944	0.3891	0.0828
2142	0.8588	0.9096	0.8987	0.8881	1.4654
2147	0.5880	0.6090	0.6487	0.6764	0.4820
2148	0.4894	0.4845	0.4798	0.4759	0.0168
2168	0.4091	0.4036	0.3977	0.3918	0.0191
2172	0.4579	0.4541	0.4534	0.4529	0.8747
2175	0.5672	0.5722	0.5813	0.5882	0.4931
2191	0.4138	0.4079	0.4014	0.3949	0.2640

ลูกค้า (คนที่)	ค่าพยากรณ์จำนวนครั้งการเรียกร้องค่าสินไหมทดแทน				
	QP	CMP	|ZIP	ZINB	RF
2193	0.7164	0.7422	0.7520	0.7551	0.5699
2195	0.5693	0.5735	0.5804	0.5857	0.2912
2196	0.5950	0.6026	0.6119	0.6185	0.6695
2206	0.5578	0.5605	0.5661	0.5706	0.5139
2214	0.4089	0.4034	0.3975	0.3917	0.0590
2216	0.4590	0.4582	0.4648	0.4711	0.9617
2217	0.5209	0.5229	0.5317	0.5392	0.1956
2218	0.5939	0.6028	0.6150	0.6235	0.5211
2225	0.4914	0.4904	0.4946	0.4986	0.3597
2226	0.3969	0.3902	0.3807	0.3709	0.6130
2228	0.4176	0.4115	0.4044	0.3975	0.7670
2232	0.6164	0.6270	0.6377	0.6447	0.4132
2243	0.4878	0.4864	0.4900	0.4935	1.2472
2244	0.3927	0.3882	0.3844	0.3806	0.2499
2246	0.3961	0.3914	0.3872	0.3829	1.2335
2247	0.4849	0.4847	0.4916	0.4979	0.4967
2248	0.5013	0.4998	0.5024	0.5049	0.8720
2256	0.5321	0.5362	0.5483	0.5584	1.2375
2262	0.4343	0.4313	0.4328	0.4344	0.2279
2265	0.5030	0.5038	0.5115	0.5185	0.9057
2282	0.4183	0.4122	0.4050	0.3979	0.3128
2286	0.4583	0.4557	0.4578	0.4599	0.0871
2290	0.5842	0.5904	0.5988	0.6049	0.6997
2298	0.5865	0.5933	0.6027	0.6095	0.8150
2300	0.5812	0.5874	0.5963	0.6028	0.1275
2312	0.4089	0.4034	0.3975	0.3917	0.0590
2313	0.4108	0.4052	0.3991	0.3929	0.4544
2315	0.4581	0.4548	0.4553	0.4560	0.0146
2316	0.5824	0.5884	0.5967	0.6027	1.0060

ลูกค้า (คนที่)	ค่าพยากรณ์จำนวนครั้งการเรียกร้องค่าสินไหมทดแทน				
	QP	CMP	|ZIP	ZINB	RF
2320	0.4068	0.4014	0.3959	0.3903	0.2582
2323	0.4584	0.4560	0.4588	0.4615	0.9955
2329	0.4515	0.4488	0.4507	0.4527	0.3060
2330	0.4056	0.4003	0.3949	0.3894	1.0989
2344	0.4840	0.4835	0.4897	0.4955	0.7148
2345	0.5484	0.5499	0.5543	0.5579	0.5910
2351	0.4866	0.4835	0.4835	0.4837	0.6022
2363	0.3971	0.3924	0.3881	0.3837	1.4198
2365	0.4591	0.4585	0.4656	0.4724	0.1015
2371	0.5284	0.5317	0.5427	0.5520	0.1525
2374	0.4049	0.3996	0.3943	0.3890	0.1859
2378	0.3816	0.3755	0.3663	0.3568	1.3175
2380	0.4594	0.4596	0.4689	0.4777	1.6098
2383	0.4170	0.4104	0.4021	0.3938	0.4005
2385	0.5115	0.5119	0.5178	0.5231	0.7029
2389	0.3987	0.3939	0.3893	0.3847	0.0117
2393	0.4181	0.4119	0.4048	0.3977	0.0187
2399	0.4152	0.4093	0.4026	0.3959	0.1369
2401	0.4580	0.4546	0.4547	0.4550	0.8567
2407	0.4060	0.4007	0.3952	0.3897	0.0328
2408	0.4927	0.4927	0.4996	0.5059	0.1426
2411	0.5050	0.5048	0.5102	0.5151	0.5934
2422	0.3930	0.3886	0.3847	0.3808	0.1041
2424	0.4405	0.4530	0.4952	0.5348	0.6644
2430	0.6082	0.6177	0.6279	0.6348	0.7468
2432	0.4017	0.3966	0.3917	0.3868	1.0508
2434	0.4889	0.4867	0.4884	0.4901	0.5154
2436	0.5133	0.5139	0.5203	0.5261	0.1245
2444	0.4122	0.4064	0.4001	0.3939	0.0993

ลูกค้า (คนที่)	ค่าพยากรณ์จำนวนครั้งการเรียกร้องค่าสินไหมทดแทน				
	QP	CMP	ZIP	ZINB	RF
2449	0.6853	0.6982	0.6946	0.6907	1.2030
2451	0.4047	0.3995	0.3942	0.3888	0.0110
2465	0.4008	0.3959	0.3911	0.3862	0.5527
2476	0.4588	0.4576	0.4630	0.4683	0.8774
2482	0.5180	0.5195	0.5274	0.5343	0.3158
2486	0.5382	0.5387	0.5420	0.5451	0.5799
2490	0.4512	0.4489	0.4519	0.4550	0.1057
2496	0.3170	0.3171	0.3189	0.3221	0.5399
2502	0.4480	0.4501	0.4642	0.4777	1.0920
2509	0.5593	0.5625	0.5689	0.5740	0.6022
2520	0.4940	0.4935	0.4989	0.5040	0.7956
2525	0.4242	0.4176	0.4096	0.4017	0.2630
2540	0.5671	0.5711	0.5778	0.5830	1.2356
2541	0.4066	0.4013	0.3957	0.3901	0.0783
2546	0.5569	0.5691	0.5959	0.6163	1.1415
2547	0.4514	0.4488	0.4511	0.4534	0.0778
2554	0.6459	0.6608	0.6724	0.6791	0.1704
2563	0.5129	0.5143	0.5224	0.5296	0.8406
2565	0.4582	0.4552	0.4563	0.4576	0.2805
2567	0.4585	0.4565	0.4601	0.4636	0.5106
2568	0.4093	0.4038	0.3979	0.3920	0.0743
2577	0.6437	0.6583	0.6699	0.6767	0.5665
2593	0.4509	0.4490	0.4529	0.4567	0.4806
2597	0.4579	0.4541	0.4534	0.4529	0.8739
2609	0.5766	0.5817	0.5894	0.5952	0.9603
2613	0.4102	0.4046	0.3986	0.3926	0.4715
2624	0.4116	0.4059	0.3997	0.3935	0.2078
2627	0.4016	0.3966	0.3917	0.3868	0.0494
2640	0.4144	0.4069	0.3961	0.3854	0.0661

ลูกค้า (คนที่)	ค่าพยากรณ์จำนวนครั้งการเรียกร้องค่าสินไหมทดแทน				
	QP	CMP	ZIP	ZINB	RF
2645	0.6059	0.6156	0.6268	0.6343	0.3599
2656	0.4149	0.4089	0.4023	0.3956	0.6110
2658	0.4067	0.4014	0.3958	0.3902	1.7879
2659	0.4180	0.4119	0.4048	0.3977	0.1082
2660	0.4580	0.4547	0.4550	0.4554	0.5119
2662	0.5035	0.5044	0.5124	0.5195	0.5657
2663	0.4417	0.4398	0.4439	0.4480	0.2474
2667	0.4586	0.4569	0.4611	0.4652	0.1719
2669	0.3972	0.3924	0.3881	0.3837	0.6654
2671	0.5268	0.5296	0.5394	0.5478	0.6225
2675	0.7233	0.7406	0.7329	0.7257	0.5301
2678	0.5584	0.5622	0.5700	0.5762	0.9897
2687	0.5387	0.5491	0.5750	0.5955	0.9601
2708	0.5037	0.5033	0.5082	0.5127	0.3350
2725	0.4126	0.4062	0.3984	0.3905	0.3224
2729	0.3529	0.3510	0.3509	0.3513	1.5067
2734	0.4586	0.4566	0.4604	0.4641	0.0633
2738	0.4415	0.4366	0.4331	0.4299	1.0444
2760	0.3819	0.3764	0.3691	0.3614	0.2013
2764	0.4070	0.4016	0.3960	0.3904	0.9781
2767	0.4600	0.4618	0.4751	0.4875	1.5062
2772	0.5235	0.5260	0.5356	0.5438	0.5063
2781	0.5841	0.5906	0.5998	0.6065	0.3428
2787	0.4999	0.4987	0.5022	0.5056	0.6152
2788	0.5785	0.5839	0.5918	0.5977	2.1072
2795	0.4509	0.4490	0.4531	0.4571	0.3560
2799	0.3993	0.3944	0.3898	0.3851	0.1414
2800	0.4584	0.4560	0.4587	0.4615	1.0195
2805	0.7694	0.7924	0.7776	0.7656	0.8172

ลูกค้า (คนที่)	ค่าพยากรณ์จำนวนครั้งการเรียกร้องค่าสินไหมทดแทน				
	QP	CMP	ZIP	ZINB	RF
2809	0.4584	0.4561	0.4590	0.4619	0.1235
2815	0.4340	0.4311	0.4328	0.4346	0.0435
2821	0.4000	0.3951	0.3904	0.3857	0.1551
2823	0.2998	0.3009	0.3030	0.3072	0.1215
2831	0.3923	0.3879	0.3841	0.3803	0.0950
2832	0.3848	0.3809	0.3780	0.3750	0.1308
2834	0.4091	0.4036	0.3977	0.3918	0.0191
2837	0.5666	0.5716	0.5805	0.5875	0.1672
2839	0.5390	0.5404	0.5456	0.5502	0.5715
2846	0.3475	0.3459	0.3462	0.3471	0.0931
2849	0.5241	0.5267	0.5364	0.5448	0.6832
2851	0.4166	0.4106	0.4036	0.3968	0.4851
2853	0.3990	0.3941	0.3896	0.3850	0.0243
2856	0.4643	0.4781	0.5219	0.5613	1.6568
2859	0.4087	0.4032	0.3974	0.3915	0.0148
2861	0.5152	0.5161	0.5231	0.5294	0.1908
2866	0.4048	0.3995	0.3942	0.3889	0.0291
2875	0.4579	0.4543	0.4540	0.4539	0.2481
2895	0.5394	0.5448	0.5590	0.5706	0.4610
2902	0.5117	0.5120	0.5180	0.5233	0.5570
2905	0.4164	0.4104	0.4035	0.3967	0.0692
2942	0.4580	0.4545	0.4545	0.4547	0.5604
2957	0.4104	0.4047	0.3987	0.3926	0.2420
2966	0.5161	0.5180	0.5272	0.5353	0.1882
2967	0.4582	0.4551	0.4562	0.4574	0.5448
2978	0.6175	0.6288	0.6405	0.6481	0.4817
2981	0.4047	0.3995	0.3942	0.3888	0.0170
2984	0.4580	0.4547	0.4551	0.4557	0.0786
2986	0.5051	0.5050	0.5104	0.5153	0.5638

ประวัติผู้เขียน

ชื่อ	นายธนสิทธิ์ พูลสมบัติ
วันเดือนปีเกิด	30 มิถุนายน พ.ศ. 2538
วุฒิการศึกษา	ปีการศึกษา 2560: วิทยาศาสตรบัณฑิต สาขาสถิติ คณะวิทยาศาสตร์ มหาวิทยาลัยสงขลานครินทร์

ผลงานทางวิชาการ

ธนสิทธิ์ พูลสมบัติ, แสงดาว วงค์สาย และ ธีระวัฒน์ สิมมาจันทร์. (2563). การประชุมวิชาการระดับชาติ วิทยาศาสตร์ เทคโนโลยีและนวัตกรรม (มหาวิทยาลัยแม่โจ้) ครั้งที่ 1. *การประยุกต์ใช้ตัวแบบจำลองคอนเวย์-แม็กซ์เวลล์-ปัวซง กับจำนวนผู้เสียชีวิตจากอุบัติเหตุทางถนนในเขตกรุงเทพมหานคร* (หน้า 194-306). แม่โจ้: มหาวิทยาลัยแม่โจ้.