วิธีการประมาณค่าสูญหายในตัวแบบประมาณค่าสมการทั่วไปของข้อมูลระยะยาว

นางสาวนฤมล คุ้มปิยะผล

วิทยานิพนธ์นี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตรปริญญาสถิติศาสตรมหาบัณฑิต
สาขาวิชาสถิติ ภาควิชาสถิติ
คณะพาณิชยศาสตร์และการบัญชี จุฬาลงกรณ์มหาวิทยาลัย
ปีการศึกษา 2553

THE ESTIMATION METHODS FOR MISSING DATA IN GENERALIZED ESTIMATING EQUATIONS MODEL OF LONGITUDINAL DATA

Miss Narumol Kumpiyaphol

A Thesis Submitted in Partial Fulfillment of the Requirements
for the Degree of Master of Science Program in Statistics

Department of Statistics
Faculty of Commerce and Accountancy
Chulalongkorn University

Academic Year 2010

หัวข้อวิทยานิพนธ์	วิธีการประมาณค่าสูญหายในตัวแบบประมาณค่าสมการทั่วไปของข้อมูลระยะยาว
โดย	นางสาวนฤมล คุ้มปิยะผล
สาขาวิชา	สถิติ
อาจารย์ที่ปรึกษาวิทยานิพนธ์หลัก	รองศาสตราจารย์ ดร. กัลยา วานิชย์บัญชา

คณะพาณิชยศาสตร์และการบัญชี จุฬาลงกรณ์มหาวิทยาลัย อนุมัติให้นับวิทยานิพนธ์ฉบับนี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตรปริญญามหาบัณฑิต

..... คณบดีคณะพาณิชยศาสตร์และการบัญชี

(รองศาสตราจารย์ ดร.อรรณพ ตันละมัย)

คณะกรรมการสอบวิทยานิพนธ์

..... ประธานกรรมการ

(รองศาสตราจารย์ ดร. ธีระพร วีระถาวร)

..... อาจารย์ที่ปรึกษาวิทยานิพนธ์หลัก

(รองศาสตราจารย์ ดร. กัลยา วานิชย์บัญชา)

..... กรรมการ

(รองศาสตราจารย์ ดร. สุพล ดุรงค์วัฒนา)

..... กรรมการภายนอกมหาวิทยาลัย

(อาจารย์ ดร.อรุณี กำลัง)

นฤมล คุ้มปิยะผล : วิธีการประมาณค่าสูญหายในตัวแบบประมาณค่าสมการทั่วไปของข้อมูลระยะยาว. (THE ESTIMATION METHODS FOR MISSING DATA IN GENERALIZED ESTIMATING EQUATIONS MODEL OF LONGITUDINAL DATA) อ. ที่ปรึกษาวิทยานิพนธ์หลัก : รศ. ดร. กัลยา วานิชย์บัญชา, 67 หน้า.

งานวิจัยครั้งนี้มีวัตถุประสงค์เพื่อศึกษาและเปรียบเทียบวิธีการประมาณค่าสูญหายของตัวแปรตามในตัวแบบประมาณค่าสมการทั่วไป (Generalized Estimating Equations) ของข้อมูลระยะยาว เมื่อตัวแปรตามเกิดอัตตสหสัมพันธ์อันดับที่หนึ่ง (First Order Autoregressive : AR(1)) โดยทำการประมาณค่าสูญหายด้วยวิธี Last Observation Carried Forward (LOCF) วิธี Previous Row Mean และวิธี Multiple Imputation (MI) ซึ่งการเปรียบเทียบกระทำภายใต้เงื่อนไขของค่าอัตตสหสัมพันธ์ 0.2, 0.5 และ 0.9 ขนาดตัวอย่าง 60 และ 90 ระยะเวลาที่ทำการเก็บข้อมูลซ้ำ 3 และ 5 คาบเวลา ร้อยละการสูญหายของตัวแปรเป็น 10 และ 15 ตามลำดับ ซึ่งข้อมูลที่ใช้ในการวิจัยครั้งนี้ได้จากการจำลองด้วยเทคนิคมอนติคาร์โลด้วยโปรแกรม R โดยใช้เกณฑ์การเปรียบเทียบด้วยวิธีค่าเฉลี่ยร้อยละความคลาดเคลื่อนสัมบูรณ์ (Mean Absolute Percentage Error : (MAPE))

ผลการวิจัยสรุปได้ดังนี้ กรณีที่อัตตสหสัมพันธ์ระดับต่ำ (ρ เท่ากับ 0.2) เมื่อร้อยละการสูญหายเป็น 10 วิธี Previous Row Mean จะให้ค่า MAPE ต่ำที่สุด ยกเว้นที่ขนาดตัวอย่าง 90 และระยะเวลาในการเก็บข้อมูลซ้ำ 5 คาบเวลา วิธี Multiple Imputation จะให้ค่า MAPE ต่ำที่สุด แต่เมื่อร้อยละการสูญหายเพิ่มขึ้นเป็น 15 วิธี Multiple Imputation จะให้ค่า MAPE ต่ำที่สุด กรณีที่อัตตสหสัมพันธ์ระดับปานกลาง (ρ เท่ากับ 0.5) เมื่อขนาดตัวอย่าง 60 วิธี Last Observation Carried Forward จะให้ค่า MAPE ต่ำที่สุด แต่เมื่อขนาดตัวอย่างเป็น 90 วิธี Multiple Imputation จะให้ค่า MAPE ต่ำที่สุด ยกเว้น ที่ระยะเวลาในการเก็บข้อมูลซ้ำ 3 คาบเวลา ร้อยละการสูญหายเป็น 10 ที่ทุกขนาดตัวอย่าง วิธี Previous Row Mean จะให้ค่า MAPE ต่ำที่สุด และกรณีที่อัตตสหสัมพันธ์ระดับสูง (ρ เท่ากับ 0.9) วิธี Multiple Imputation จะให้ค่า MAPE ต่ำที่สุด ทุกขนาดตัวอย่าง ทุกร้อยละการสูญหาย และทุกระยะเวลาที่ทำการเก็บข้อมูลซ้ำ

ภาควิชา.....สถิติ..... ลายมือชื่อนิสิต.....นฤมล คุ้มปิยะผล.....

สาขาวิชา.....สถิติ..... ลายมือชื่อ อ.ที่ปรึกษาวิทยานิพนธ์หลัก.....

ปีการศึกษา.....2553.....

5181833226 : MAJOR STATISTICS

KEYWORDS : MISSING DATA / GENERALIZED ESTIMATING EQUATIONS / LONGITUDINAL DATA / FIRST ORDER AUTOREGRESSIVE

NARUMOL KUMPIYAPHOL : THE ESTIMATION METHODS FOR MISSING DATA IN GENERALIZED ESTIMATING EQUATIONS MODEL OF LONGITUDINAL DATA. THESIS ADVISOR : ASSOC.PROF. KANLAYA VANICHBUNCHA, Ph.D., 67 pp.

The purpose of this research is to study and compare the estimation methods for missing data of the dependent variable in Generalized Estimating Equations model of longitudinal data when dependent variable follow a first order auto-regressive (AR(1)) process. The methods used to estimate missing data are Last Observation Carried Forward (LOCF), Previous Row Mean (PRM) and Multiple Imputation (MI) method. The study is compared under the condition of auto-regressive of 0.2, 0.5 and 0.9; sample size of 60 and 90; 3 and 5 periods of replicate; percentage of missing data of 10 and 15. The data are simulated by R program. Mean Absolute Percentage Error (MAPE) is used as criterion for determination. The results for low autoregressive level (0.2), the lowest MAPE of PRM which percentage of missing data is 10, except that the sample size is 90 and periods of replicate are 5, the MAPE of MI is the lowest. However, If the percentage of missing data is 15, the MAPE of MI is the lowest. In the case when medium autoregressive level (0.5), the lowest MAPE of LOCF which sample size is 60 and the lowest MAPE of MI which sample size is 90, except that the periods of replicate is 3, percentage of missing data is 10 and all sample size, the MAPE of PRM is the lowest. In the case when high autoregressive level (0.9), it is found that the lowest MAPE of MI which all sample size , all periods of replicate data and all percentage of missing data.

Department : Statistics Student's Signature นฤมล กุ้มปิยะผล

Field of Study : Statistics Advisor's Signature Kanlaya Vanichbuncha

Academic Year : 2010

กิตติกรรมประกาศ

วิทยานิพนธ์ฉบับนี้สำเร็จลุล่วงไปได้ด้วยความกรุณาและได้รับความช่วยเหลืออย่างดียิ่งจาก รองศาสตราจารย์ ดร. กัลยา วานิชย์บัญชา อาจารย์ที่ปรึกษาวิทยานิพนธ์ ผู้วิจัยขอกราบขอบพระคุณเป็นอย่างสูง ที่กรุณาให้คำแนะนำ คำปรึกษาตลอดจนช่วยเหลือตรวจสอบแก้ไขข้อบกพร่องต่าง ๆ จนกระทั่งวิทยานิพนธ์ฉบับนี้เสร็จสมบูรณ์

ผู้วิจัยขอกราบขอบพระคุณ รองศาสตราจารย์ ดร. ธีระพร วีระถาวร ประธานกรรมการ รองศาสตราจารย์ ดร. สุพล ดุรงค์วัฒนา กรรมการสอบวิทยานิพนธ์ และ อาจารย์ ดร. อรุณี กำลัง กรรมการภายนอกมหาวิทยาลัย ที่กรุณาตรวจสอบวิทยานิพนธ์ฉบับนี้ให้สมบูรณ์ยิ่งขึ้น และขอกราบขอบคุณคณาจารย์ทุกท่านที่กรุณาถ่ายทอดความรู้แก่ผู้วิจัย จนกระทั่งสำเร็จการศึกษา

ผู้วิจัยขอกราบขอบพระคุณ บิดา มารดา ซึ่งสนับสนุนด้านการเรียน ให้คำแนะนำและให้กำลังใจแก่ผู้วิจัยเสมอมาจนสำเร็จการศึกษา และขอขอบคุณเพื่อน ๆ พี่ ๆ ทุกคนที่คอยช่วยเหลือผู้วิจัย ให้คำปรึกษา รวมทั้งให้กำลังใจในเรื่องต่าง ๆ มา ณ โอกาสนี้ด้วย

สารบัญ

บทที่ หน้า

สารบัญตาราง

สารบัญภาพ

บทที่ 1

บทนำ

ความเป็นมาและความสำคัญของปัญหา

ในองค์กรต่าง ๆ การทำงานหรือดำเนินงานจะต้องมีการวางแผนเพื่อการตัดสินใจที่ถูกต้องและลดความเสี่ยงจากความไม่แน่นอนของเหตุการณ์ที่จะเกิดขึ้น ในการวางแผนจึงต้องมีการเก็บรวบรวมข้อมูล เพื่ออาศัยความรู้ทางสถิติมาช่วยในการวิเคราะห์สิ่งที่จะเกิดขึ้นในอนาคต เพื่อให้องค์กรนำผลที่ได้จากการพยากรณ์ไปใช้กับงานของตนตรงตามวัตถุประสงค์ได้อย่างมีประสิทธิภาพ

ข้อมูลจึงเป็นสิ่งสำคัญมากสำหรับการทำงานหรือดำเนินงานในองค์กรต่าง ๆ แต่ในบางครั้งการเก็บข้อมูลต้องใช้ระยะเวลานานในการเก็บเพื่อการพยากรณ์ที่ถูกต้องแม่นยำ ไม่ว่าจะเป็นข้อมูลทางการแพทย์ การเกษตร หรือการศึกษา การเก็บข้อมูลจะต้องทำการเก็บซ้ำจากหน่วยตัวอย่างเดียวกันมากกว่า 1 ครั้งในระยะเวลาที่ต่างกัน ข้อมูลประเภทนี้เรียกว่า ข้อมูลระยะยาว (Longitudinal data) ซึ่งข้อมูลประเภทนี้จะมีความสัมพันธ์กันตามเวลา เพราะเก็บมาจากหน่วยตัวอย่างเดียวกัน ซึ่งในความเป็นจริงทางปฏิบัติส่วนใหญ่ ข้อมูลที่ได้ในปัจจุบันจะมีความสัมพันธ์กับข้อมูลในอดีตที่ใกล้กันมากกว่าข้อมูลในอดีตที่ไกลออกไป จึงทำให้เกิดปัญหาความคลาดเคลื่อนอัตตสหสัมพันธ์ต่อกัน ซึ่งลักษณะโครงสร้างของความคลาดเคลื่อนโดยทั่วไปที่เกิดปัญหานี้คือ อัตตสหสัมพันธ์อันดับที่หนึ่ง (First Order Autoregressive : AR(1))

การพยากรณ์ต้องอาศัยความรู้ทั้งทางด้านคณิตศาสตร์และสถิติมาสร้างสมการพยากรณ์เพื่อใช้ในการอธิบายความสัมพันธ์ระหว่างตัวแปรหรือทำนายสิ่งที่จะเกิดขึ้นในอนาคต ซึ่งในการวิเคราะห์ข้อมูลส่วนใหญ่นั้นมักจะใช้วิธีการวิเคราะห์การถดถอยเชิงเส้น สำหรับข้อมูลที่ค่าสังเกตเป็นอิสระกันและไม่มีการวัดซ้ำ แต่สำหรับข้อมูลระยะยาว ข้อมูลค่าสังเกตจะมีการวัดซ้ำและมีความสัมพันธ์กัน ดังนั้น การวิเคราะห์ข้อมูลก็ควรจะเหมาะสมกับลักษณะของข้อมูลด้วย และวิธีการวิเคราะห์ที่นำมาพิจารณาคือ Generalized Estimating Equations ซึ่งนำโครงสร้างความสัมพันธ์ของข้อมูลที่เก็บซ้ำของหน่วยตัวอย่างเดียวกันที่เรียกว่า Working Correlation Matrix เข้ามาพิจารณาในขั้นตอน ของการประมาณพารามิเตอร์ด้วย

สำหรับการวิเคราะห์ข้อมูลระยะยาว สิ่งที่เกิดขึ้นบ่อยครั้งและแทบจะหลีกเลี่ยงไม่ได้เลยนั่นคือ การเกิดข้อมูลสูญหาย (Missing Data) ซึ่งในทางปฏิบัติส่วนใหญ่มักเลือกที่จะตัดข้อมูลทิ้งไป และนำเพียงส่วนที่สมบูรณ์มาใช้ในการวิเคราะห์แทน ซึ่งทำให้ข้อมูลมีจำนวนลดลงและอาจมีผลทำให้สูญเสียรายละเอียดบางอย่างไป ทำให้กระทบต่อผลสรุปในการวิเคราะห์ได้ ดังนั้น ใน

งานวิจัยฉบับนี้จะพิจารณาวิธีการจัดการกับข้อมูลที่สูญหายไป เพื่อสามารถนำข้อมูลที่ได้จากการจัดการไปประมาณค่าต่อไปได้อย่างมีประสิทธิภาพ

นักสถิติหลายท่านได้ทำการศึกษางานวิจัยที่เกี่ยวกับการประมาณค่าสูญหายในการวิเคราะห์ข้อมูลระยะยาว ดังนี้

ในปี ค.ศ. 2002 Jean Mundahl Engels และ Paula Diehr ได้ทำการศึกษาเปรียบเทียบวิธีการประมาณค่าสูญหายในข้อมูลระยะยาว 14 วิธีดังนี้ Column mean , Class mean , Column median , Class median , Hot Deck , Regression , Regression with error , Previous row mean , Previous row median , Last observation carried forward , Row mean, Row median , Next observation carried backwad และ Last & Next

ในปี ค.ศ. 2004 Jos W.R. Twisk ได้ทำการศึกษาและเปรียบเทียบตัวแบบสองตัวคือ Generalized estimating equations และ random coefficient analysis เมื่อข้อมูลเกิดการสูญหายในระยะยาว

ศุภลักษณ์ กรรณิกา (2549) ได้ศึกษาเปรียบเทียบวิธีการประมาณค่าสูญหายในการวางแผนการทดลองแบบจัตุรัสละติน 3 วิธีคือ การประมาณค่าวิธีกำลังสองน้อยสุด การประมาณค่าวิธีค่าคาดหวังสูงสุด และการประมาณค่าวิธีมัลติเพิล อิมพิวเทชัน การเปรียบเทียบกระทำภายใต้สถานการณ์ของจำนวนวิธีทดลอง สัมประสิทธิ์ความผันแปร จำนวนข้อมูลสูญหาย และหาค่าความคลาดเคลื่อนสัมบูรณ์สูงสุด (Maximum Absolute Error (MAE)) ของค่าพยากรณ์ทั้ง 3 วิธี พบว่า เมื่อเปอร์เซ็นต์ข้อมูลสูญหายและสัมประสิทธิ์ความแปรผันมีค่ามาก วิธีมัลติเพิล อิมพิวเทชัน จะให้ค่าความคลาดเคลื่อนสัมบูรณ์สูงสุดมีค่าต่ำกว่าวิธีค่าคาดหวังสูงสุด และวิธีกำลังสองน้อยสุด แต่สำหรับกรณีที่เปอร์เซ็นต์ข้อมูลสูญหายและสัมประสิทธิ์ความแปรผันมีค่าน้อยพบว่าค่าความคลาดเคลื่อนสัมบูรณ์สูงสุดของทั้ง 3 วิธี มีค่าใกล้เคียงกันมาก ดังนั้น จึงควรเลือกใช้วิธีกำลังสองน้อยสุดในการประมาณค่าสูญหาย เนื่องจากสะดวกและรวดเร็วกว่า

ศิริญญา ธีระอนันต์ชัย และ ลี่ลี อิงศรีสว่าง (2552) ได้ทำการศึกษาเพื่อหาตัวแบบทางสถิติที่เหมาะสมสำหรับข้อมูลจำนวนการเรียกค่าสินไหมทดแทนการประกันภัยรถยนต์ในกรุงเทพมหานครซึ่งเป็นข้อมูลติดตามระยะยาวเป็นระยะเวลา 5 ปี ด้วยตัวแบบ Generalized Estimating Equation (GEE) เมื่อกำหนดโครงสร้างความสัมพันธ์ของข้อมูลเป็นแบบ First-order Autoregresssive (AR(1)) และ Compound Symmetry (CS) และตัวแบบผสมเชิงเส้นวางนัยทั่วไป (Generalized Linear Mixed Model (GLMM)) เมื่อกำหนดโครงสร้างความแปรปรวนร่วมของข้อมูลเป็นแบบ First-order Autoregresssive (AR(1)) และ Compound Symmetry (CS) ตามลำดับ พบว่าตัวแบบ GEE เมื่อโครงสร้างความสัมพันธ์ของข้อมูลเป็นแบบ AR(1) จะมีความ

เหมาะสมสำหรับข้อมูลมากกว่ารูปแบบ CS ส่วนตัวแบบ GLMM โครงสร้างความแปรปรวนร่วมของข้อมูลเป็นแบบ AR(1) ก็จะมีความเหมาะสม สำหรับข้อมูลมากกว่ารูปแบบ CS เช่นกัน

ในการวิจัยครั้งนี้ ผู้วิจัยต้องการศึกษาวิธีการประมาณค่าสูญหายของตัวแปรตามในข้อมูลระยะยาว ในตัวแบบ Generalized Estimating Equations ภายใต้สถานการณ์ที่ต่างกัน ซึ่งวิธีประมาณค่าสูญหายที่สนใจศึกษาคือ

1) Last observation carried forward (LOCF)
2) Previous Row Mean
3) Multiple Imputation (MI)

1.2 วัตถุประสงค์ของการวิจัย

เพื่อศึกษาและเปรียบเทียบวิธีการประมาณค่าสูญหายของข้อมูลระยะยาวเพื่อการพยากรณ์ สำหรับตัวแบบ Generalized Estimating Equations

1.3 ข้อตกลงเบื้องต้น

ในการวิเคราะห์ครั้งนี้ผู้วิจัยได้กำหนดข้อตกลงเบื้องต้นดังนี้

1. ตัวแปรที่สนใจศึกษาคือ x และ y ภายใต้การวิเคราะห์ของข้อมูลระยะยาวเพื่อการพยากรณ์ ที่ข้อมูลมีการวัดซ้ำและมีความสัมพันธ์กัน ซึ่งตัวแบบที่ใช้ในการวิเคราะห์คือ

ตัวแบบ Generalized Estimating Equations

$$y_{i(t)} = \beta_0 + \beta_1 t + \beta_{2j}\sum_{j=1}^{J} x_{ijt} + [corr] + \varepsilon_{i(t)}$$

$$; \ i = 1,2,...,n \ , \ j = 1,2,...,J \ , \ t = 1,2,...,p$$

เมื่อ

$y_{i(t)}$ คือ ตัวแปรตาม (Outcome Variable) ของหน่วยตัวอย่างที่ i ระยะเวลาที่ t

x_{ijt} คือ ตัวแปรอิสระ (Predictor Variable) ที่ j ของหน่วยตัวอย่างที่ i ระยะเวลาที่ t

$\beta_0, \beta_1, \beta_{2j}, \beta_{3m}$ คือ พารามิเตอร์ที่ไม่ทราบค่า (Unknown Parameter)

[corr] คือ โครงสร้างสหสัมพันธ์ของ $y_{i(t)}$

$\varepsilon_{i(t)}$ คือ ความคลาดเคลื่อนของหน่วยตัวอย่างที่ i ระยะเวลาที่ t

n คือ ขนาดตัวอย่าง

J คือ จำนวนตัวแปรอิสระ

p คือ ระยะเวลาที่ทำการเก็บข้อมูลซ้ำ

ซึ่งในงานวิจัยสนใจศึกษาโครงสร้างอัตตสหสัมพันธ์อันดับที่หนึ่ง (AR(1)) จากลักษณะของความคลาดเคลื่อนที่มีโครงสร้างอัตตสหสัมพันธ์อันดับที่หนึ่ง (AR(1)) ดังนี้

$$\varepsilon_{i(t)} = \rho\varepsilon_{i(t-1)} + u_{i(t)}$$

$$; \ i = 1,2,...,n \quad , \quad t = 1,2,..., p$$

เมื่อ ρ คือ สัมประสิทธิ์สหสัมพันธ์ระหว่าง $\varepsilon_{i(t)}$ กับ $\varepsilon_{i(t-1)}$ โดยที่ $|\rho| < 1$ และข้อตกลงเบื้องต้นของ $u_{i(t)}$ คือ

$$E(u_{i(t)}) = 0 \quad , \quad Var(u_{i(t)}) = \sigma_u^2$$

ดังนั้นจะได้ว่า

$$E(\varepsilon_{i(t)}) = 0$$

$$Var(\varepsilon_{i(t)}) = \sigma_\varepsilon^2 = \frac{\sigma_u^2}{1-\rho^2}$$

$$Cov(\varepsilon_{i(t)}, \varepsilon_{i(t-r)}) = \rho^r \sigma_\varepsilon^2 \qquad ; \quad r > 0$$

2. กำหนดตัวแปรอิสระไม่มีความสัมพันธ์กัน นั่นคือ กำหนดให้ค่าสหสัมพันธ์ (correlation) มีค่าเท่ากับ 0
3. ลักษณะของตัวแปรตามที่สนใจคือตัวแปรตามต่อเนื่อง (continuous outcome variable)
4. ข้อมูลมีการวัดซ้ำด้วยระยะห่างของเวลาเท่ากัน
5. กำหนดการสูญหายเกิดขึ้นในตัวแปรตาม เป็นการสุ่มหายแบบสุ่ม (missing at random) ที่เกิดขึ้นอย่างน้อยที่สุดที่ $t = 2$
6. ภายในหน่วยตัวอย่างเดียวกันข้อมูลมีความสัมพันธ์กัน แต่ระหว่างหน่วยตัวอย่างข้อมูลเป็นอิสระกัน

1.4 ขอบเขตของการวิจัย

ในการวิจัยครั้งนี้กระทำภายใต้ขอบเขตดังนี้

1. ลักษณะของความความเคลื่อนมีการแจกแจงปกติที่
   - $u_{i(t)}$ เป็นค่าความคลาดเคลื่อนในตัวแบบอัตตสหสัมพันธ์อันดับที่หนึ่ง ซึ่งมีค่าเฉลี่ยเป็น 0 และความแปรปรวนมีค่าคงที่เท่ากับ σ_u^2
   - $\varepsilon_{i(t)}$ เป็นค่าความคลาดเคลื่อนในตัวแบบของข้อมูลระยะยาวที่มีค่าเฉลี่ยเป็น 0 และมีเมทริซ์ความแปรปรวนร่วมเท่ากับ $\sum$
2. จำนวนตัวแปรอิสระที่ศึกษาเท่ากับ 3
3. ขนาดตัวอย่างที่ศึกษาเท่ากับ 60 และ 90
4. ศึกษาเมื่อระยะเวลาที่ทำการเก็บข้อมูลซ้ำเท่ากับ 3 และ 5 คาบเวลา
5. ข้อมูลตัวแปรตามที่สูญหายคิดเป็นร้อยละ 10 และ 15
6. ศึกษาเมื่อค่าสัมประสิทธิ์สหสัมพันธ์ (ρ) ของ $y_{i(t)}$ เท่ากับ 0.2 , 0.5 และ 0.9
7. กำหนดการประมวลผลในแต่ละสถานการณ์จนกว่า $\left|MAPE_K - MAPE_{K+1}\right|$<0.001

1.5 เกณฑ์ที่ใช้ในการตัดสินใจ

ในการวิจัยครั้งนี้จะพิจารณาค่าความคลาดเคลื่อนระหว่างค่าพยากรณ์ของตัวแปรตามกับค่าจริง ในรูปแบบ Mean absolute percentage error (MAPE) ซึ่งมีสูตรในการคำนวณ ดังนี้

$$APE = \frac{\sum_{t=1}^{p}\sum_{i=1}^{n}\left|\frac{(y_{i(t)} - \hat{y}_{i(t)})}{y_{i(t)}}\right|}{np} \times 100$$

$$MAPE = \sum_{k=1}^{K}\frac{APE}{k}$$

$$i = 1,2,...,n \ , \ t = 1,2,...,p \ , \ k = 1,2,...,K$$

เมื่อ

$y_{i(t)}$ คือ ค่าจริงของข้อมูลตัวแปรตามของหน่วยตัวอย่างที่ i คาบเวลาที่ t

$\hat{y}_{i(t)}$ คือ ค่าประมาณของข้อมูลตัวแปรตามของหน่วยตัวอย่างที่ i คาบเวลาที่ t

n คือ ขนาดตัวอย่าง

p คือ ระยะเวลาที่ทำการเก็บข้อมูลซ้ำ

K คือ จำนวนรอบ

การตัดสินใจว่าวิธีการประมาณค่าสูญหายใดที่ดีกว่า ดูได้จากการเปรียบเทียบค่าความคลาดเคลื่อนระหว่างค่าพยากรณ์ของตัวแปรตามกับค่าจริง ด้วย Mean absolute percentage error (MAPE) ว่าวิธีใดให้ค่า MAPE ต่ำกว่า แสดงว่าเป็นวิธีการประมาณที่ดีกว่า

1.6 วิธีดำเนินการวิจัย

1. สร้างข้อมูลตัวแปรอิสระและสร้างข้อมูลความคลาดเคลื่อนให้เกิดอัตตสหสัมพันธ์แบบ AR(1)
2. สร้างข้อมูลตัวแปรตามในตัวแบบ Generalized Estimating Equations
3. สร้างข้อมูลตัวแปรตามให้เกิดการสูญหายโดยเกิดขึ้นอย่างสุ่ม ภายใต้สถานการณ์ที่กำหนด
4. ประมาณค่าข้อมูลด้วยวิธีประมาณค่าสูญหายทั้ง 3 วิธี เพื่อแทนที่ข้อมูลที่สูญหายในตัวแปรตาม
5. ประมาณค่าสัมประสิทธิ์การถดถอยในตัวแบบ Generalized Estimating Equations
6. คำนวณหาค่าความคลาดเคลื่อนระหว่างค่าพยากรณ์ของตัวแปรตามกับค่าจริง (MAPE) ของแต่ละวิธีการประมาณค่าสูญหาย
7. เปรียบเทียบค่าความคลาดเคลื่อนระหว่างค่าพยากรณ์ของตัวแปรตามกับค่าจริงของแต่ละวิธี
8. สรุปผลการวิจัยในแต่ละสถานการณ์

1.7 ประโยชน์ที่คาดว่าจะได้รับ

ประโยชน์ที่คาดว่าจะได้รับในงานวิจัยครั้งนี้ คือ

1. เพื่อเป็นแนวทางในการตัดสินใจเลือกวิธีประมาณค่าสูญหายของตัวแปรตามเมื่อข้อมูลที่นำมาวิเคราะห์เป็นข้อมูลระยะยาว
2. เป็นแนวทางในการศึกษา วิธีการประมาณค่าสูญหายในการวิเคราะห์ข้อมูลในตัวแบบรูปอื่น ๆ ต่อไป

บทที่ 2

แนวคิดและทฤษฎี

ในงานวิจัยครั้งนี้ได้ทำการศึกษาวิธีประมาณค่าสูญหายทั้งหมด 3 วิธี เพื่อพิจารณาเปรียบเทียบหาวิธีการประมาณค่าสูญหายที่เหมาะสมที่สุดสำหรับข้อมูลระยะยาว ซึ่งวิธีที่เลือกมาทำการศึกษาเปรียบเทียบในงานวิจัยได้แก่ วิธี Last Observation Carried Forward วิธี Previous Row Mean และวิธี Multiple Imputation เมื่อเกิดปัญหาข้อมูลสูญหายในตัวแปรตามสำหรับข้อมูลระยะยาวซึ่งมีความสัมพันธ์กันตามเวลาแบบอัตตสหสัมพันธ์อันดับที่หนึ่ง (AR(1)) ด้วยตัวแบบ Generalized Estimating Equations ซึ่งมีรายละเอียดดังต่อไปนี้

2.1 ตัวแบบที่ศึกษา

การวิจัยระยะยาวนั้นต้องเลือกใช้วิธีการทางสถิติที่มีลักษณะเฉพาะ เนื่องจากข้อมูลที่เก็บจากหน่วยตัวอย่างเดียวกันมีแนวโน้มที่จะเกิดสหสัมพันธ์ในตัวเอง การอนุมานทางสถิติจึงจำเป็นต้องคำนึงถึงสหสัมพันธ์ที่เกิดขึ้นนี้ด้วย ซึ่งวิธีการวิเคราะห์ Generalized Estimating Equations นี้ ได้นำโครงสร้างความสัมพันธ์ของข้อมูลที่เก็บซ้ำของหน่วยตัวอย่างเดียวกัน (Working Correlation Matrix) เข้ามาพิจารณาด้วย ดังนี้

$$y_{i(t)} = \beta_0 + \beta_1 t + \beta_{2j} \sum_{j=1}^{J} x_{ijt} + [corr] + \varepsilon_{i(t)}$$

$$; \ i = 1,2,...,n \ , \ j = 1,2,...,J \ , \ t = 1,2,...,p$$

เมื่อ

$y_{i(t)}$ คือ ตัวแปรตาม (Outcome Variable) ของหน่วยตัวอย่างที่ i ระยะเวลาที่ t

x_{ijt} คือ ตัวแปรอิสระ (Predictor Variable) ที่ j ของหน่วยตัวอย่างที่ i ระยะเวลาที่ t

$\beta_0, \beta_1, \beta_{2j}, \beta_{3m}$ คือ พารามิเตอร์ที่ไม่ทราบค่า (Unknown Parameter)

[corr] คือ โครงสร้างสหสัมพันธ์ของ $y_{i(t)}$

$\varepsilon_{i(t)}$ คือ ความคลาดเคลื่อนของหน่วยตัวอย่างที่ i ระยะเวลาที่ t

n คือ ขนาดตัวอย่าง

J คือ จำนวนตัวแปรอิสระ

p คือ ระยะเวลาที่ทำการเก็บข้อมูลซ้ำ

2.2 รูปแบบของอัตตสหสัมพันธ์

โครงสร้างสหสัมพันธ์ เป็นสหสัมพันธ์ของความคลาดเคลื่อนภายในหน่วยตัวอย่างเดียวกันกับตำแหน่งของกาลเวลา ซึ่งอาจจะมีโครงสร้าง (structure) ได้หลายแบบ แต่ในงานวิจัยครั้งนี้สนใจศึกษาโครงสร้างสหสัมพันธ์แบบอัตตสหสัมพันธ์อันดับที่หนึ่ง (AR(1))

First-order Autoregresssive (AR(1)) เป็นความสัมพันธ์ของค่าสังเกตภายในหน่วยตัวอย่างเดียวกัน จะมีค่าลดลงตามระยะห่างของช่วงเวลาที่เก็บข้อมูลซ้ำ ซึ่งสัมประสิทธิ์สหสัมพันธ์จะอยู่ในรูปของ $Cov(y_{i(t)}, y_{i(t-r)}) = \rho^{|r|} \quad ; \quad t \neq r \, , t = 1,2,..., p$ โดยที่สามารถเขียนในรูปเมทริกซ์ คือ

$$\begin{pmatrix} 1 & \rho & \rho^2 & \cdots & \rho^{p-1} \\ \rho & 1 & \rho & \cdots & \rho^{p-2} \\ \rho^2 & \rho & 1 & \cdots & \rho^{p-3} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \rho^{p-1} & \rho^{p-2} & \rho^{p-3} & \cdots & 1 \end{pmatrix}$$

ซึ่งในงานวิจัยสนใจศึกษาโครงสร้างอัตตสหสัมพันธ์อันดับที่หนึ่ง (AR(1)) จากลักษณะของความคลาดเคลื่อนที่มีโครงสร้างอัตตสหสัมพันธ์อันดับที่หนึ่ง (AR(1)) ดังนี้

$$\varepsilon_{i(t)} = \rho \varepsilon_{i(t-1)} + u_{i(t)}$$

$$; \; i = 1,2,..., n \quad , \quad t = 1,2,..., p$$

เมื่อ ρ คือ สัมประสิทธิ์สหสัมพันธ์ระหว่าง $\varepsilon_{i(t)}$ กับ $\varepsilon_{i(t-1)}$ โดยที่ $|\rho| < 1$

$u_{i(t)}$ คือ ค่าความคลาดเคลื่อนในตัวแบบอัตตสหสัมพันธ์อันดับที่หนึ่ง ซึ่งมีค่าเฉลี่ยเป็น 0 และความแปรปรวนมีค่าคงที่เท่ากับ และไม่เกิดอัตตสหสัมพันธ์ต่อกันหรือความแปรปรวนร่วมเป็น 0

เพราะว่า $u_{i(t)} \sim N(0, \sigma_u^2)$

$$Cov(u_{i(t)}, u_{i(t-r)}) = 0 \quad ; \quad t \neq r \, , t = 1,2,..., p$$

ดังนั้นจะมี

$$E(\varepsilon_{i(t)}) = 0$$

$$Var(\varepsilon_{i(t)}) = \sigma_\varepsilon^2 = \frac{\sigma_u^2}{1 - \rho^2}$$

$$Cov(\varepsilon_{i(t)}, \varepsilon_{i(t-r)}) = \rho^r \sigma_\varepsilon^2 \qquad ; \quad r > 0$$

จะทำให้เมทริกซ์ความแปรปรวนร่วม (Covariance Matrix) ของคลาดคลาดเคลื่อน คือ

$$\Sigma = \begin{bmatrix} \Sigma_1 & 0 & 0 & \dots & 0 \\ 0 & \Sigma_2 & 0 & \dots & 0 \\ 0 & 0 & \Sigma_3 & \dots & 0 \\ 0 & 0 & 0 & \dots & \Sigma_n \end{bmatrix}$$

ดังนั้น

$$\Sigma_i = \frac{\sigma_u^2}{1-\rho^2} \begin{bmatrix} 1 & \rho & \rho^2 & \rho^3 & \dots & \rho^{p-1} \\ \rho & 1 & \rho & \rho^2 & \dots & \rho^{p-2} \\ \rho^2 & \rho & 1 & \rho & \dots & \rho^{p-3} \\ \vdots & & \vdots & & \ddots & \\ \rho^{p-1} & \rho^{p-2} & \rho^{p-3} & \rho^{p-4} & \dots & 1 \end{bmatrix} = \sigma_\varepsilon^2 V_i \quad ; i = 1,2,...,n$$

$$\therefore \ \Sigma = \sigma_\varepsilon^2 \begin{bmatrix} V_1 & 0 & 0 & \dots & 0 \\ 0 & V_2 & 0 & \dots & 0 \\ 0 & 0 & V_3 & \dots & 0 \\ 0 & 0 & 0 & \dots & V_n \end{bmatrix} = \sigma_\varepsilon^2 V$$

2.3 วิธีประมาณค่าพารามิเตอร์

การวิเคราะห์ด้วย Generalized Estimating Equations ประมาณค่าสัมประสิทธิ์การถดถอยและค่าความคลาดเคลื่อนมาตรฐานด้วยวิธี Quasi – Likelihood ถูกพัฒนาจาก Weddderburn (1974) เป็นวิธีที่มีคุณสมบัติเหมือนกับวิธี Maximum Likelihood ซึ่งมีรายละเอียดดังนี้

โครงสร้างของ Quasi - Likelihood

พิจารณาส่วนประกอบเดียวของตัวแปรตอบสนอง Y โดยไม่มี subscript ภายใต้เงื่อนไขดังนี้

$$U = u(\mu|;Y) = \frac{Y-\mu}{\sigma^2 V(\mu)}$$

โดยที่ Quasi – Likelihood Function หาได้จาก $Q(\mu; y) = \int_y^\mu u(y|y)dt = \int_y^\mu \frac{y-t}{\sigma^2 V(t)} dt$

Quasi – likelihood สำหรับข้อมูลที่สมบูรณ์เป็นผลรวมของแต่ละ $Q_i(\mu_i; y_i)$ คือ

$$Q(\mu; y) = \sum Q_i(\mu_i; y_i)$$

ให้ ฟังก์ชัน quasi – deviance สำหรับค่าสังเกตเดียว คือ

$$D(\mu; y) = 2\sigma^2 Q(\mu; y) = 2\int_y^{\mu} \frac{y-t}{V(t)} dt$$

ที่เป็นอย่างถูกต้องที่เชื่อถือได้ ยกเว้น $y = \mu$ ผลรวมของ deviance $D(\mu; y)$ หามาได้จากผลรวมของแต่ละส่วนประกอบคือ คำนวณจากฟังก์ชันที่ขึ้นอยู่กับ y และ μ แต่ไม่ขึ้นกับ σ^2

การประมาณพารามิเตอร์

Quasi - Likelihood Estimating Equation สำหรับพารามิเตอร์ (β) สามารถหาได้จาก $Q(\mu; y)$ อาจจะเขียนอยู่ในรูปของ $U(\beta) = 0$ ซึ่งจะได้

$$U(\beta) = \sum D^T V^{-1}(Y - \mu) = 0$$

ซึ่งเรียกว่า quasi – score function โดยที่

$$D_i = \frac{\partial \mu_i}{\partial \beta_t}$$

$$V_i = (A_i^{1/2} R_i A_i^{1/2})\phi$$

$$\phi = \frac{1}{n-p} \sum_i \sum_j \frac{y_{it} - \mu_{it}}{\sqrt{\text{var}(\mu_{it})}}$$

เมื่อ V_i คือ เมทริกซ์โครงสร้างความแปรปรวนร่วมของ Y_i

A_i คือ diagonal matrix ความแปรปรวนของ Y_i โดยเขียนให้อยู่ในรูปเมทริกซ์

R_i คือ เมทริกซ์โครงสร้างความสัมพันธ์ของ Y_i

ϕ คือ overdispersion parameter

ซึ่ง $A_i = diag\{v(\mu_{i(t)})\}$ เป็นเมทริกซ์ทแยงมุม ที่ประกอบด้วยฟังก์ชั่นความแปรปรวน $v(\mu_{i(t)})$ ในลำดับทแยงมุมที่ t

$$A_i = \begin{bmatrix} v(\mu_{1t}) & 0 & \dots & \dots & 0 \\ 0 & v(\mu_{2t}) & 0 & \dots & 0 \\ \dots & 0 & \dots & \dots & 0 \\ \dots & \dots & \dots & \dots & 0 \\ 0 & 0 & 0 & \dots & v(\mu_{nt}) \end{bmatrix}$$

R_i เป็นเมทริกซ์ของความสัมพันธ์ $y_{i(t)}$ หรือที่เรียกว่า Working Correlation Matrix กรณี R_i มีโครงสร้างสหสัมพันธ์แบบ AR(1)

$$R_i = \begin{bmatrix} 1 & \rho & \rho^2 & \dots & \rho^{p-1} \\ \rho & 1 & \rho & \dots & \rho^{p-2} \\ \rho^2 & \rho & 1 & \dots & \rho^{p-3} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \rho^{p-1} & \rho^{p-2} & \rho^{p-3} & \dots & 1 \end{bmatrix}$$

เนื่องจากในงานวิจัยนี้สนใจศึกษาอัตตสัมพันธ์แบบ AR(1)

ϕ คือ overdispersion parameter = 1

2.4 **วิธี** Last Observation Carried Forward

วิธีนี้เป็นวิธีที่ง่ายที่สุดในการแก้ปัญหาข้อมูลสูญหายของข้อมูลระยะยาว ทำได้โดยแทนค่าสูญหายด้วยค่าก่อนหน้าที่ไม่ได้เกิดข้อมูลสูญหายในหน่วยตัวอย่างเดียวกัน โดยที่

$$\hat{y}_{i(t)} = y_{i(t-1)} \quad ; \quad i = 1,2,...,n \quad , \quad t = 1,2,...,p$$

2.5 **วิธี** Previous Row Mean

วิธีนี้หาค่าเฉลี่ยของข้อมูลค่าสังเกตก่อนหน้าในหน่วยตัวอย่างเดียวกัน โดยแทนค่าสูญหายด้วยค่าเฉลี่ยนี้ แล้วนำมาวิเคราะห์ในตัวแบบ Generalized Estimating Equations ต่อไปโดย

$$\hat{y}_{i(t)} = \bar{y}_{i(t^*)}$$

$$; \quad i = 1,2,...,n \quad , \quad t = 1,2,...,p$$

และ t^* คือ ระยะเวลาก่อนหน้าทั้งหมด

2.6 วิธี Multiple Imputation

วิธีนี้จะประมาณค่าสูญหายคล้าย ๆ กับวิธีอื่น ๆ ใน Single Imputation แต่ดำเนินการมากกว่า 1 ครั้ง ซึ่งข้อมูลที่ถูกประมาณค่าขึ้นมาใหม่จะสะท้อนให้เห็นถึงการกระจายสุ่ม (Sampling Variability) ของค่าจริงของ Y โดยข้อมูลที่สูญหายจะถูกแทนที่ด้วยชุดของค่าที่เป็นไปได้มากกว่า 1 (m>1) เพื่อที่จะสร้างชุดข้อมูลที่สมบูรณ์ m ชุด ซึ่งจำนวน m ที่เหมาะสมที่จะได้ชุดข้อมูลที่ดีคือ m ตั้งแต่ 3 ถึง 5 (Sandip Sinharay, Hal S. Stern and Daniel Russell ; 2001) จากนั้นทำการวิเคราะห์ข้อมูลจากชุดต่าง ๆ แล้วบันทึกผลการวิเคราะห์ที่ได้ โดยผลการวิเคราะห์ที่ได้เหล่านี้จะถูกรวมเข้าด้วยกันเพื่อนำค่าที่ได้ไปใช้เป็นค่าประมาณที่สูญหายต่อไป

แนวคิดพื้นฐานในการสร้างชุดข้อมูล m ชุดนั้นคือ การหา Predictive distribution สำหรับค่าสังเกตที่สูญหายจากค่าสังเกตที่มีอยู่ โดยให้ $Y=(Y_{obs},Y_{mis})$ เมื่อ Y_{obs} แทนค่าสังเกตที่มีอยู่ และ Y_{mis} แทนค่าสังเกตที่สูญหาย

สมมติให้ Y มีการแจกแจง $p(Y|\theta)$ เมื่อ θ เป็นพารามิเตอร์ทั้งหมดของตัวแบบ Predictive distribution หาได้จาก

$$\begin{aligned} p(Y_{mis}|Y_{obs}) &= \int p(Y_{mis},\theta|Y_{obs})d\theta \\ &= \int p(Y_{mis}|Y_{obs},\theta)p(\theta|Y_{obs})d\theta \end{aligned}$$

จะเห็นว่าเราสามารถประมาณค่าสูญหาย จากการจำลองค่าพารามิเตอร์จากค่าสังเกตที่มีอยู่ ซึ่งมีการแจกแจงโดยประสบการณ์ $p(\theta|Y_{obs})$ แล้วจำลองค่าสูญหายจากการแจกแจงโดยประสบการณ์ที่มีเงื่อนไข $p(Y_{mis}|Y_{obs},\theta)$

ดังนั้น การประมาณค่าสูญหายโดยวิธี MI ในขั้นตอนของการสร้างข้อมูล m ชุดนั้น ทำได้โดย

กำหนดค่าเริ่มต้นสำหรับการประมาณค่าสูญหายด้วยวิธีภาวะน่าจะเป็นสูงสุด (Maximim Likelihood : ML)

Data augmentation (DA) algorithm แบ่งออกเป็น 2 ขั้นตอนคือ

1. Imputation Step (I step) : $Y_{mis}^{(r)} \sim p(Y_{mis}|Y_{obs},\theta^{(r-1)})$
2. Posterior Step (P step) : $\theta^{(r)} \sim p(\theta|Y_{obs},Y_{mis}^{(r)})$

โดยที่ $(Y_{mis}^{(r)},\theta^{(r)});r=1,2,...$ ทำวนซ้ำจนค่าลู่เข้าสู่การแจกแจง $p(Y_{mis},\theta|Y_{obs})$ จะได้ค่าประมาณค่าสังเกตที่สูญหายค่าใหม่ จากนั้นทำการวนซ้ำตามขั้นตอนดังกล่าวข้างต้นจนค่าลู่เข้าสู่การแจกแจง $p(Y_{mis},\theta|Y_{obs})$ ไปจนครบ m ครั้ง ก็จะได้ค่าประมาณค่าสูญหายทั้งหมด m ชุด จากนั้นนำไปหาค่าเฉลี่ยของค่าประมาณที่ได้ เพื่อนำค่าที่ได้ ไปใช้เป็นค่าประมาณที่สูญหาย

บทที่ 3

วิธีดำเนินการวิจัย

ในการวิจัยครั้งนี้เป็นการวิจัยเชิงทดลอง เพื่อเปรียบเทียบวิธีการประมาณค่าสูญหายของตัวแปรตามสำหรับข้อมูลระยะยาว ในตัวแบบ Generalized Estimating Equations เมื่อข้อมูลระยะยาวมีโครงสร้างสหสัมพันธ์รูปแบบอัตตสหสัมพันธ์อันดับที่หนึ่ง AR(1) โดยทำการเปรียบเทียบวิธีการประมาณค่าสูญหาย 3 วิธี คือ

1. Last Observation Carried Forward
2. Previous Row Mean
3. Multiple Imputation (MI)

การเปรียบเทียบจะเปรียบเทียบความคลาดเคลื่อนของแต่ละวิธีด้วยค่าความคลาดเคลื่อนระหว่างค่าพยากรณ์ของตัวแปรตามกับค่าจริง ในรูปแบบ Mean absolute percentage error (MAPE) ด้วยวิธีการต่าง ๆ ที่ขนาดตัวอย่าง 2 ระดับ ระยะเวลาในการเก็บข้อมูลซ้ำ 2 ระดับ ค่าอัตตสหสัมพันธ์ 3 ระดับ และร้อยละการสูญหาย 2 ระดับ โดยแต่ละสถานการณ์จะทำจนกว่าค่าที่ใช้ในการเปรียบเทียบของรอบก่อนหน้าและรอบถัดไปห่างกันไม่เกิน 0.001

เทคนิคที่ใช้ในการจำลองข้อมูลครั้งนี้อาศัยเทคนิคการจำลองแบบมอนติคาร์โล (Monte Carlo Simulation Technique) ทำการจำลองในแต่ละสถานการณ์ ดังนั้นในส่วนแรกจะกล่าวถึงวิธีการจำลองโดยใช้เทคนิคการจำลองแบบมอนติคาร์โล และแสดงรายละเอียดของขั้นตอนการวิจัยในส่วนถัดไป

3.1 เทคนิคการจำลองแบบมอนติคาร์โล

เทคนิคที่ใช้ในการจำลองข้อมูลครั้งนี้อาศัยเทคนิคการจำลองแบบมอนติคาร์โล (Monte Carlo Simulation Technique) ซึ่งหลักของการจำลองโดยใช้เทคนิคนี้ จะใช้ตัวเลขสุ่ม (Random Numbers) ในการหาคำตอบของปัญหาที่ต้องการศึกษา ซึ่งขั้นตอนของวิธีการจำลองด้วยเทคนิคมอนติคาร์โลแบ่งออกเป็น 3 ขั้นตอน คือ

1. การสร้างตัวเลขสุ่ม ซึ่งการสร้างตัวเลขสุ่มเป็นสิ่งที่สำคัญมากในเทคนิคนี้ เพราะว่าหลักการของการจำลองแบบมอนติคาร์โล จะใช้ตัวเลขสุ่มมาช่วยในการหาคำตอบของปัญหา โดยตัวเลขสุ่มที่สร้างขึ้นนี้จะมีการแจกแจงสม่ำเสมอในช่วง (0,1) ตัวเลขสุ่มแต่ละตัวจะเป็นอิสระกันและมีช่วงยาวก่อนจะเกิดการสุ่มซ้ำ

2. นำตัวเลขสุ่มที่สร้างขึ้นมาประยุกต์ใช้กับปัญหาที่ต้องการศึกษา ซึ่งขั้นตอนนี้ขึ้นอยู่กับลักษณะของปัญหา บางปัญหาอาจจะไม่ใช้ตัวเลขสุ่มโดยตรง แต่บางปัญหาอาจจะต้องมีขั้นตอนอื่นๆ อีกหลายขั้นตอน โดยที่มีการใช้ตัวเลขสุ่มในบางขั้นตอนเท่านั้น
3. การทดลองกระทำ เมื่อนำตัวเลขสุ่มมาประยุกต์ให้เข้ากับปัญหาที่ต้องการศึกษาได้แล้ว ขั้นต่อไปคือ การทดลองโดยใช้กระบวนการของการสุ่ม (Random Process) มากระทำในลักษณะซ้ำ ๆ กันหลาย ๆ ครั้ง เพื่อหาคำตอบที่ต้องการ

3.2 แผนการดำเนินการวิจัย

ในการวิจัยครั้งนี้ได้ทำการจำลองการทดลองตามสถานการณ์ต่าง ๆ โดยเขียนโปรแกรมคอมพิวเตอร์ด้วยโปรแกรม R เพื่อสร้างข้อมูลให้เป็นไปตามการวิจัยในแต่ละสถานการณ์ที่กำหนดตามขั้นตอนดังต่อไปนี้

3.2.1 การสร้างข้อมูล

ในการวิจัยครั้งนี้ตัวแปรที่สนใจศึกษาคือ x และ y ภายใต้การวิเคราะห์ของข้อมูลระยะยาว ที่ข้อมูลมีการวัดซ้ำและมีความสัมพันธ์กัน ซึ่งตัวแบบที่ใช้ในการวิเคราะห์ คือ

ตัวแบบ Generalized Estimating Equations

$$y_{i(t)} = \beta_0 + \beta_1 t + \beta_{2j} \sum_{j=1}^{J} x_{ijt} + [corr] + \varepsilon_{i(t)}$$

$$; \ i = 1,2,...,n \ , \ j = 1,2,...,J \ , \ t = 1,2,...,p$$

เมื่อ

$y_{i(t)}$ คือ ตัวแปรตาม (Outcome Variable) ของหน่วยตัวอย่างที่ i ระยะเวลาที่ t

x_{ijt} คือ ตัวแปรอิสระ (Predictor Variable) ที่ j ของหน่วยตัวอย่างที่ i ระยะเวลาที่ t

$\beta_0, \beta_1, \beta_{2j}, \beta_{3m}$ คือ พารามิเตอร์ที่ไม่ทราบค่า (Unknown Parameter)

[corr] คือ โครงสร้างสหสัมพันธ์ของ $y_{i(t)}$

$\varepsilon_{i(t)}$ คือ ความคลาดเคลื่อนของหน่วยตัวอย่างที่ i ระยะเวลาที่ t

n คือ ขนาดตัวอย่าง

J คือ จำนวนตัวแปรอิสระ

p คือ ระยะเวลาที่ทำการเก็บข้อมูลซ้ำ

ซึ่งมีขั้นตอนในการสร้างข้อมูล ดังนี้

1) จำลองข้อมูลตัวแปรอิสระ ให้มีการแจกแจงปกติ โดยกำหนดในงานวิจัยครั้งนี้ให้ mean = 20 , var = 9 ซึ่งมีรูปแบบอัตตสหสัมพันธ์อันดับที่หนึ่ง (AR(1)) และค่าอัตตสหสัมพันธ์ 3 ระดับ คือ 0.2, 0.5 และ 0.9

2) สร้างความคลาดเคลื่อนตามรูปแบบอัตตสหสัมพันธ์อันดับที่หนึ่ง (AR(1)) ดังนี้

$$\varepsilon_{i(t)} = \rho\varepsilon_{i(t-1)} + u_{i(t)}$$

$$; \ i = 1,2,...,n \quad , \quad t = 1,2,...,p$$

กำหนดค่าอัตตสหสัมพันธ์ 3 ระดับคือ 0.2, 0.5 และ 0.9 ซึ่ง $u_{i(t)}$ เป็นความคลาดเคลื่อนสุ่มที่มีการแจกแจงปกติที่มีค่าเฉลี่ยเป็น 0 และความแปรปรวน มีค่าคงที่เท่ากับ σ_u^2 โดยกำหนดในงานวิจัยครั้งนี้เท่ากับ 9 จากนั้นจึงสร้าง $\varepsilon_{i(t)}$ ให้มีการแจกแจงปกติที่มีค่าเฉลี่ยเป็น 0 และเมทริกซ์ความแปรปรวนร่วมเท่ากับ $\sum$

3) สร้างข้อมูลตัวแปรตาม จากตัวแปรอิสระและค่าความคลาดเคลื่อนในตัวแบบ Generalized Estimating Equations โดย $\beta_0, \beta_1, \beta_2, \beta_3, \beta_4$ แต่ละตัวถูกกำหนดขึ้นในงานวิจัยครั้งนี้เท่ากับ 0.5

โดยจำลองข้อมูลให้มีสถานการณ์ที่ขนาดตัวอย่าง ระยะเวลาเก็บข้อมูลซ้ำ และร้อยละการสูญหายที่ใช้แตกต่างกัน

3.2.2 สุ่มตำแหน่งการสูญหายของข้อมูล

ทำการสุ่มตำแหน่งที่สูญหายตามร้อยละการสูญหายที่กำหนด โดยใช้การสุ่มตัดข้อมูลออกแบบสุ่ม แต่ไม่มีรูปแบบที่แน่นอนสำหรับข้อมูลระยะยาว และข้อมูลที่ถูกตัดออกไปนั้นจะเก็บไว้เพื่อนำมาเปรียบเทียบกับค่าใหม่ที่ประมาณด้วยวิธีประมาณค่าสูญหายทั้ง 3 วิธี

3.2.3 ประมาณค่าสูญหายด้วยวิธีการทั้ง 3 วิธี

1) Last Observation Carried Forward (LOCF) เป็นวิธีการประมาณค่าสูญหายจากค่าสังเกตก่อนหน้าที่ไม่ได้สูญหายในหน่วยตัวอย่างเดียวกัน

2) Previous Row Mean เป็นวิธีการประมาณค่าสูญหายด้วยค่าเฉลี่ยของข้อมูลค่าสังเกตก่อนหน้าในหน่วยตัวอย่างเดียวกัน

3) Multiple Imputation (MI) เป็นวิธีการประมาณค่าสูญหายด้วยค่าเฉลี่ยของชุดข้อมูล m ชุด ที่สร้างขึ้นมาจาก Predictive Distribution ของค่าสังเกตที่มีอยู่

3.2.4 ประมาณพารามิเตอร์ใหม่

เมื่อแทนค่าข้อมูลสูญหายจากค่าประมาณที่ได้แล้ว จะทำการประมาณค่าสัมประสิทธิ์การถดถอยใหม่ด้วยวิธี Quasi-Likelihood เพื่อหาสมการถดถอยในตัวแบบ Generalized Estimating Equations จากวิธีการประมาณค่าสูญหายทั้ง 3 วิธี

3.2.5 หาค่าเฉลี่ยร้อยละความคลาดเคลื่อนสัมบูรณ์

หาความคลาดเคลื่อนระหว่างค่าพยากรณ์ของตัวแปรตามกับค่าจริง ในรูปแบบ Mean absolute percentage error (MAPE) ซึ่งมีขั้นตอนดังนี้

1) ประมาณค่าพารามิเตอร์ $\hat{y}$ ด้วยพารามิเตอร์ใหม่ที่ประมาณโดยสมการถดถอยจากวิธีการประมาณค่าสูญหายทั้ง 3 วิธี
2) นำมาเปรียบเทียบกับ y ที่สร้างไว้ในตอนต้น โดยใช้ โดยมีสูตรการคำนวณดังนี้

$$APE = \frac{\sum_{t=1}^{p}\sum_{i=1}^{n}\left|\frac{(y_{i(t)} - \hat{y}_{i(t)})}{y_{i(t)}}\right|}{np} \times 100$$

$$i = 1,2,...,n \ , \ t = 1,2,...,p \ , \ k = 1,2,...,K$$

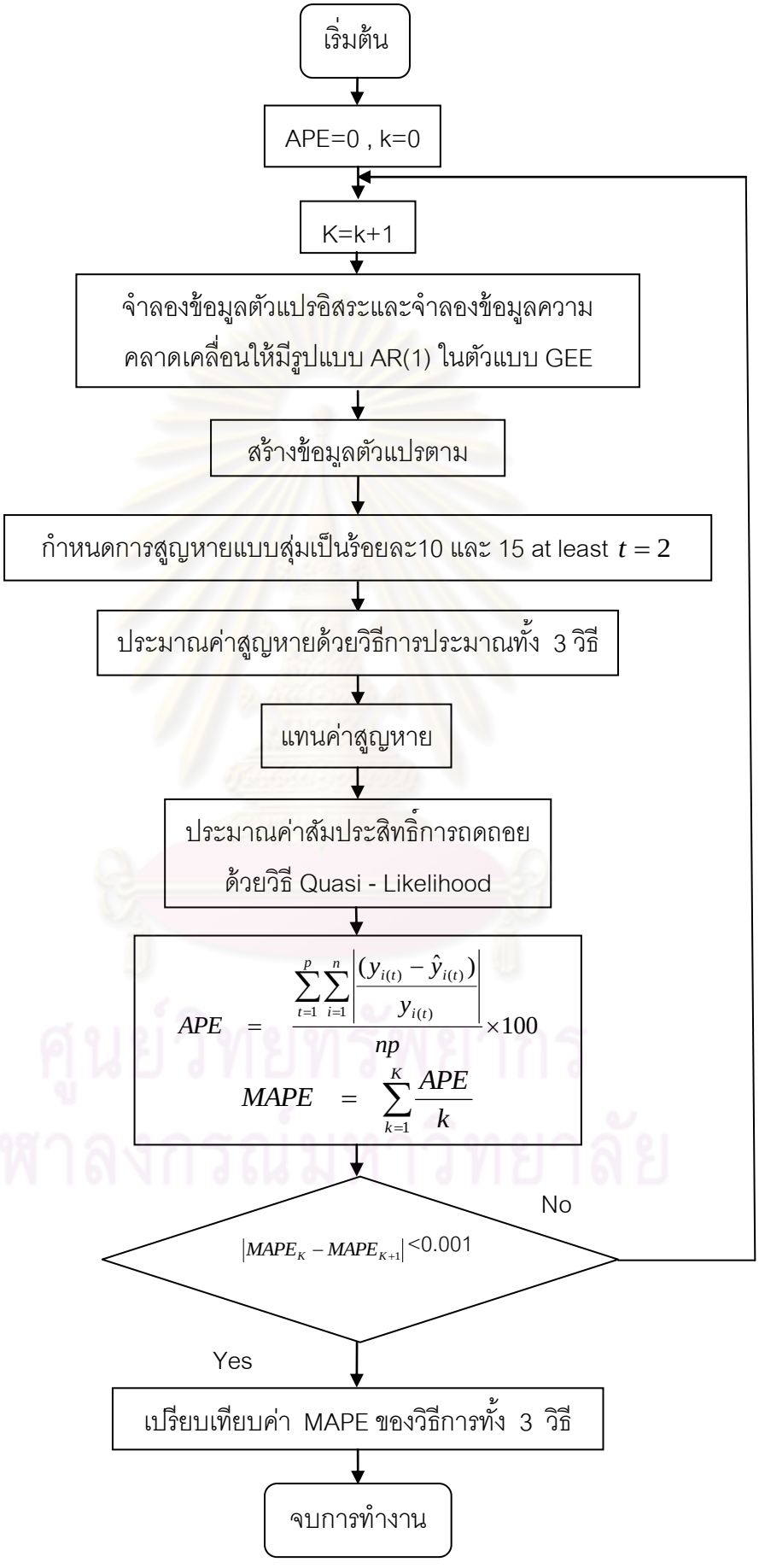
3) หาค่า MAPE เฉลี่ยจากการจำลองข้อมูลตามสถานการณ์ต่าง ๆ

$$MAPE = \sum_{k=1}^{K}\frac{APE}{k}$$

4) เปรียบเทียบค่า MAPE จากการประมาณค่าสูญหายแต่ละวิธี และสรุปผลการเปรียบเทียบ

โดยทำการทดลองจนกว่าค่า MAPE ของรอบก่อนหน้าและรอบถัดไปห่างกันไม่เกิน 0.001 และทำตามขั้นตอนต่างๆ โดยจะเปลี่ยนขนาดตัวอย่าง ระยะเวลาที่ทำการเก็บข้อมูลซ้ำ อัตตสหสัมพันธ์ และร้อยละการสูญหาย จนครบทุกสถานการณ์ จึงนำค่า Mean absolute percentage error (MAPE) ของวิธีประมาณค่าสูญหายทั้ง 3 วิธีมาเปรียบเทียบกัน และสรุปผลการเปรียบเทียบ

3.3 ขั้นตอนการทำงานของโปรแกรม



บทที่ 4

ผลการวิจัย

การวิจัยครั้งนี้มีวัตถุประสงค์เพื่อศึกษาและเปรียบเทียบวิธีการประมาณค่าสูญหายของตัวแปรตามสำหรับข้อมูลระยะยาวในตัวแบบ Generalized Estimating Equations เมื่อข้อมูลระยะยาวมีอัตตสหสัมพันธ์ในตัวเองรูปแบบอัตตสหสัมพันธ์อันดับที่หนึ่ง (First order autoregressive : AR(1)) โดยทำการเปรียบเทียบวิธีการประมาณค่าสูญหาย 3 วิธีคือ วิธี Last Observation Carried Forward วิธี Previous Row Mean และวิธี Multiple Imputation ซึ่งพิจารณาปัจจัยต่าง ๆ คือ ขนาดตัวอย่าง ระยะเวลาในการเก็บข้อมูลซ้ำ ร้อยละการสูญหาย และระดับอัตตสหสัมพันธ์ (First order autoregressive : AR(1)) การวิจัยครั้งนี้จำลองข้อมูลโดยอาศัยเทคนิคการจำลองแบบมอนติคาร์โล (Monte Carlo Simulation Technique) ทำการจำลองในแต่ละสถานการณ์ โดยใช้เกณฑ์เปรียบเทียบความคลาดเคลื่อนของแต่ละวิธีด้วยค่าความคลาดเคลื่อนระหว่างค่าพยากรณ์ของตัวแปรตามกับค่าจริงในรูปแบบ Mean absolute percentage error (MAPE) ซึ่งแต่ละสถานการณ์จะทำจนกว่าค่าที่ใช้ในการเปรียบเทียบของรอบก่อนหน้าและรอบถัดไปห่างกันไม่เกิน 0.001

การเสนอผลการวิจัยในบทนี้ จะนำเสนอผลการวิจัยในรูปแบบตารางและกราฟ เพื่อความสะดวกในการอธิบายจึงใช้สัญลักษณ์ต่อไปนี้เพื่อแทนความหมายต่าง ๆ

n	หมายถึง	ขนาดตัวอย่าง
t	หมายถึง	ระยะเวลาในการเก็บข้อมูลซ้ำ
rho	หมายถึง	ค่าอัตตสหสัมพันธ์
pm	หมายถึง	ร้อยละการสูญหายของข้อมูล
LOCF	หมายถึง	วิธี Last Observation Carried Forward
PRM	หมายถึง	วิธี Previous Row Mean
MI	หมายถึง	วิธี Multiple Imputation
MAPE	หมายถึง	ค่าความคลาดเคลื่อนระหว่างค่าพยากรณ์ของตัวแปรตามกับค่าจริงในรูปแบบ Mean absolute percentage error

4.1 การเปรียบเทียบค่าเฉลี่ยร้อยละความคลาดเคลื่อนสัมบูรณ์ของวิธีการประมาณค่าสูญหายทั้ง 3 วิธี ในแต่ละสถานการณ์

ในการประมาณค่าสูญหายทั้ง 3 วิธี ผู้วิจัยได้ทำการศึกษาที่ขนาดตัวอย่าง 2 ระดับ คือ 60 และ 90 ระยะเวลาในการเก็บข้อมูลซ้ำ 2 ระดับ คือ 3 และ 5 คาบเวลา อัตตสหสัมพันธ์ 3 ระดับ คือ 0.2, 0.5 และ 0.9 และร้อยละการสูญหาย 2 ระดับเป็น 10 และ 15 ตามลำดับ

ซึ่งผลการวิจัยได้นำเสนอในตารางที่ 4.1 – 4.9 ดังต่อไปนี้

4.1.1 แสดงค่าเฉลี่ยร้อยละความคลาดเคลื่อนสัมบูรณ์ เมื่อระดับอัตตสหสัมพันธ์เปลี่ยนแปลง แต่ขนาดตัวอย่าง ระยะเวลาในการเก็บข้อมูลซ้ำ และร้อยละการสูญหายคงที่ นำเสนอในตารางที่ 4.1 - 4.2

4.1.2 แสดงค่าเฉลี่ยร้อยละความคลาดเคลื่อนสัมบูรณ์ เมื่อขนาดตัวอย่างเปลี่ยนแปลง แต่ระยะเวลาในการเก็บข้อมูลซ้ำ ร้อยละการสูญหาย และระดับอัตตสหสัมพันธ์คงที่ นำเสนอในตารางที่ 4.3 - 4.4

4.1.3 แสดงค่าเฉลี่ยร้อยละความคลาดเคลื่อนสัมบูรณ์ เมื่อระยะเวลาในการเก็บข้อมูลซ้ำเปลี่ยนแปลง แต่ขนาดตัวอย่าง ร้อยละการสูญหาย และระดับอัตตสหสัมพันธ์คงที่ นำเสนอในตารางที่ 4.5 – 4.6

4.1.4 แสดงค่าเฉลี่ยร้อยละความคลาดเคลื่อนสัมบูรณ์ เมื่อร้อยละการสูญหายเปลี่ยนแปลง แต่ขนาดตัวอย่าง ระยะเวลาในการเก็บข้อมูลซ้ำ และระดับอัตตสหสัมพันธ์คงที่ นำเสนอในตารางที่ 4.7 – 4.9

4.1.1 แสดงค่าเฉลี่ยร้อยละความคลาดเคลื่อนสัมบูรณ์ เมื่อระดับอัตตสหสัมพันธ์เปลี่ยนแปลง แต่ขนาดตัวอย่าง ระยะเวลาในการเก็บข้อมูลซ้ำ และร้อยละการสูญหายคงที่

ซึ่งผลการวิจัยนำเสนอดังตารางที่ 4.1 - 4.2 ดังนี้

ตารางที่ 4.1 แสดงค่าเฉลี่ยร้อยละความคลาดเคลื่อนสัมบูรณ์ เมื่อขนาดตัวอย่างเท่ากับ 60 ระยะเวลาในการเก็บข้อมูลซ้ำเท่ากับ 3 และ 5 คาบเวลา ร้อยละการสูญหายเท่ากับ 10 และ 15 โดยจำแนกตามอัตตสหสัมพันธ์

n	t	pm	rho	MAPE		
				LOCF	PRM	MI
60	3	10	0.2	0.00087	0.00067*	0.00099
			0.5	0.00288	0.00284*	0.00290
			0.9	0.00315	0.00618	0.00312*
		15	0.2	0.00075	0.00068	0.00062*
			0.5	0.00276*	0.00315	0.00284
			0.9	0.00384	0.00592	0.00331*
	5	10	0.2	0.00052	0.00047*	0.00061
			0.5	0.00203*	0.00272	0.00238
			0.9	0.00306	0.00542	0.00301*
		15	0.2	0.00058	0.00055	0.00049*
			0.5	0.00201*	0.00264	0.00240
			0.9	0.00336	0.00459	0.00298*

หมายเหตุ * หมายถึง วิธีการประมาณค่าสูญหายที่มีค่า MAPE ต่ำสุด

จากตารางที่ 4.1 เมื่อพิจารณาค่า MAPE โดยที่ขนาดตัวอย่างเท่ากับ 60 ระยะเวลาในการเก็บข้อมูลซ้ำเท่ากับ 3 และ 5 คาบเวลา ร้อยละการสูญหายเท่ากับ 10 และ 15 พบว่า เมื่อค่าอัตตสหสัมพันธ์เพิ่มสูงขึ้น ค่าของ MAPE มีแนวโน้มเพิ่มสูงขึ้น

เมื่อพิจารณาเปรียบเทียบค่า MAPE กรณีที่ขนาดตัวอย่างเท่ากับ 60 พบว่า

เมื่อระยะเวลาในการเก็บข้อมูลซ้ำเท่ากับ 3 คาบเวลา ที่ร้อยละการสูญหายเท่ากับ 10 พบว่า เมื่ออัตตสหสัมพันธ์ระดับต่ำ (0.2) และอัตตสหสัมพันธ์ระดับปานกลาง (0.5) วิธี PRM จะให้ค่า MAPE ต่ำที่สุด แต่เมื่ออัตตสหสัมพันธ์ระดับสูง (0.9) วิธี MI จะให้ค่า MAPE ต่ำที่สุด ซึ่งถ้า

ร้อยละการสูญหายเพิ่มสูงขึ้นเท่ากับ 15 พบว่า เมื่ออัตตสหสัมพันธ์ระดับต่ำ (0.2) และอัตตสหสัมพันธ์ระดับสูง (0.9) วิธี MI จะให้ค่า MAPE ต่ำที่สุด แต่เมื่ออัตตสหสัมพันธ์ระดับปานกลาง (0.5) วิธี LOCF จะให้ค่า MAPE ต่ำที่สุด

เมื่อระยะเวลาในการเก็บข้อมูลซ้ำเท่ากับ 5 คาบเวลา ที่ร้อยละการสูญหายเท่ากับ 10 พบว่า เมื่ออัตตสหสัมพันธ์ระดับต่ำ (0.2) วิธี PRM จะให้ค่า MAPE ต่ำที่สุด เมื่ออัตตสหสัมพันธ์ระดับปานกลาง (0.5) วิธี LOCF จะให้ค่า MAPE ต่ำที่สุด และเมื่อและอัตตสหสัมพันธ์ระดับสูง (0.9) วิธี MI จะให้ค่า MAPE ต่ำที่สุด ซึ่งถ้าร้อยละการสูญหายเพิ่มสูงขึ้นเท่ากับ 15 พบว่า เมื่ออัตตสหสัมพันธ์ระดับต่ำ (0.2) และอัตตสหสัมพันธ์ระดับสูง (0.9) วิธี MI จะให้ค่า MAPE ต่ำที่สุด แต่เมื่ออัตตสหสัมพันธ์ระดับปานกลาง (0.5) วิธี LOCF จะให้ค่า MAPE ต่ำที่สุด

ตารางที่ 4.2 แสดงค่าเฉลี่ยร้อยละความคลาดเคลื่อนสัมบูรณ์ เมื่อขนาดตัวอย่างเท่ากับ 90 ระยะเวลาในการเก็บข้อมูลซ้ำเท่ากับ 3 และ 5 คาบเวลา ร้อยละการสูญหายเท่ากับ 10 และ 15 โดยจำแนกตามอัตตสหสัมพันธ์

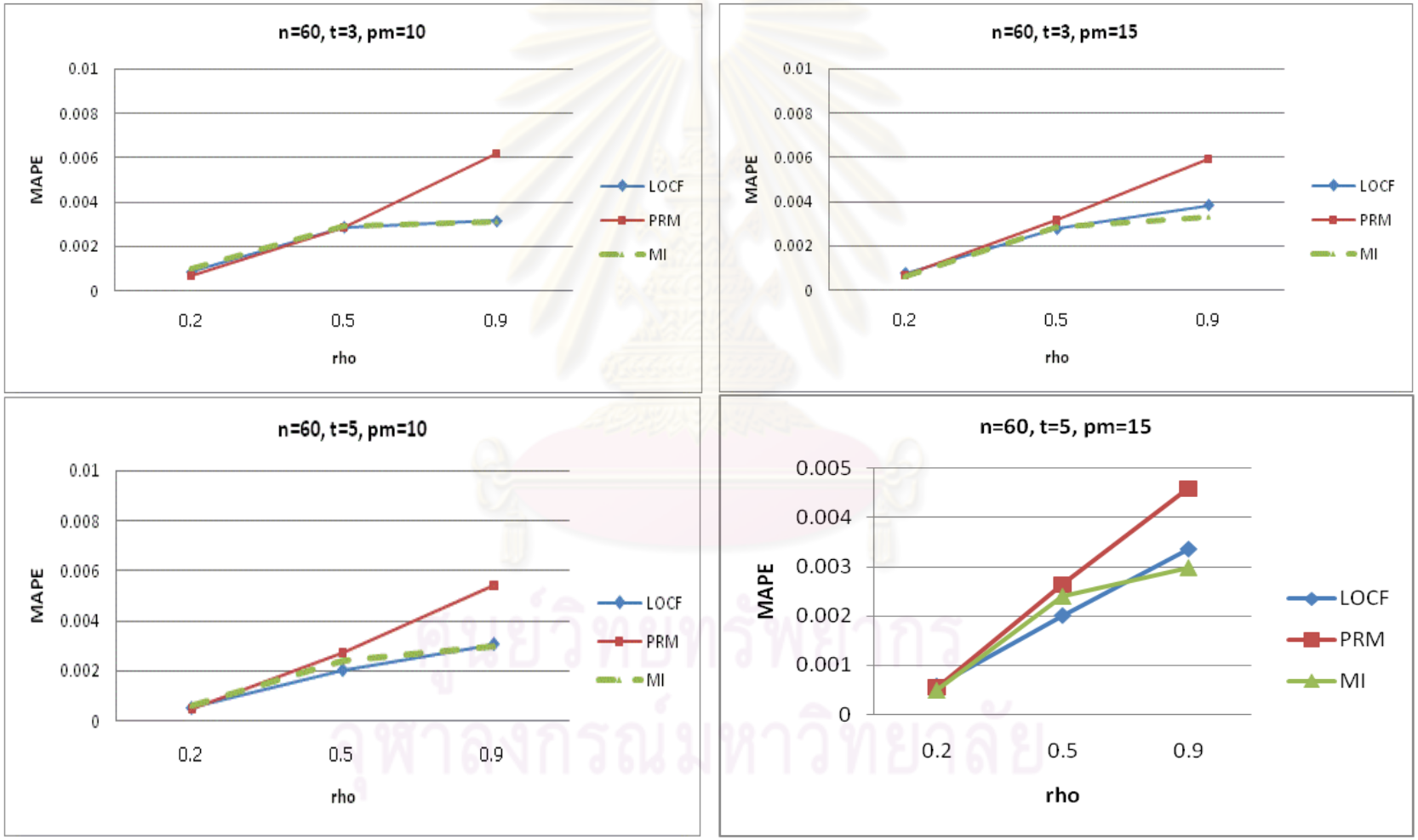
n	t	pm	rho	MAPE		
				LOCF	PRM	MI
90	3	10	0.2	0.00086	0.00067*	0.00098
			0.5	0.00283	0.00281*	0.00284
			0.9	0.00309	0.00607	0.00301*
		15	0.2	0.00075	0.00065	0.00061*
			0.5	0.00275	0.00311	0.00273*
			0.9	0.00379	0.00462	0.00330*
	5	10	0.2	0.00052	0.00045	0.00044*
			0.5	0.00198	0.00236	0.00192*
			0.9	0.00293	0.00490	0.00287*
		15	0.2	0.00057	0.00049	0.00045*
			0.5	0.00199	0.00251	0.00187*
			0.9	0.00318	0.00442	0.00292*

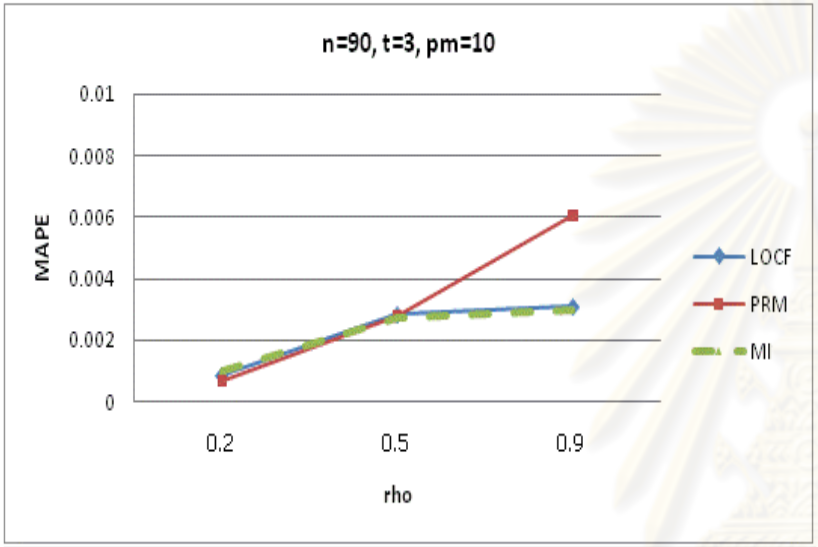
จากตารางที่ 4.2 เมื่อพิจารณาค่า MAPE โดยที่ขนาดตัวอย่างเท่ากับ 90 ระยะเวลาในการเก็บข้อมูลซ้ำเท่ากับ 3 และ 5 คาบเวลา ร้อยละการสูญหายเท่ากับ 10 และ 15 พบว่า เมื่อค่าอัตตสหสัมพันธ์เพิ่มสูงขึ้น ค่าของ MAPE มีแนวโน้มเพิ่มสูงขึ้น

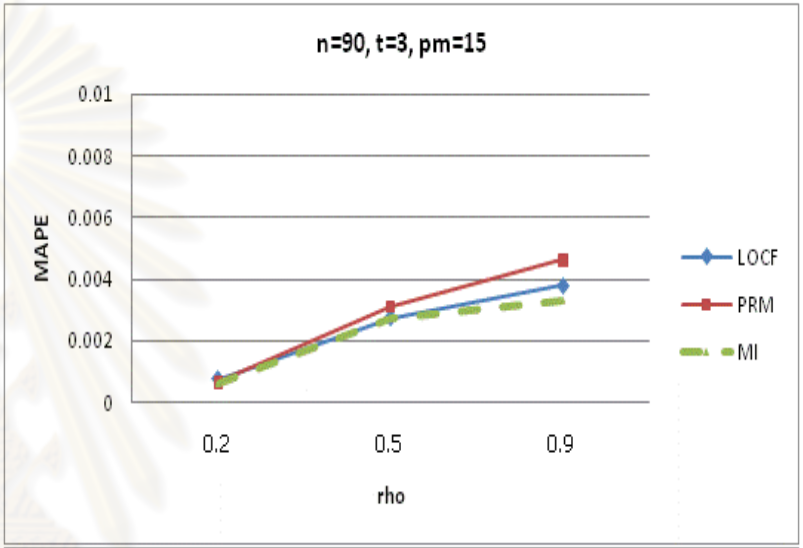
เมื่อพิจารณาเปรียบเทียบค่า MAPE กรณีที่ขนาดตัวอย่างเท่ากับ 90 พบว่า

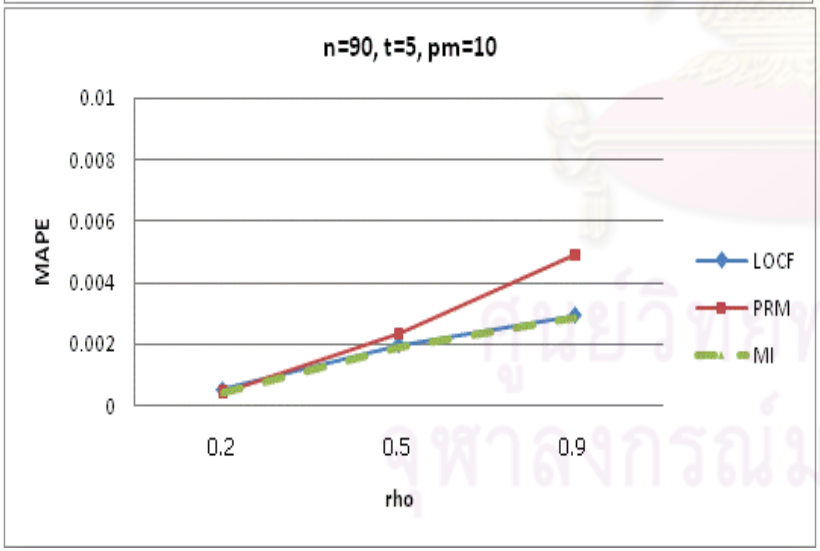
เมื่อระยะเวลาในการเก็บข้อมูลซ้ำเท่ากับ 3 คาบเวลา ที่ร้อยละการสูญหายเท่ากับ 10 พบว่า เมื่ออัตตสหสัมพันธ์ระดับต่ำ (0.2) และอัตตสหสัมพันธ์ระดับปานกลาง (0.5) วิธี PRM จะให้ค่า MAPE ต่ำที่สุด แต่เมื่ออัตตสหสัมพันธ์ระดับสูง (0.9) วิธี MI จะให้ค่า MAPE ต่ำที่สุด ซึ่งถ้าร้อยละการสูญหายเพิ่มสูงขึ้นเท่ากับ 15 พบว่า ทุกระดับอัตตสหสัมพันธ์ทั้งระดับต่ำ ระดับปานกลาง และระดับสูง วิธี MI จะให้ค่า MAPE ต่ำที่สุด

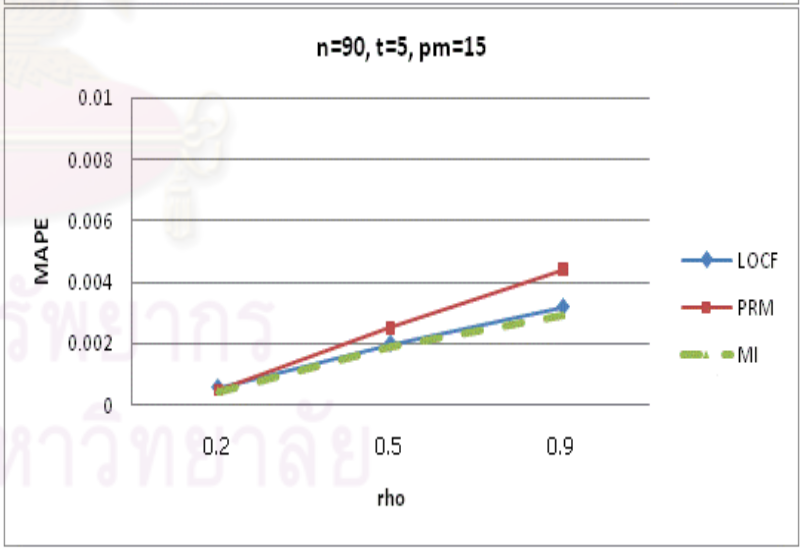
รูปที่ 4.1 แสดงค่าเฉลี่ยร้อยละความคลาดเคลื่อนสัมบูรณ์ เมื่อระดับอัตตสหสัมพันธ์เปลี่ยนแปลง แต่ขนาดตัวอย่าง ระยะเวลาในการเก็บข้อมูลซ้ำ และร้อยละการสูญหายคงที่











4.1.2 แสดงค่าเฉลี่ยร้อยละความคลาดเคลื่อนสัมบูรณ์ เมื่อขนาดตัวอย่างเปลี่ยนแปลง แต่ระยะเวลาในการเก็บข้อมูลซ้ำ ร้อยละการสูญหาย และระดับอัตตสหสัมพันธ์คงที่

ซึ่งผลการวิจัยนำเสนอดังตารางที่ 4.3 - 4.4 ดังนี้

ตารางที่ 4.3 แสดงค่าเฉลี่ยร้อยละความคลาดเคลื่อนสัมบูรณ์ เมื่อระยะเวลาในการเก็บข้อมูลซ้ำเท่ากับ 3 คาบเวลา ระดับอัตตสหสัมพันธ์เท่ากับ 0.2, 0.5 และ 0.9 ร้อยละการสูญหายเท่ากับ 10 และ 15 โดยจำแนกตามขนาดตัวอย่าง

t	rho	pm	n	MAPE		
				LOCF	PRM	MI
3	0.2	10	60	0.00087	0.00067*	0.00099
			90	0.00086	0.00067*	0.00098
		15	60	0.00075	0.00068	0.00062*
			90	0.00075	0.00065	0.00061*
	0.5	10	60	0.00288	0.00284*	0.00290
			90	0.00283	0.00281*	0.00284
		15	60	0.00276*	0.00315	0.00284
			90	0.00275	0.00311	0.00273*
	0.9	10	60	0.00315	0.00618	0.00312*
			90	0.00309	0.00607	0.00301*
		15	60	0.00384	0.00592	0.00331*
			90	0.00379	0.00462	0.00330*

จากตารางที่ 4.3 เมื่อพิจารณาค่า MAPE โดยที่ระยะเวลาในการเก็บข้อมูลซ้ำเท่ากับ 3 คาบเวลา ระดับอัตตสหสัมพันธ์เท่ากับ 0.2, 0.5 และ 0.9 ร้อยละการสูญหายเท่ากับ 10 และ 15 พบว่า เมื่อขนาดตัวอย่างเพิ่มสูงขึ้น ค่าของ MAPE มีแนวโน้มลดลง

เมื่อพิจารณาเปรียบเทียบค่า MAPE กรณีที่ระยะเวลาในการเก็บข้อมูลซ้ำเท่ากับ 3 คาบเวลา พบว่า

เมื่ออัตตสหสัมพันธ์ระดับต่ำ (0.2) ร้อยละการสูญหายเท่ากับ 10 พบว่า วิธี PRM จะให้ค่า MAPE ต่ำที่สุดที่ขนาดตัวอย่างทั้ง 2 ระดับคือ 60 และ 90 แต่เมื่อร้อยละการสูญหายเพิ่มสูงขึ้นเท่ากับ 15 พบว่า วิธี MI จะให้ค่า MAPE ต่ำที่สุดที่ขนาดตัวอย่างทั้ง 2 ระดับคือ 60 และ 90

เมื่ออัตตสหสัมพันธ์ระดับปานกลาง (0.5) ร้อยละการสูญหายเท่ากับ 10 พบว่า วิธี PRM จะให้ค่า MAPE ต่ำที่สุดที่ขนาดตัวอย่างทั้ง 2 ระดับคือ 60 และ 90 แต่เมื่อร้อยละการสูญหายเพิ่มสูงขึ้นเท่ากับ 15 พบว่า วิธี LOCF จะให้ค่า MAPE ต่ำที่สุดที่ขนาดตัวอย่างเท่ากับ 60 และเมื่อขนาดตัวอย่าเท่ากับ 90 วิธี MI จะให้ค่า MAPE ต่ำที่สุด

เมื่ออัตตสหสัมพันธ์ระดับสูง (0.9) ทุกร้อยละการสูญหาย ทุกขนาดตัวอย่าง วิธี MI จะให้ค่า MAPE ต่ำที่สุด

ตารางที่ 4.4 แสดงค่าเฉลี่ยร้อยละความคลาดเคลื่อนสัมบูรณ์ เมื่อระยะเวลาในการเก็บข้อมูลซ้ำเท่ากับ 5 คาบเวลา ระดับอัตตสหสัมพันธ์เท่ากับ 0.2, 0.5 และ 0.9 ร้อยละการสูญหายเท่ากับ 10 และ 15 โดยจำแนกตามขนาดตัวอย่าง

t	rho	pm	n	MAPE		
				LOCF	PRM	MI
5	0.2	10	60	0.00052	0.00047*	0.00061
			90	0.00052	0.00045	0.00044*
		15	60	0.00058	0.00055	0.00049*
			90	0.00057	0.00049	0.00045*
	0.5	10	60	0.00203*	0.00272	0.00238
			90	0.00198	0.00236	0.00192*
		15	60	0.00201*	0.00264	0.00240
			90	0.00199	0.00251	0.00187*
	0.9	10	60	0.00306	0.00542	0.00301*
			90	0.00293	0.00490	0.00287*
		15	60	0.00336	0.00459	0.00298*
			90	0.00318	0.00442	0.00292*

จากตารางที่ 4.4 เมื่อพิจารณาค่า MAPE โดยที่ระยะเวลาในการเก็บข้อมูลซ้ำเท่ากับ 5 คาบเวลา ระดับอัตตสหสัมพันธ์เท่ากับ 0.2, 0.5 และ 0.9 ร้อยละการสูญหายเท่ากับ 10 และ 15 พบว่า เมื่อขนาดตัวอย่างเพิ่มสูงขึ้น ค่าของ MAPE มีแนวโน้มลดลง

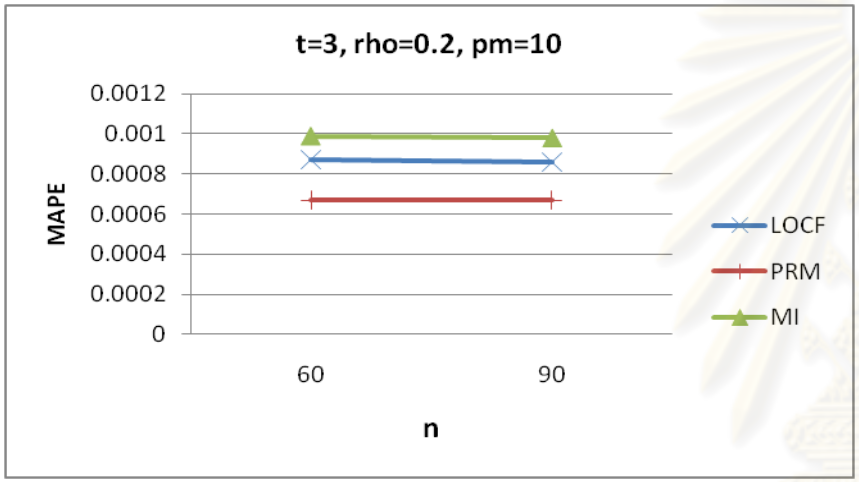
เมื่อพิจารณาเปรียบเทียบค่า MAPE กรณีที่ระยะเวลาในการเก็บข้อมูลซ้ำเท่ากับ 5 คาบเวลา พบว่า

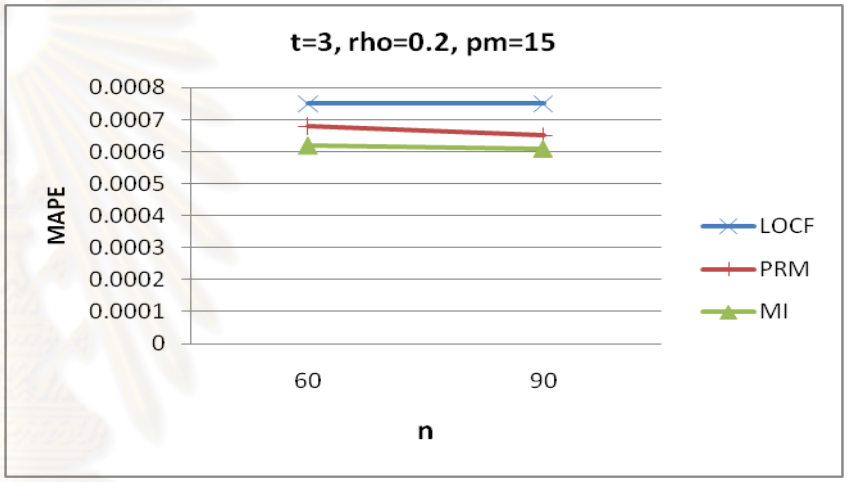
เมื่ออัตตสหสัมพันธ์ระดับต่ำ (0.2) ร้อยละการสูญหายเท่ากับ 10 พบว่า วิธี PRM จะให้ค่า MAPE ต่ำที่สุดที่ขนาดตัวอย่างเท่ากับ 60 และเมื่อขนาดตัวอย่างเท่ากับ 90 วิธี MI จะให้ค่า MAPE ต่ำที่สุด แต่เมื่อร้อยละการสูญหายเพิ่มสูงขึ้นเท่ากับ 15 พบว่า วิธี MI จะให้ค่า MAPE ต่ำที่สุดที่ขนาดตัวอย่างทั้ง 2 ระดับคือ 60 และ 90

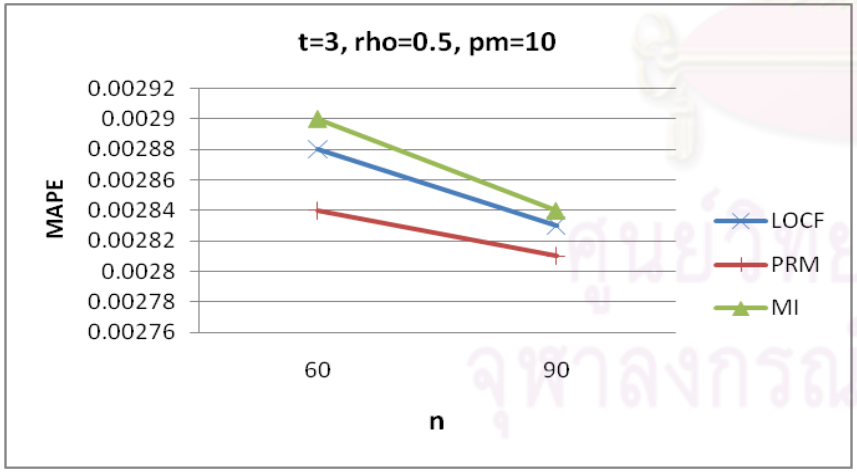
เมื่ออัตตสหสัมพันธ์ระดับปานกลาง (0.5) ที่ร้อยละการสูญหายทั้ง 2 ระดับคือ ร้อยละ 10 และร้อยละ 15 พบว่า วิธี LOCF จะให้ค่า MAPE ต่ำที่สุดที่ขนาดตัวอย่างเท่ากับ 60 และเมื่อขนาดตัวอย่างเท่ากับ 90 วิธี MI จะให้ค่า MAPE ต่ำที่สุด

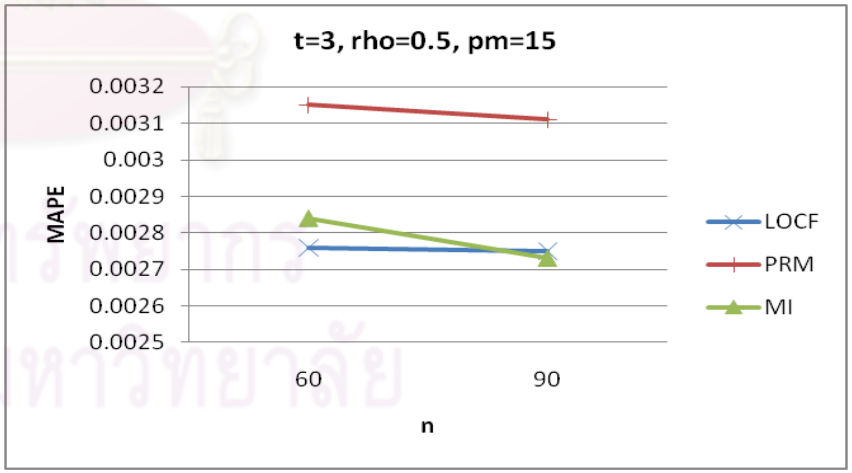
เมื่ออัตตสหสัมพันธ์ระดับสูง (0.9) ทุกร้อยละการสูญหาย ทุกขนาดตัวอย่าง วิธี MI จะให้ค่า MAPE ต่ำที่สุด

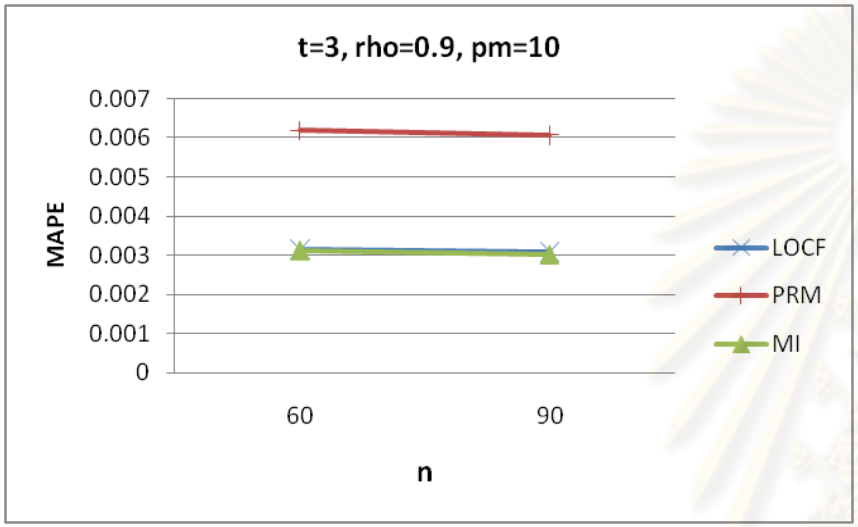
รูปที่ 4.2 แสดงค่าเฉลี่ยร้อยละความคลาดเคลื่อนสัมบูรณ์ เมื่อขนาดตัวอย่างเปลี่ยนแปลง แต่ระยะเวลาในการเก็บข้อมูลซ้ำ ร้อยละการสูญหาย และระดับอัตตสหสัมพันธ์คงที่

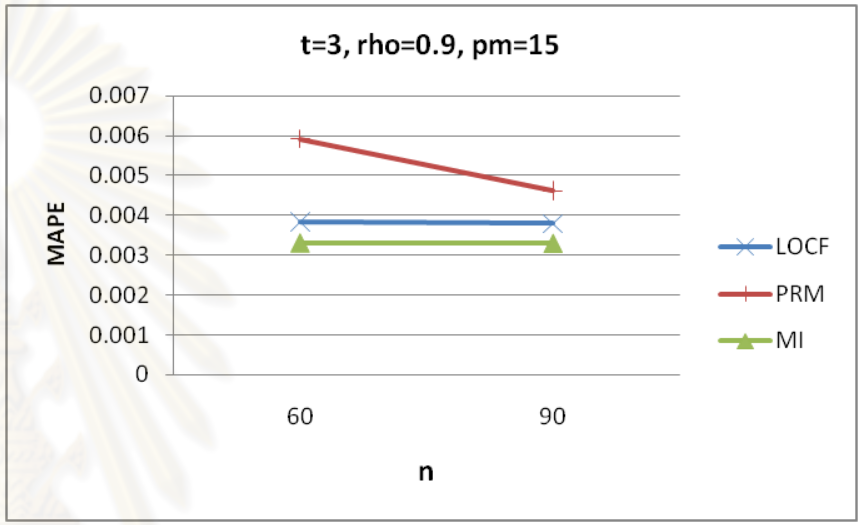


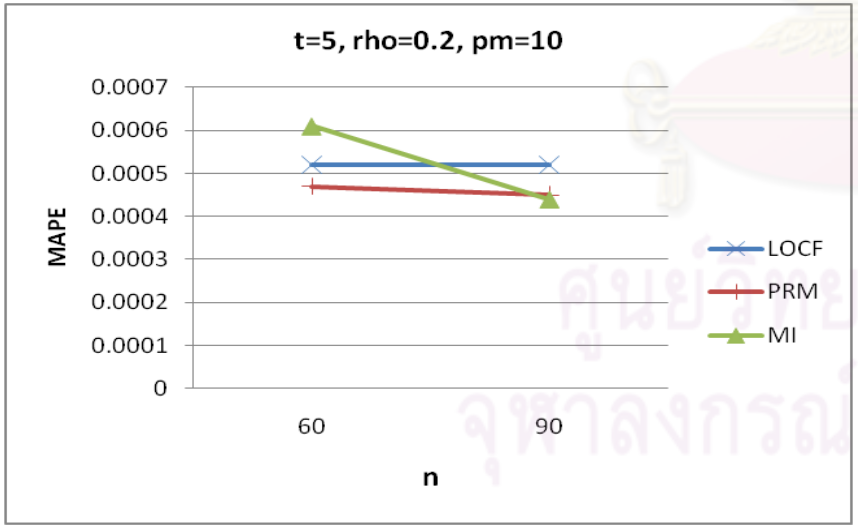


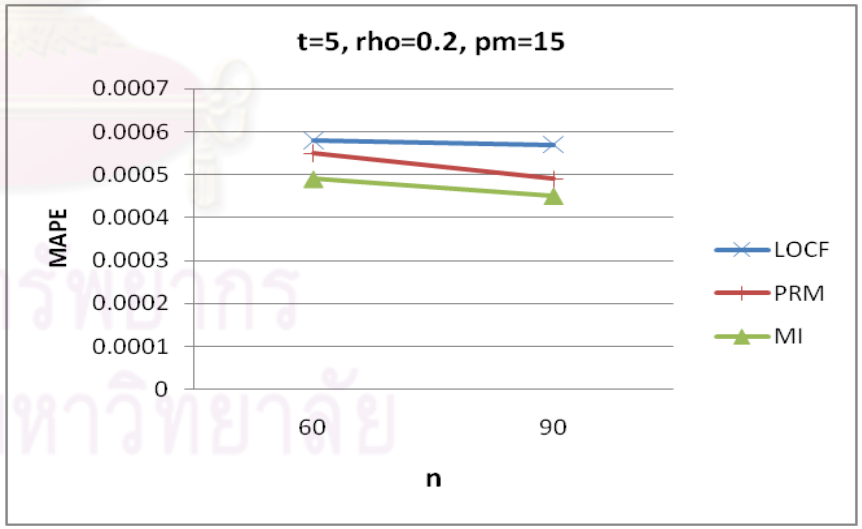


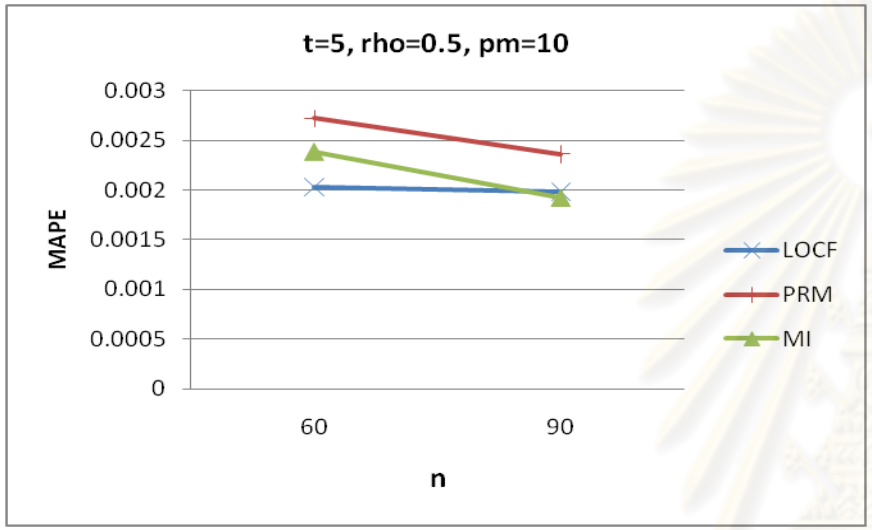


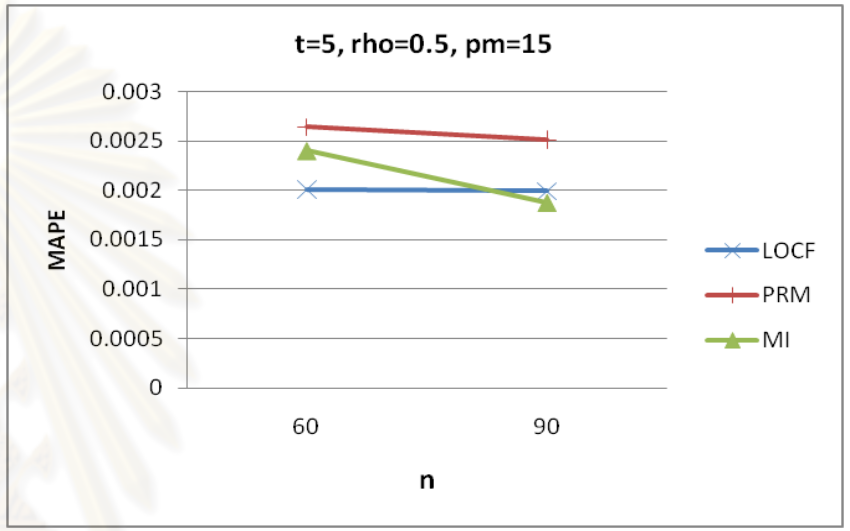


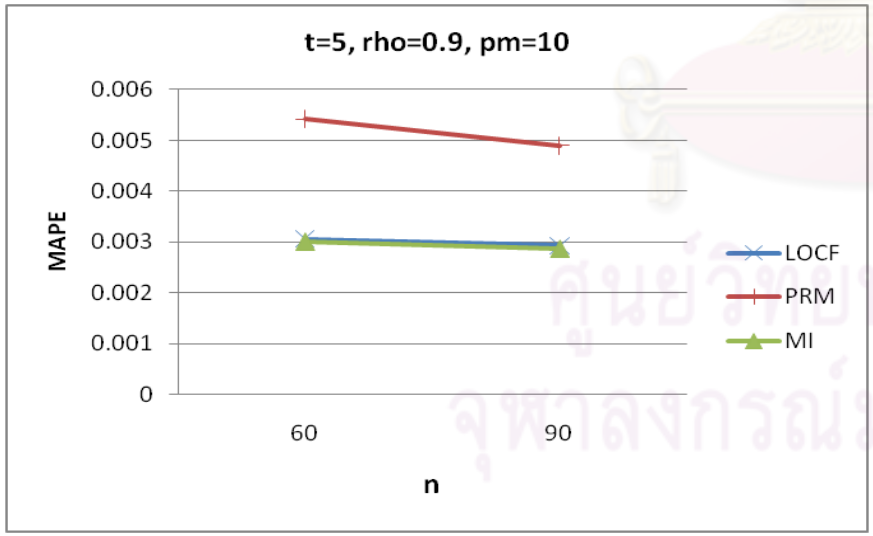


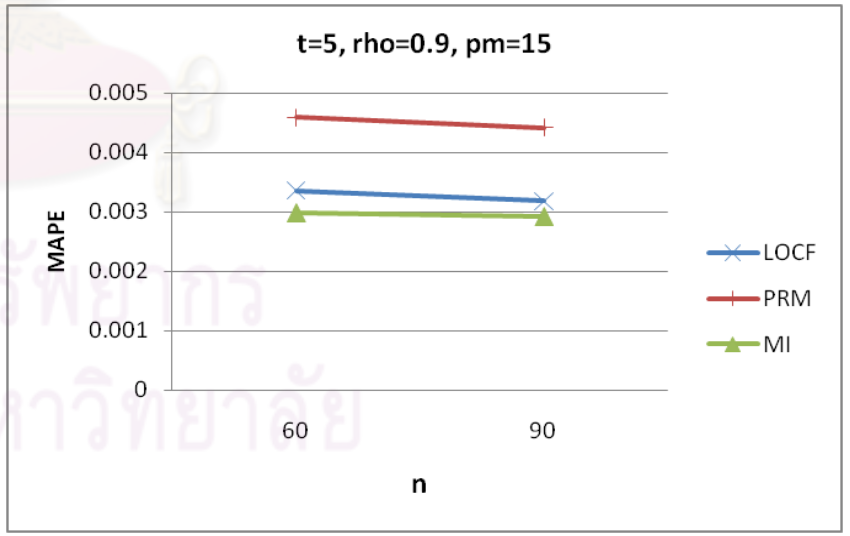












4.1.3 แสดงค่าเฉลี่ยร้อยละความคลาดเคลื่อนสัมบูรณ์ เมื่อระยะเวลาในการเก็บข้อมูลซ้ำเปลี่ยนแปลง แต่ขนาดตัวอย่าง ร้อยละการสูญหาย และระดับอัตตสหสัมพันธ์คงที่

ซึ่งผลการวิจัยนำเสนอดังตารางที่ 4.5 - 4.6 ดังนี้

ตารางที่ 4.5 แสดงค่าเฉลี่ยร้อยละความคลาดเคลื่อนสัมบูรณ์ เมื่อขนาดตัวอย่างเท่ากับ 60 ระดับอัตตสหสัมพันธ์เท่ากับ 0.2, 0.5 และ 0.9 ร้อยละการสูญหายเท่ากับ 10 และ 15 โดยจำแนกตามระยะเวลาในการเก็บข้อมูลซ้ำ

n	rho	pm	t	MAPE		
				LOCF	PRM	MI
60	0.2	10	3	0.00087	0.00067*	0.00099
			5	0.00052	0.00047*	0.00061
		15	3	0.00075	0.00068	0.00062*
			5	0.00058	0.00055	0.00049*
	0.5	10	3	0.00288	0.00284*	0.00290
			5	0.00203*	0.00272	0.00238
		15	3	0.00276*	0.00315	0.00284
			5	0.00201*	0.00264	0.00240
	0.9	10	3	0.00315	0.00618	0.00312*
			5	0.00306	0.00542	0.00301*
		15	3	0.00384	0.00592	0.00331*
			5	0.00336	0.00459	0.00298*

จากตารางที่ 4.5 เมื่อพิจารณาค่า MAPE โดยที่ขนาดตัวอย่างเท่ากับ 60 ระดับอัตตสหสัมพันธ์เท่ากับ 0.2, 0.5 และ 0.9 ร้อยละการสูญหายเท่ากับ 10 และ 15 พบว่า เมื่อระยะเวลาในการเก็บข้อมูลซ้ำเพิ่มสูงขึ้น ค่าของ MAPE มีแนวโน้มลดลง

เมื่อพิจารณาเปรียบเทียบค่า MAPE กรณีที่ขนาดตัวอย่างเท่ากับ 60 พบว่า

เมื่ออัตตสหสัมพันธ์ระดับต่ำ (0.2) ร้อยละการสูญหายเท่ากับ 10 พบว่า วิธี PRM จะให้ค่า MAPE ต่ำที่สุดที่ระยะเวลาในการเก็บข้อมูลซ้ำทั้ง 2 ระดับคือ 3 และ 5 คาบเวลา แต่เมื่อร้อยละ

การสูญหายเพิ่มสูงขึ้นเท่ากับ 15 พบว่า วิธี MI จะให้ค่า MAPE ต่ำที่สุดที่ระยะเวลาในการเก็บข้อมูลซ้ำทั้ง 2 ระดับคือ 3 และ 5 คาบเวลา

เมื่ออัตตสหสัมพันธ์ระดับปานกลาง (0.5) ร้อยละการสูญหายเท่ากับ 10 พบว่า วิธี PRM จะให้ค่า MAPE ต่ำที่สุดเมื่อระยะเวลาในการเก็บข้อมูลซ้ำเท่ากับ 3 คาบเวลา และวิธี LOCF จะให้ค่า MAPE ต่ำที่สุดเมื่อระยะเวลาในการเก็บข้อมูลซ้ำเท่ากับ 5 คาบเวลา แต่เมื่อร้อยละการสูญหายเพิ่มสูงขึ้นเท่ากับ 15 พบว่า วิธี LOCF จะให้ค่า MAPE ต่ำที่สุดที่ระยะเวลาในการเก็บข้อมูลซ้ำทั้ง 2 ระดับคือ 3 และ 5 คาบเวลา

เมื่ออัตตสหสัมพันธ์ระดับสูง (0.9) ทุกร้อยละการสูญหาย ทุกระยะเวลาในการเก็บข้อมูลซ้ำ วิธี MI จะให้ค่า MAPE ต่ำที่สุด

ตารางที่ 4.6 แสดงค่าเฉลี่ยร้อยละความคลาดเคลื่อนสัมบูรณ์ เมื่อขนาดตัวอย่างเท่ากับ 90 ระดับอัตตสหสัมพันธ์เท่ากับ 0.2, 0.5 และ 0.9 ร้อยละการสูญหายเท่ากับ 10 และ 15 โดยจำแนกตามระยะเวลาในการเก็บข้อมูลซ้ำ

n	rho	pm	t	MAPE		
				LOCF	PRM	MI
90	0.2	10	3	0.00086	0.00067*	0.00098
			5	0.00052	0.00045	0.00044*
		15	3	0.00075	0.00065	0.00061*
			5	0.00057	0.00049	0.00045*
	0.5	10	3	0.00283	0.00281*	0.00284
			5	0.00198	0.00236	0.00192*
		15	3	0.00275	0.00311	0.00273*
			5	0.00199	0.00251	0.00187*
	0.9	10	3	0.00309	0.00607	0.00301*
			5	0.00293	0.0049	0.00287*
		15	3	0.00379	0.00462	0.0033*
			5	0.00318	0.00442	0.00292*

จากตารางที่ 4.6 เมื่อพิจารณาค่า MAPE โดยที่ขนาดตัวอย่างเท่ากับ 90 ระดับอัตตสหสัมพันธ์เท่ากับ 0.2, 0.5 และ 0.9 ร้อยละการสูญหายเท่ากับ 10 และ 15 พบว่า เมื่อระยะเวลาในการเก็บข้อมูลซ้ำเพิ่มสูงขึ้น ค่าของ MAPE มีแนวโน้มลดลง

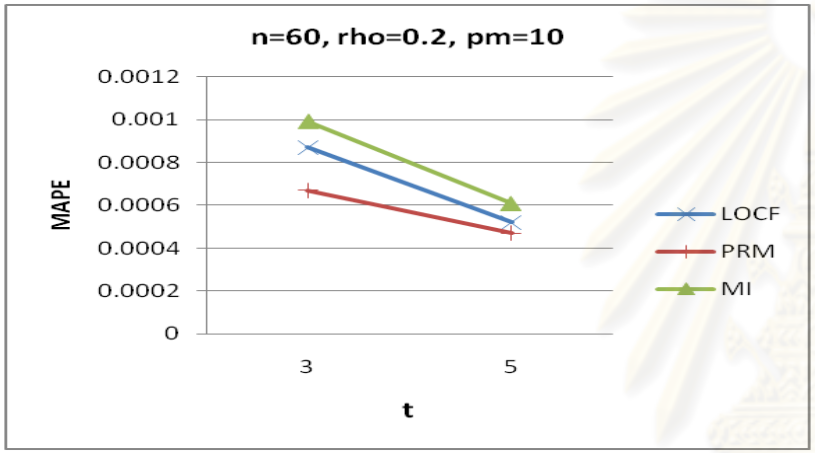
เมื่อพิจารณาเปรียบเทียบค่า MAPE กรณีที่ขนาดตัวอย่างเท่ากับ 90 พบว่า

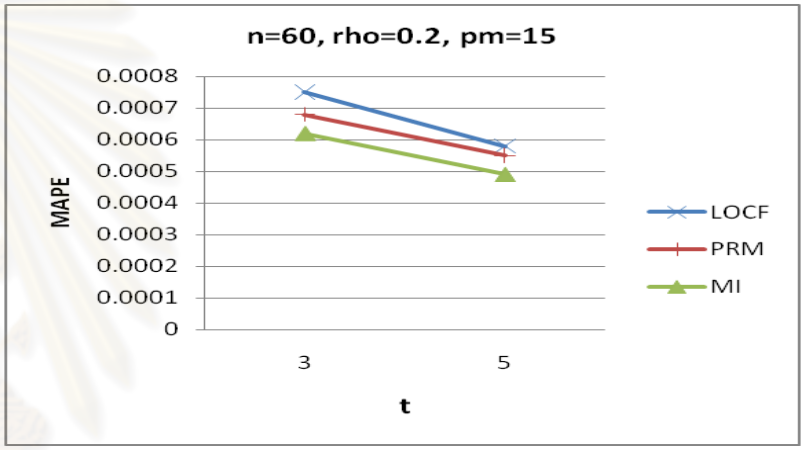
เมื่ออัตตสหสัมพันธ์ระดับต่ำ (0.2) ร้อยละการสูญหายเท่ากับ 10 พบว่า วิธี PRM จะให้ค่า MAPE ต่ำที่สุดเมื่อระยะเวลาในการเก็บข้อมูลซ้ำเท่ากับ 3 คาบเวลา และวิธี MI จะให้ค่า MAPE ต่ำที่สุดเมื่อระยะเวลาในการเก็บข้อมูลซ้ำเท่ากับ 5 คาบเวลา แต่เมื่อร้อยละการสูญหายเพิ่มสูงขึ้นเท่ากับ 15 พบว่า วิธี MI จะให้ค่า MAPE ต่ำที่สุดที่ระยะเวลาในการเก็บข้อมูลซ้ำทั้ง 2 ระดับคือ 3 และ 5 คาบเวลา

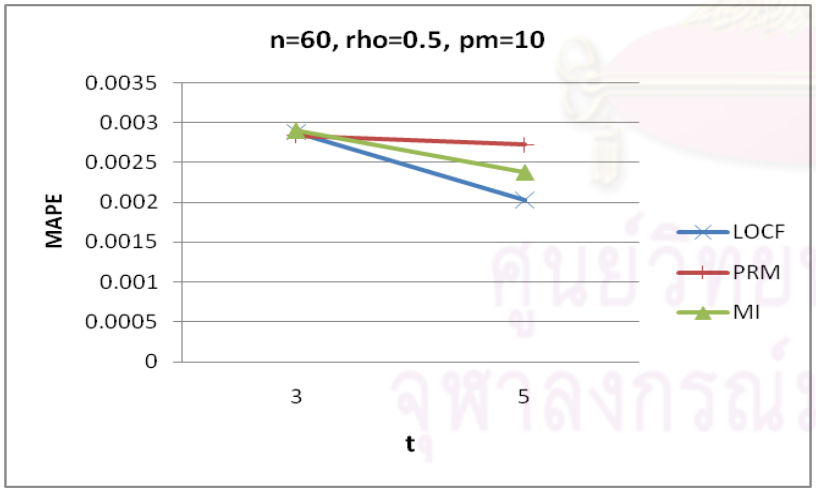
เมื่ออัตตสหสัมพันธ์ระดับปานกลาง (0.5) ร้อยละการสูญหายเท่ากับ 10 พบว่า วิธี PRM จะให้ค่า MAPE ต่ำที่สุดเมื่อระยะเวลาในการเก็บข้อมูลซ้ำเท่ากับ 3 คาบเวลา และวิธี MI จะให้ค่า MAPE ต่ำที่สุดเมื่อระยะเวลาในการเก็บข้อมูลซ้ำเท่ากับ 5 คาบเวลา แต่เมื่อร้อยละการสูญหายเพิ่มสูงขึ้นเท่ากับ 15 พบว่า วิธี MI จะให้ค่า MAPE ต่ำที่สุดที่ระยะเวลาในการเก็บข้อมูลซ้ำทั้ง 2 ระดับคือ 3 และ 5 คาบเวลา

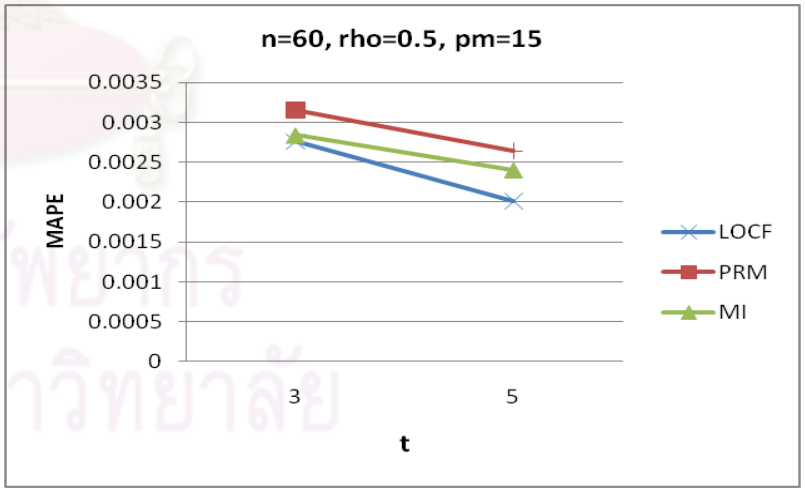
เมื่ออัตตสหสัมพันธ์ระดับสูง (0.9) ทุกร้อยละการสูญหาย ทุกระยะเวลาในการเก็บข้อมูลซ้ำ วิธี MI จะให้ค่า MAPE ต่ำที่สุด

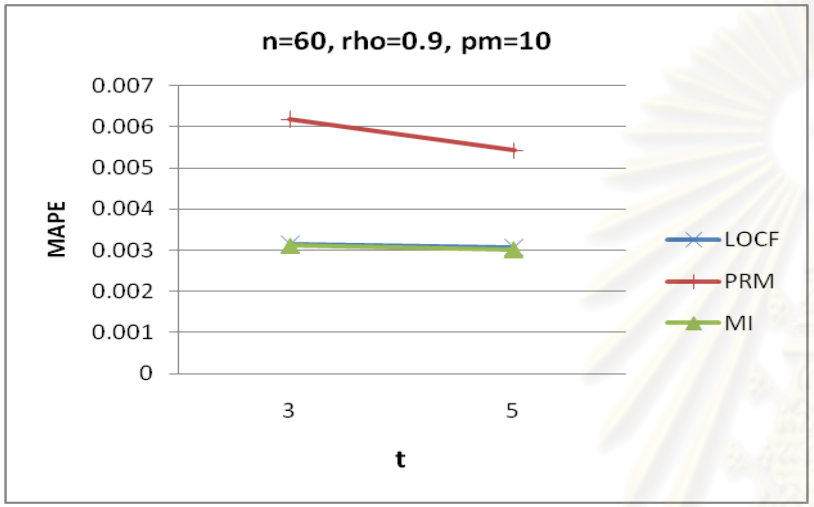
รูปที่ 4.3 แสดงค่าเฉลี่ยร้อยละความคลาดเคลื่อนสัมบูรณ์ เมื่อระยะเวลาในการเก็บข้อมูลซ้ำเปลี่ยนแปลง แต่ขนาดตัวอย่าง ร้อยละการสูญหาย และระดับอัตตสหสัมพันธ์คงที่

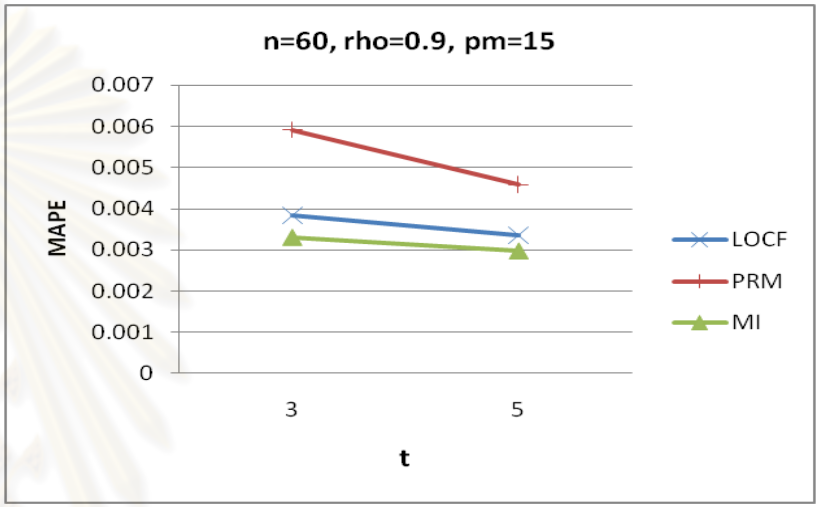


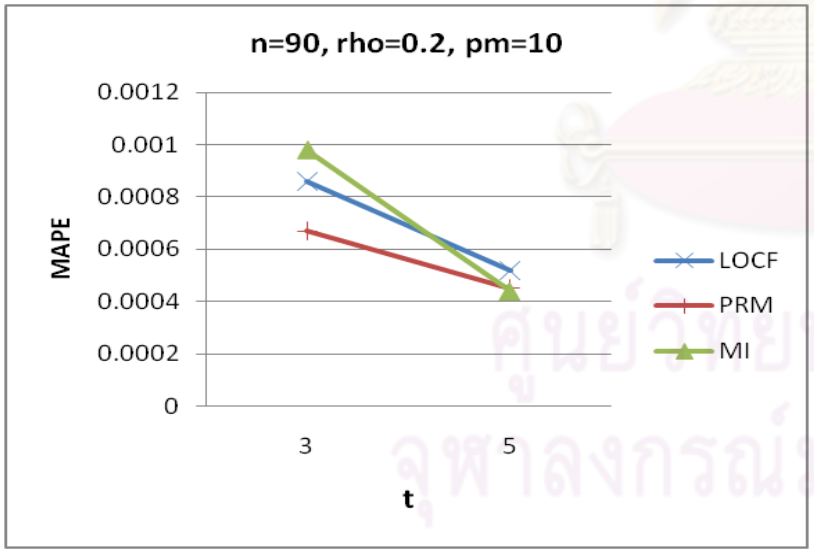


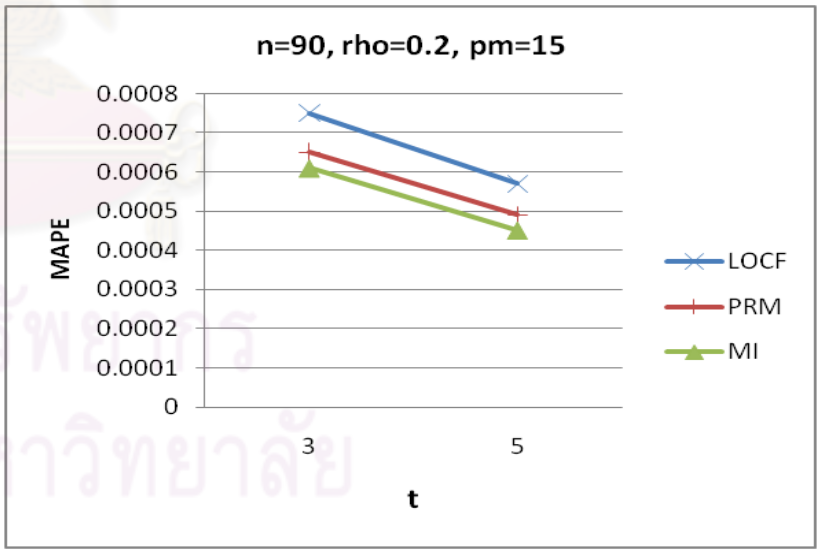


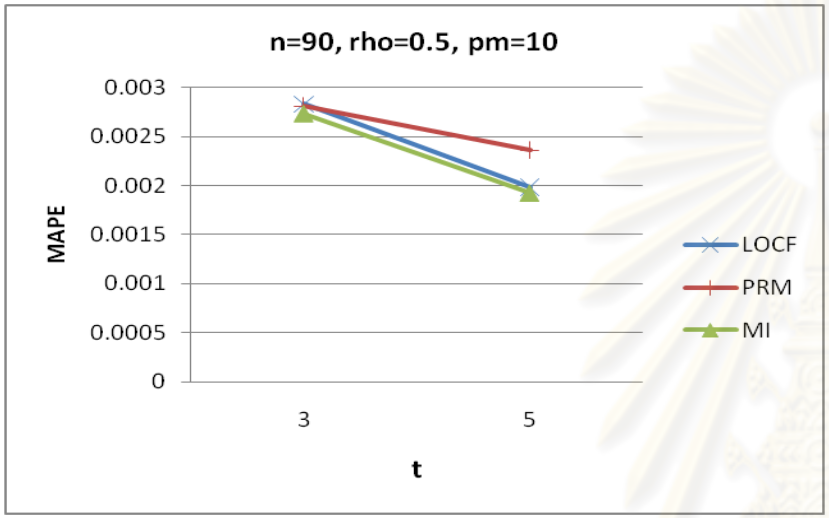


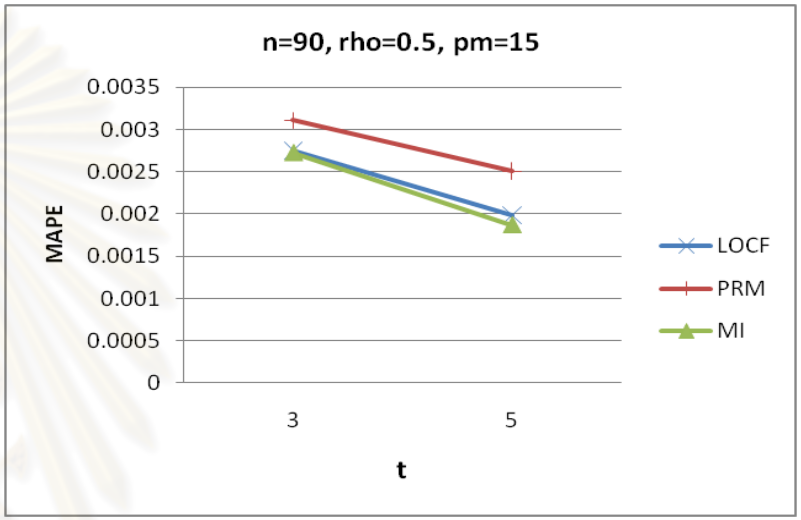


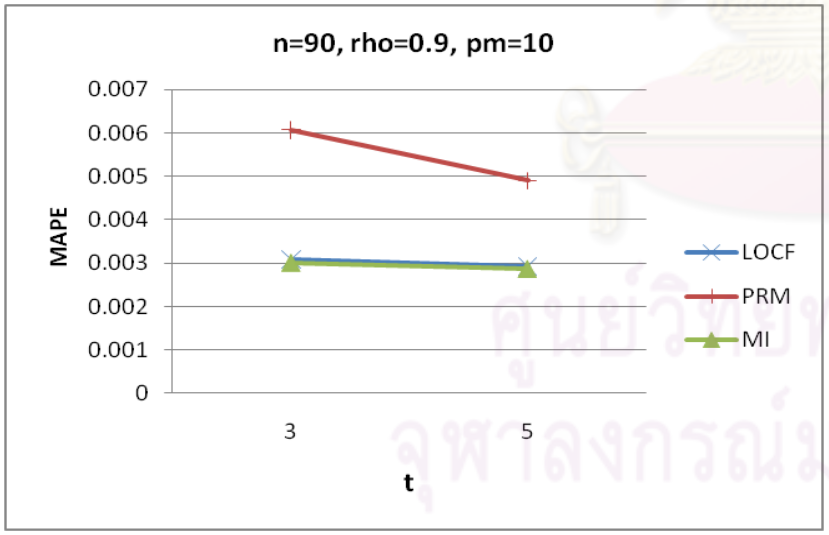


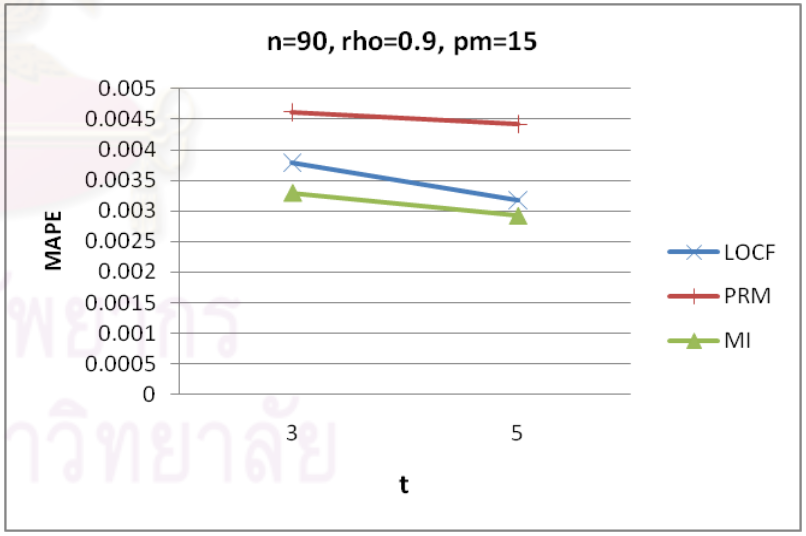












4.1.4 แสดงค่าเฉลี่ยร้อยละความคลาดเคลื่อนสัมบูรณ์ เมื่อร้อยละการสูญหายเปลี่ยนแปลง แต่ขนาดตัวอย่าง ระยะเวลาในการเก็บข้อมูลซ้ำ และระดับอัตตสหสัมพันธ์คงที่

ซึ่งผลการวิจัยนำเสนอดังตารางที่ 4.7 – 4.9 ดังนี้

ตารางที่ 4.7 แสดงค่าเฉลี่ยร้อยละความคลาดเคลื่อนสัมบูรณ์ เมื่ออัตตสหสัมพันธ์ระดับต่ำ (0.2) ขนาดตัวอย่างเท่ากับ 60 และ 90 ระยะเวลาในการเก็บข้อมูลซ้ำเท่ากับ 3 และ 5 คาบเวลา โดยจำแนกตามร้อยละการสูญหาย

rho	n	t	pm	MAPE		
				LOCF	PRM	MI
0.2	60	3	10	0.00087	0.00067*	0.00099
			15	0.00075	0.00068	0.00062*
		5	10	0.00052	0.00047*	0.00061
			15	0.00058	0.00055	0.00049*
	90	3	10	0.00086	0.00067*	0.00098
			15	0.00075	0.00065	0.00061*
		5	10	0.00052	0.00045	0.00044*
			15	0.00057	0.00049	0.00045*

จากตารางที่ 4.7 เมื่อพิจารณาค่า MAPE โดยที่อัตตสหสัมพันธ์ระดับต่ำ (0.2) ขนาดตัวอย่างเท่ากับ 60 และ 90 ระยะเวลาในการเก็บข้อมูลซ้ำเท่ากับ 3 และ 5 คาบเวลา พบว่า เมื่อระยะเวลาในการเก็บข้อมูลซ้ำเพิ่มสูงขึ้น ค่าของ MAPE มีแนวโน้มเพิ่มขึ้นหรือลดลงไม่คงที่

เมื่อพิจารณาเปรียบเทียบค่า MAPE กรณีที่อัตตสหสัมพันธ์ระดับต่ำ (0.2) พบว่า

เมื่อขนาดตัวอย่างเท่ากับ 60 ที่ร้อยละการสูญหายเท่ากับ 10 วิธี PRM จะให้ค่า MAPE ต่ำที่สุด และที่วิธี MI จะให้ค่า MAPE ต่ำที่สุดที่ร้อยละการสูญหายเท่ากับ 15 ทุกระยะเวลาในการเก็บข้อมูลซ้ำ

เมื่อขนาดตัวอย่างเพิ่มสูงขึ้นเท่ากับ 90 ที่ระยะเวลาในการเก็บข้อมูลซ้ำเท่ากับ 3 คาบเวลา พบว่า เมื่อร้อยละการสูญหายเท่ากับ 10 วิธี PRM จะให้ค่า MAPE ต่ำที่สุด และเมื่อร้อยละการสูญหายเท่ากับ 15 วิธี MI จะให้ค่า MAPE ต่ำที่สุด ซึ่งเมื่อระยะเวลาในการเก็บข้อมูลซ้ำ

เพิ่มขึ้นเป็น 5 คาบเวลา พบว่า วิธี MI จะให้ค่า MAPE ต่ำที่สุด ที่ร้อยละการสูญหายทั้ง 2 ระดับคือ ร้อยละ 10 และร้อยละ 15

ตารางที่ 4.8 แสดงค่าเฉลี่ยร้อยละความคลาดเคลื่อนสัมบูรณ์ เมื่ออัตตสหสัมพันธ์ระดับปานกลาง (0.5) ขนาดตัวอย่างเท่ากับ 60 และ 90 ระยะเวลาในการเก็บข้อมูลซ้ำเท่ากับ 3 และ 5 คาบเวลา โดยจำแนกตามร้อยละการสูญหาย

rho	n	t	pm	MAPE		
				LOCF	PRM	MI
0.5	60	3	10	0.00288	0.00284*	0.00290
			15	0.00276*	0.00315	0.00284
		5	10	0.00203*	0.00272	0.00238
			15	0.00201*	0.00264	0.00240
	90	3	10	0.00283	0.00281*	0.00284
			15	0.00275	0.00311	0.00273*
		5	10	0.00198	0.00236	0.00192*
			15	0.00199	0.00251	0.00187*

จากตารางที่ 4.8 เมื่อพิจารณาค่า MAPE โดยที่อัตตสหสัมพันธ์ระดับปานกลาง (0.5) ขนาดตัวอย่างเท่ากับ 60 และ 90 ระยะเวลาในการเก็บข้อมูลซ้ำเท่ากับ 3 และ 5 คาบเวลา พบว่า เมื่อระยะเวลาในการเก็บข้อมูลซ้ำเพิ่มสูงขึ้น ค่าของ MAPE มีแนวโน้มเพิ่มขึ้นหรือลดลงไม่คงที่

เมื่อพิจารณาเปรียบเทียบค่า MAPE กรณีที่อัตตสหสัมพันธ์ระดับปานกลาง (0.5) พบว่า

เมื่อขนาดตัวอย่างเท่ากับ 60 ที่ระยะเวลาในการเก็บข้อมูลซ้ำเท่ากับ 3 คาบเวลา พบว่า เมื่อร้อยละการสูญหายเท่ากับ 10 วิธี PRM จะให้ค่า MAPE ต่ำที่สุด และเมื่อร้อยละการสูญหายเท่ากับ 15 วิธี LOCF จะให้ค่า MAPE ต่ำที่สุด ซึ่งเมื่อระยะเวลาในการเก็บข้อมูลซ้ำเพิ่มขึ้นเป็น 5 คาบเวลา พบว่า วิธี LOCF จะให้ค่า MAPE ต่ำที่สุด ที่ร้อยละการสูญหายทั้ง 2 ระดับคือ ร้อยละ 10 และร้อยละ 15

เมื่อขนาดตัวอย่างเพิ่มสูงขึ้นเท่ากับ 90 ที่ระยะเวลาในการเก็บข้อมูลซ้ำเท่ากับ 3 คาบเวลา พบว่า เมื่อร้อยละการสูญหายเท่ากับ 10 วิธี PRM จะให้ค่า MAPE ต่ำที่สุด และเมื่อร้อยละการสูญหายเท่ากับ 15 วิธี MI จะให้ค่า MAPE ต่ำที่สุด ซึ่งเมื่อระยะเวลาในการเก็บข้อมูลซ้ำ

เพิ่มขึ้นเป็น 5 คาบเวลา พบว่า วิธี MI จะให้ค่า MAPE ต่ำที่สุด ที่ร้อยละการสูญหายทั้ง 2 ระดับคือ ร้อยละ 10 และร้อยละ 15

ตารางที่ 4.9 แสดงค่าเฉลี่ยร้อยละความคลาดเคลื่อนสัมบูรณ์ เมื่ออัตตสหสัมพันธ์ระดับสูง (0.9) ขนาดตัวอย่างเท่ากับ 60 และ 90 ระยะเวลาในการเก็บข้อมูลซ้ำเท่ากับ 3 และ 5 คาบเวลา โดยจำแนกตามร้อยละการสูญหาย

rho	n	t	pm	MAPE		
				LOCF	PRM	MI
0.9	60	3	10	0.00315	0.00618	0.00312*
			15	0.00384	0.00592	0.00331*
		5	10	0.00306	0.00542	0.00301*
			15	0.00336	0.00459	0.00298*
	90	3	10	0.00309	0.00607	0.00301*
			15	0.00379	0.00462	0.00330*
		5	10	0.00293	0.00490	0.00287*
			15	0.00318	0.00442	0.00292*

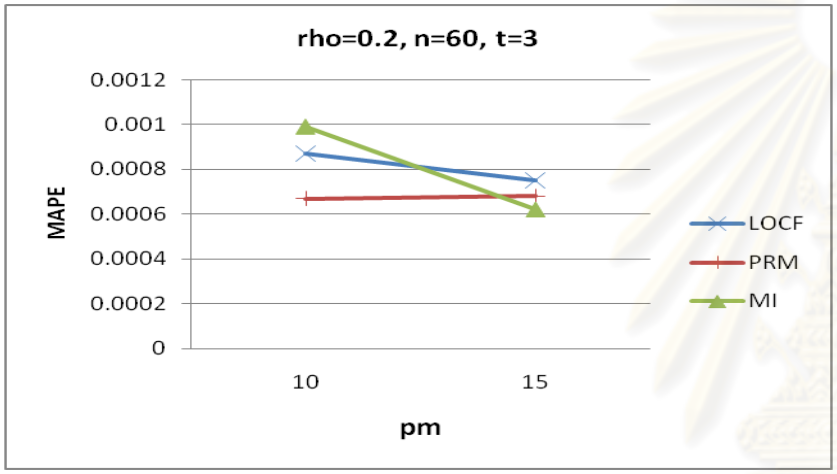
จากตารางที่ 4.9 เมื่อพิจารณาค่า MAPE โดยที่อัตตสหสัมพันธ์ระดับสูง (0.9) ขนาดตัวอย่างเท่ากับ 60 และ 90 ระยะเวลาในการเก็บข้อมูลซ้ำเท่ากับ 3 และ 5 คาบเวลา พบว่า เมื่อระยะเวลาในการเก็บข้อมูลซ้ำเพิ่มสูงขึ้น ค่าของ MAPE มีแนวโน้มเพิ่มขึ้นหรือลดลงไม่คงที่

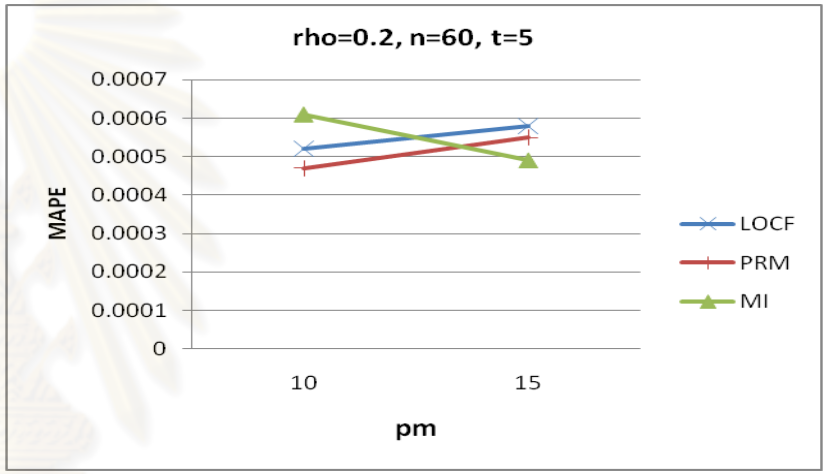
เมื่อพิจารณาเปรียบเทียบค่า MAPE กรณีที่อัตตสหสัมพันธ์ระดับสูง (0.9) พบว่า

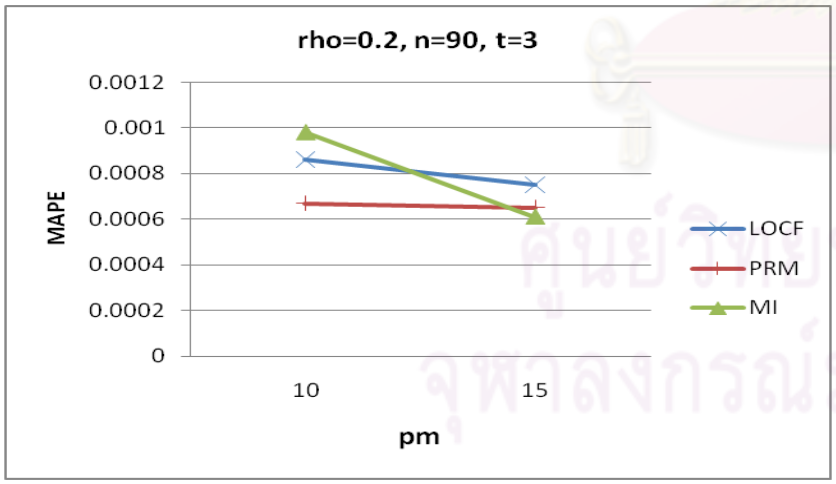
เมื่อขนาดตัวอย่างเท่ากับ 60 ทุกระยะเวลาในการเก็บข้อมูลซ้ำ และทุกร้อยละการสูญหาย วิธี MI จะให้ค่า MAPE ต่ำที่สุด

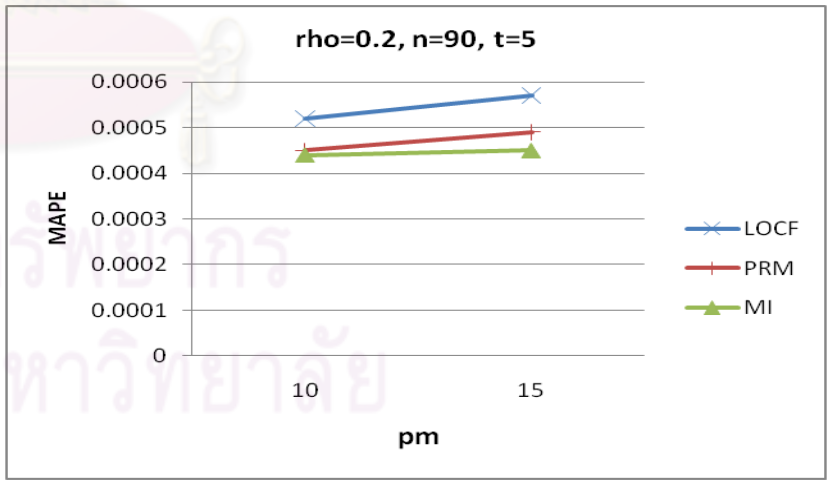
เมื่อขนาดตัวอย่างเพิ่มสูงขึ้นเท่ากับ 90 ทุกระยะเวลาในการเก็บข้อมูลซ้ำ และทุกร้อยละการสูญหาย วิธี MI จะให้ค่า MAPE ต่ำที่สุด

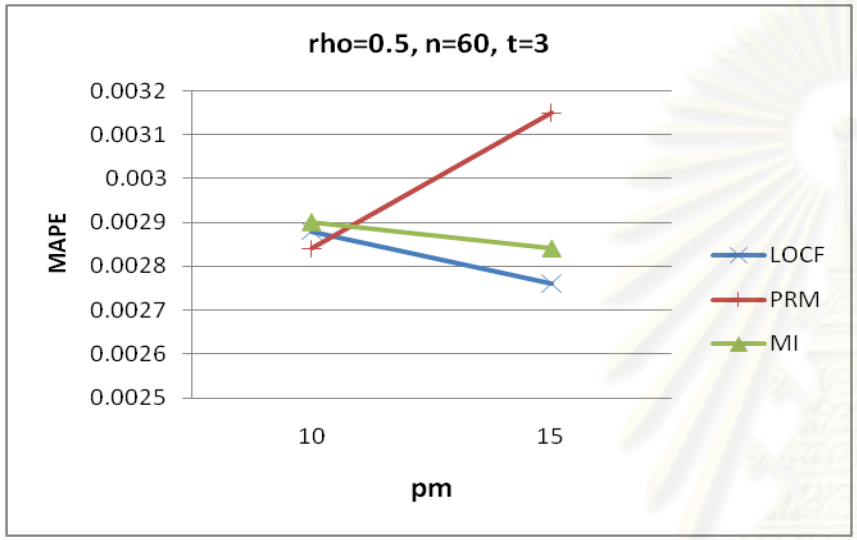
รูปที่ 4.4 แสดงค่าเฉลี่ยร้อยละความคลาดเคลื่อนสัมบูรณ์ เมื่อร้อยละการสูญหายเปลี่ยนแปลง แต่ขนาดตัวอย่าง ระยะเวลาในการเก็บข้อมูลซ้ำ และระดับอัตตสหสัมพันธ์คงที่

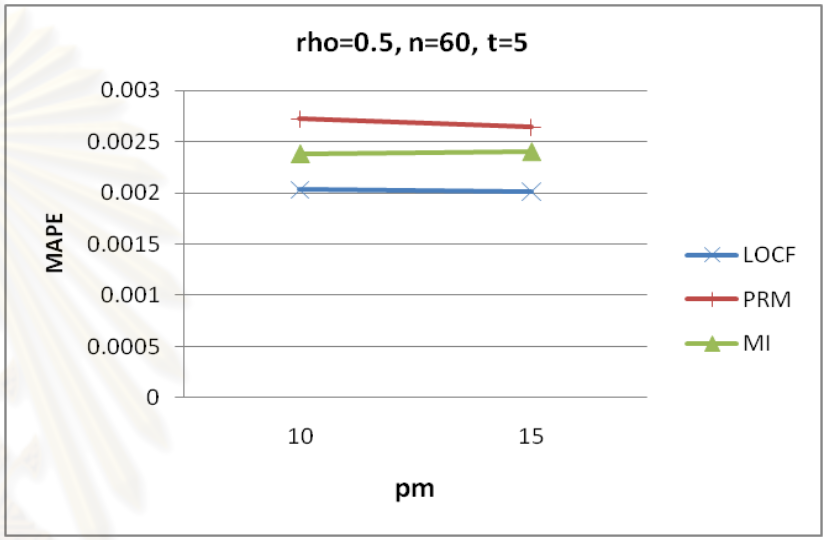


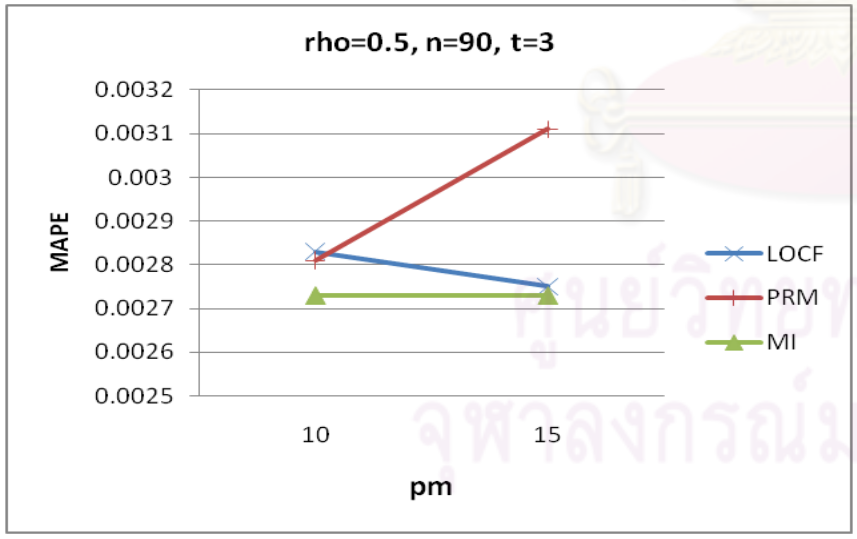


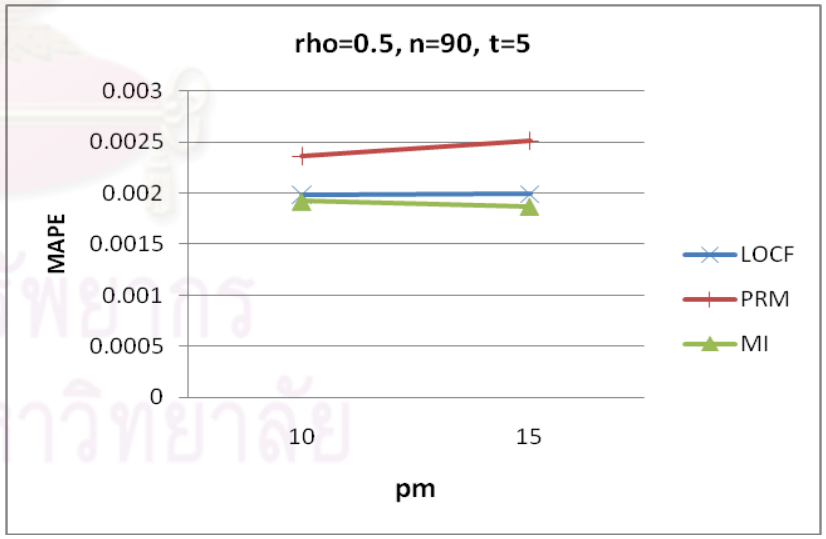


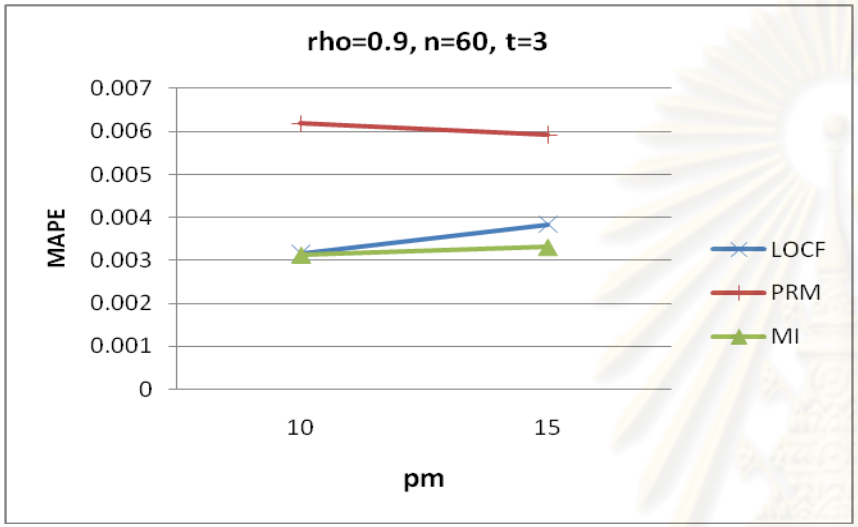


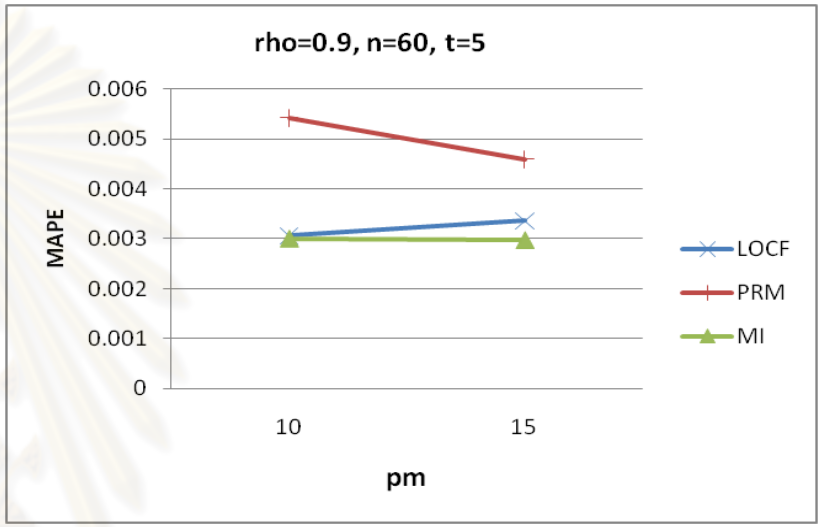


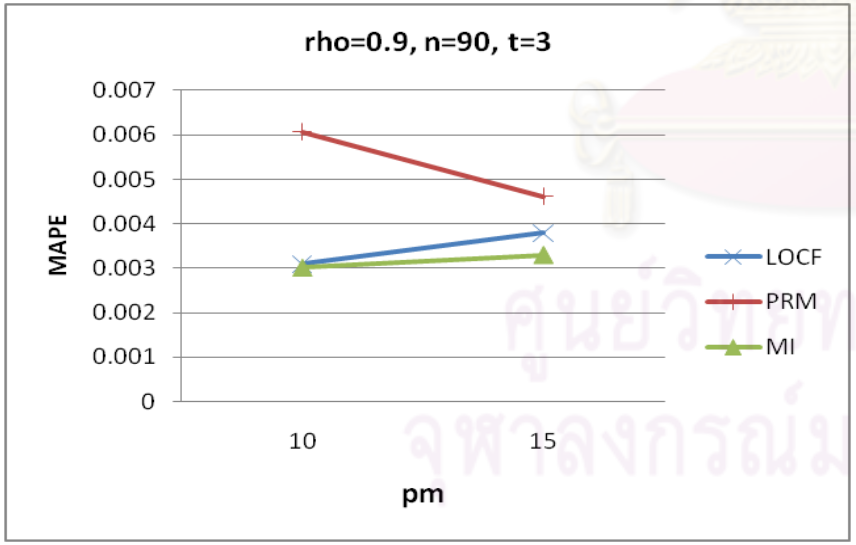


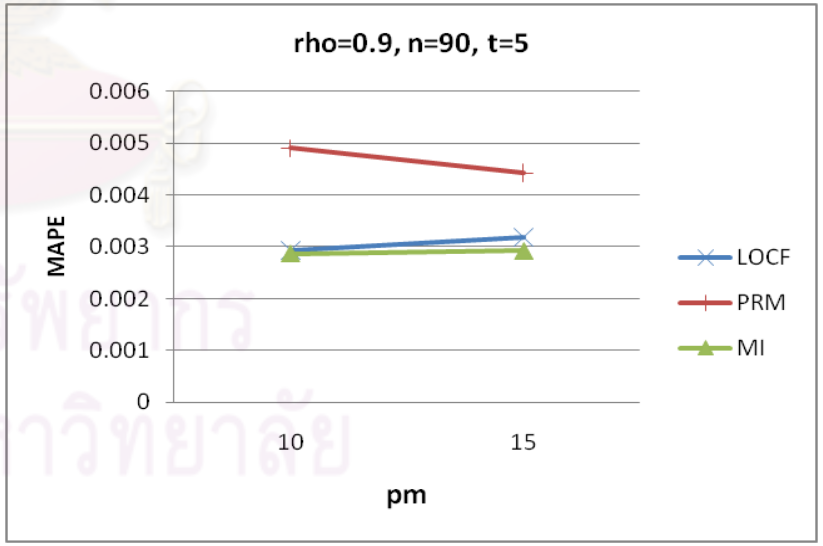












4.2 ผลสรุปการเปรียบเทียบค่าเฉลี่ยร้อยละความคลาดเคลื่อนสัมบูรณ์ของวิธีการประมาณค่าสูญหายทั้ง 3 วิธี

ตารางที่ 4.10 แสดงค่าเฉลี่ยร้อยละความคลาดเคลื่อนสัมบูรณ์ของวิธีการประมาณค่าสูญหายทั้ง 3 วิธี เมื่ออัตตสหสัมพันธ์ระดับต่ำ (0.2)

n	t	pm	MAPE			Relative ** ระหว่างวิธี LOCF และวิธี PRM	Relative** ระหว่างวิธี PRM และวิธี MI	Relative** ระหว่างวิธี LOCF และวิธี MI
			LOCF	PRM	MI			
60	3	10	0.00087	0.00067*	0.00099	0.298507463	0.47761194	0.137931034
		15	0.00075	0.00068	0.00062*	0.102941176	0.096774194	0.209677419
	5	10	0.00052	0.00047*	0.00061	0.106382979	0.29787234	0.173076923
		15	0.00058	0.00055	0.00049*	0.054545455	0.12244898	0.183673469
90	3	10	0.00086	0.00067*	0.00098	0.28358209	0.462686567	0.139534884
		15	0.00075	0.00065	0.00061*	0.153846154	0.06557377	0.229508197
	5	10	0.00052	0.00045	0.00044*	0.155555556	0.022727273	0.181818182
		15	0.00057	0.00049	0.00045*	0.163265306	0.088888889	0.266666667

จากตารางที่ 4.10 เมื่อพิจารณาเปรียบเทียบค่า MAPE กรณีที่อัตตสหสัมพันธ์ระดับต่ำ (0.2)

จากวิธีการประมาณค่าสูญหายที่มีค่า MAPE ต่ำที่สุด * พบว่าเมื่อขนาดตัวอย่าง ระยะเวลาในการเก็บข้อมูลซ้ำ และร้อยละการสูญหายเปลี่ยนแปลงไป วิธีการประมาณค่าสูญหายที่ให้ค่า MAPE ต่ำที่สุดนั้นจะมีค่า MAPE แตกต่างจากวิธีการประมาณค่าสูญหายด้วยวิธีอื่น ๆ ไม่มากนัก โดยมีร้อยละความแตกต่างตั้งแต่ร้อยละ 0.02 แต่ไม่เกิน 0.5

เนื่องจากค่าร้อยละความแตกต่างของค่า MAPE ไม่สูงนัก ดังนั้น ในทางปฏิบัติสามารถที่จะใช้วิธีการประมาณค่าสูญหายด้วยวิธีที่สามารถทำได้ง่าย และสะดวกกว่า

หมายเหตุ ** คือ การหาร้อยละความแตกต่างของค่า MAPE สำหรับวิธีการประมาณค่าสูญหาย ซึ่งคำนวณได้ดังนี้ $\frac{Max(MAPE) - Min(MAPE)}{Min(MAPE)} \times 100$

โดย $Max(MAPE)$ คือ วิธีการประมาณค่าสูญหายที่ให้ค่า MAPE สูง เมื่อทำการเปรียบเทียบทั้ง 2 วิธี

$Min(MAPE)$ คือ วิธีการประมาณค่าสูญหายที่ให้ค่า MAPE ต่ำ เมื่อทำการเปรียบเทียบทั้ง 2 วิธี

ตารางที่ 4.11 แสดงค่าเฉลี่ยร้อยละความคลาดเคลื่อนสัมบูรณ์ของวิธีการประมาณค่าสูญหายทั้ง 3 วิธี เมื่ออัตตสหสัมพันธ์ระดับปานกลาง (0.5)

n	t	pm	MAPE			Relative ระหว่างวิธี LOCF และวิธี PRM	Relative ระหว่างวิธี PRM และวิธี MI	Relative ระหว่างวิธี LOCF และวิธี MI
			LOCF	PRM	MI			
60	3	10	0.00288	0.00284*	0.0029	0.014084507	0.021126761	0.006944444
		15	0.00276*	0.00315	0.00284	0.141304348	0.10915493	0.028985507
	5	10	0.00203*	0.00272	0.00238	0.339901478	0.142857143	0.172413793
		15	0.00201*	0.00264	0.0024	0.313432836	0.1	0.194029851
90	3	10	0.00283	0.00281*	0.00284	0.007117438	0.010676157	0.003533569
		15	0.00275	0.00311	0.00273*	0.130909091	0.139194139	0.007326007
	5	10	0.00198	0.00236	0.00192*	0.191919192	0.229166667	0.03125
		15	0.00199	0.00251	0.00187*	0.261306532	0.342245989	0.064171122

จากตารางที่ 4.11 เมื่อพิจารณาเปรียบเทียบค่า MAPE กรณีที่อัตตสหสัมพันธ์ระดับปานกลาง (0.5)

จากวิธีการประมาณค่าสูญหายที่มีค่า MAPE ต่ำที่สุด * พบว่าเมื่อขนาดตัวอย่าง ระยะเวลาในการเก็บข้อมูลซ้ำ และร้อยละการสูญหายเปลี่ยนแปลงไป วิธีการประมาณค่าสูญหายที่ให้ค่า MAPE ต่ำที่สุดนั้นจะมีค่า MAPE แตกต่างจากวิธีการประมาณค่าสูญหายด้วยวิธีอื่น ๆ ไม่มากนัก โดยมีร้อยละความแตกต่างตั้งแต่ร้อยละ 0.003 แต่ไม่เกินร้อยละ 0.35

เนื่องจากค่าร้อยละความแตกต่างของค่า MAPE ไม่สูงนัก ดังนั้น ในทางปฏิบัติสามารถที่จะใช้วิธีการประมาณค่าสูญหายด้วยวิธีที่สามารถทำได้ง่าย และสะดวกกว่า

ตารางที่ 4.12 แสดงค่าเฉลี่ยร้อยละความคลาดเคลื่อนสัมบูรณ์ของวิธีการประมาณค่าสูญหายทั้ง 3 วิธี เมื่ออัตตสหสัมพันธ์ระดับสูง (0.9)

n	t	pm	MAPE			Relative ระหว่างวิธี LOCF และวิธี PRM	Relative ระหว่างวิธี PRM และวิธี MI	Relative ระหว่างวิธี LOCF และวิธี MI
			LOCF	PRM	MI			
60	3	10	0.00315	0.00618	0.00312*	0.961904762	0.980769231	0.009615385
		15	0.00384	0.00592	0.00331*	0.541666667	0.788519637	0.160120846
	5	10	0.00306	0.00542	0.00301*	0.77124183	0.800664452	0.016611296
		15	0.00336	0.00459	0.00298*	0.366071429	0.540268456	0.127516779
90	3	10	0.00309	0.00607	0.00301*	0.964401294	1.016611296	0.026578073
		15	0.00379	0.00462	0.0033*	0.218997361	0.4	0.148484848
	5	10	0.00293	0.0049	0.00287*	0.672354949	0.707317073	0.020905923
		15	0.00318	0.00442	0.00292*	0.389937107	0.51369863	0.089041096

จากตารางที่ 4.12 เมื่อพิจารณาเปรียบเทียบค่า MAPE กรณีที่อัตตสหสัมพันธ์ระดับสูง (0.9) พบว่า

จากวิธีการประมาณค่าสูญหายที่มีค่า MAPE ต่ำที่สุด * พบว่าเมื่อขนาดตัวอย่าง ระยะเวลาในการเก็บข้อมูลซ้ำ และร้อยละการสูญหายเปลี่ยนแปลงไป วิธีการประมาณค่าสูญหายที่ให้ค่า MAPE ต่ำที่สุดนั้นจะมีค่า MAPE แตกต่างจากวิธีการประมาณค่าสูญหายด้วยวิธีอื่น ๆ ไม่มากนัก โดยมีร้อยละความแตกต่างตั้งแต่ร้อยละ 0.009 แต่ไม่เกินร้อยละ 1

เนื่องจากค่าร้อยละความแตกต่างของค่า MAPE ไม่สูงนัก ดังนั้น ในทางปฏิบัติสามารถที่จะใช้วิธีการประมาณค่าสูญหายด้วยวิธีที่สามารถทำได้ง่าย และสะดวกกว่า

บทที่ 5

สรุปผลการวิจัย อภิปรายผล และข้อเสนอแนะ

การวิจัยครั้งนี้มีวัตถุประสงค์เพื่อศึกษาและเปรียบเทียบวิธีการประมาณค่าสูญหายของตัวแปรตามสำหรับข้อมูลระยะยาวในตัวแบบ Generalized Estimating Equations เมื่อข้อมูลระยะยาวมีอัตตสหสัมพันธ์ในตัวเองรูปแบบอัตตสหสัมพันธ์อันดับที่หนึ่ง (AR(1)) โดยทำการเปรียบเทียบวิธีการประมาณค่าสูญหาย 3 วิธีคือ วิธี Last Observation Carried Forward วิธี Previous Row Mean และวิธี Multiple Imputation ซึ่งศึกษาภายใต้สถานการณ์ต่าง ๆ ที่กำหนดดังนี้

1. ขนาดตัวอย่าง (n) ที่ใช้ในการศึกษาเท่ากับ 60 และ 90
2. ระยะเวลาในการเก็บข้อมูลซ้ำ (t) เท่ากับ 3 และ 5 คาบเวลา
3. สัมประสิทธิ์สหสัมพันธ์ (ρ) เท่ากับ 0.2 , 0.5 และ 0.9
4. กำหนดการสูญหายเกิดขึ้นในตัวแปรตามและเป็นการสูญหายแบบสุ่ม คิดเป็นร้อยละ 10 และ 15

สำหรับเกณฑ์ที่ใช้ในการเปรียบเทียบประสิทธิภาพของวิธีการประมาณค่าสูญหาย ผู้วิจัยพิจารณาค่าพยากรณ์ของตัวแปรตามกับค่าจริงในรูปแบบ Mean absolute percentage error (MAPE) ซึ่งวิธีการใดให้ค่า MAPE ต่ำกว่า แสดงว่าเป็นวิธีการประมาณค่าสูญหายที่ดีกว่า

5.1 สรุปผลการวิจัย

5.1.1 ผลการเปรียบเทียบค่าเฉลี่ยร้อยละความคลาดเคลื่อนสัมบูรณ์

โดยสรุปผลจำแนกตามอัตตสหสัมพันธ์ ดังนี้

- เมื่ออัตตสหสัมพันธ์ระดับต่ำ (0.2)

วิธีการประมาณค่าสูญหายด้วยวิธี PRM จะให้ค่าเฉลี่ยร้อยละความคลาดเคลื่อนสัมบูรณ์ (MAPE) ต่ำที่สุด เมื่อร้อยละการสูญหายเป็น 10 ยกเว้นที่ขนาดตัวอย่าง 90 และระยะเวลาในการเก็บข้อมูลซ้ำเท่ากับ 5 คาบเวลา วิธี MI จะให้ค่าเฉลี่ยร้อยละความคลาดเคลื่อนสัมบูรณ์ (MAPE) ต่ำที่สุด แต่เมื่อร้อยละการสูญหายเพิ่มขึ้นเป็น 15 วิธี MI จะให้ค่าเฉลี่ยร้อยละความคลาดเคลื่อนสัมบูรณ์ (MAPE) ต่ำที่สุด ทุกขนาดตัวอย่าง ทุกระยะเวลาในการเก็บข้อมูลซ้ำ

- เมื่ออัตสหสัมพันธ์ระดับปานกลาง (0.5)

โดยส่วนใหญ่ วิธีการประมาณค่าสูญหายด้วยวิธี LOCF จะให้ค่าเฉลี่ยร้อยละความคลาดเคลื่อนสัมบูรณ์ (MAPE) ต่ำที่สุด ที่ขนาดตัวอย่างเท่ากับ 60 ระยะเวลาในการเก็บข้อมูลซ้ำเท่ากับ 3 และ 5 คาบเวลา และร้อยละการสูญหายเท่ากับ 10 และ 15 ส่วนวิธีการประมาณค่าสูญหายด้วยวิธี MI จะพบว่า ส่วนใหญ่ให้ค่าเฉลี่ยร้อยละความคลาดเคลื่อนสัมบูรณ์ (MAPE) ต่ำที่สุด ที่ขนาดตัวอย่างเท่ากับ 90 ระยะเวลาในการเก็บข้อมูลซ้ำเท่ากับ 3 และ 5 คาบเวลา และร้อยละการสูญหายเท่ากับ 10 และ 15 ยกเว้นกรณีที่ระยะเวลาในการเก็บข้อมูลซ้ำเท่ากับ 3 คาบเวลา ร้อยละการสูญหายเท่ากับ 10 ทุกขนาดตัวอย่าง วิธี PRM จะให้ค่าเฉลี่ยร้อยละความคลาดเคลื่อนสัมบูรณ์ (MAPE) ต่ำที่สุด

- เมื่ออัตตสหสัมพันธ์ระดับสูง (0.9)

วิธีประมาณค่าสูญหายด้วยวิธี Multiple Imputation (MI)จะให้ค่าเฉลี่ยร้อยละความคลาดเคลื่อนสัมบูรณ์ (MAPE) ต่ำที่สุด ที่ทุกขนาดตัวอย่าง ทุกระยะเวลาในการเก็บข้อมูลซ้ำ และทุกร้อยละการสูญหาย

5.1.2 ปัจจัยที่มีผลต่อค่าเฉลี่ยร้อยละความคลาดเคลื่อนสัมบูรณ์

- **ขนาดตัวอย่าง** (n)

เมื่อขนาดตัวอย่างเพิ่มขึ้น จะส่งผลให้ค่าเฉลี่ยร้อยละความคลาดเคลื่อนสัมบูรณ์มีแนวโน้มลดลง เพราะขนาดตัวอย่างที่เพิ่มขึ้นส่งผลให้ค่าความคลาดเคลื่อนในการพยากรณ์ลดลง

- **ระยะเวลาในการเก็บข้อมูลซ้ำ** (t)

เมื่อระยะเวลาในการเก็บข้อมูลซ้ำเพิ่มสูงขึ้น ส่งผลให้ค่าเฉลี่ยร้อยละความคลาดเคลื่อนสัมบูรณ์มีแนวโน้มลดลง เพราะระยะเวลาในการเก็บข้อมูลซ้ำเพิ่มขึ้นส่งผลให้ค่าความคลาดเคลื่อนในการพยากรณ์ลดลง

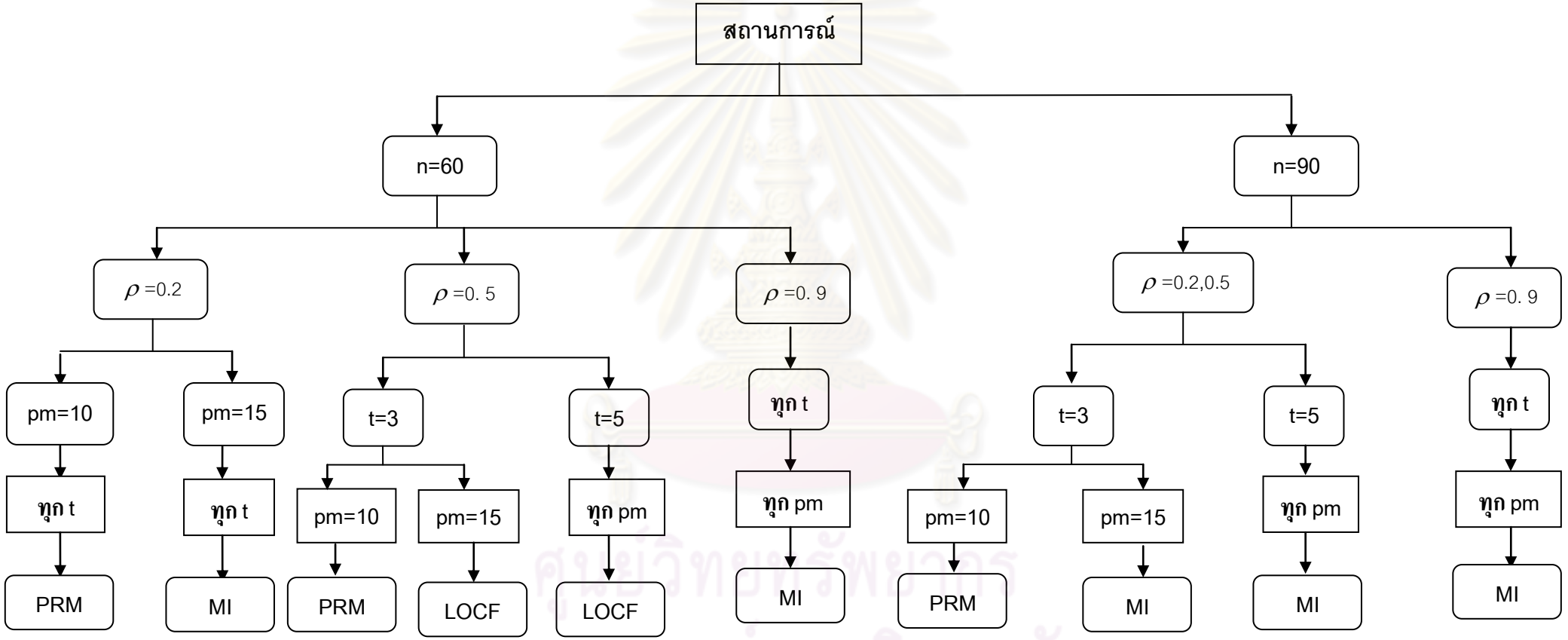
- **สัมประสิทธิ์สหสัมพันธ์** (ρ)

เมื่อสัมประสิทธิ์สหสัมพันธ์เพิ่มสูงขึ้น จะส่งผลให้ค่าเฉลี่ยร้อยละความคลาดเคลื่อนสัมบูรณ์มีแนวโน้มเพิ่มขึ้น นั่นคือค่าเฉลี่ยร้อยละความคลาดเคลื่อนสัมบูรณ์แปรผันตามสัมประสิทธิ์สหสัมพันธ์ เพราะว่าข้อมูลที่ใช้ในการประมาณมีอัตตสหสัมพันธ์ต่อกัน ความผิดพลาดก็จะเกิดขึ้น ยิ่งข้อมูลสัมพันธ์กันมากขึ้นก็จะยิ่งทำให้เกิดการผิดพลาดในการประมาณค่ามากขึ้นตามไปด้วย

- **ร้อยละการสูญหายของตัวแปรตาม**

ที่ทุกขนาดตัวอย่าง ทุกระยะเวลาในการเก็บข้อมูลซ้ำ ทุกระดับของอัตตสหสัมพันธ์ เมื่อร้อยละการสูญหายเพิ่มสูงขึ้นจะทำให้ MAPE มีค่าไม่คงที่ นั่นคือไม่แปรผันตามหรือแปรผกผันกับปัจจัยใด

รูปที่ 5.1 แสดงการเลือกใช้วิธีการประมาณค่าสูญหายสำหรับข้อมูลระยะยาวในตัวแบบ Generalized Estimating Equations



5.2 ข้อเสนอแนะ

5.2.1 ด้านการนำไปใช้

5.2.1.1 ควรตรวจสอบลักษณะข้อมูลระยะยาวว่ามีลักษณะ และรูปแบบอย่างไร ตรงตามข้อตกลงเบื้องต้นหรือไม่ เพื่อที่จะนำไปพิจารณาได้อย่างถูกต้องเหมาะสม

5.2.1.2 จากผลการวิจัย พบว่าวิธีการประมาณค่าสูญหายด้วยวิธี Last Observation Carried Forward จะเป็นวิธีการประมาณค่าสูญหายที่ดีเมื่อข้อมูลระยะยาวมีอัตตสหสัมพันธ์ระดับปานกลาง และมีขนาดตัวอย่างไม่มากนัก ซึ่งพบว่าเมื่อระยะเวลาในการเก็บข้อมูลซ้ำเพิ่มมากขึ้น วิธีนี้จะประมาณได้ดีขึ้นด้วย

5.2.1.3 วิธี Previous Row Mean จะเป็นวิธีการประมาณค่าสูญหายที่ดี เมื่อข้อมูลมีขนาดตัวอย่างไม่มาก ระยะเวลาในการเก็บข้อมูลซ้ำน้อย ร้อยละการสูญหายไม่สูงมากนัก อัตตสหสัมพันธ์ระดับต่ำและปานกลาง

5.2.1.4 วิธี Multiple Imputation (MI) จะเป็นวิธีการประมาณค่าสูญหายที่ดี เมื่อข้อมูลมีขนาดตัวอย่างสูงขึ้น ทุกระยะเวลาในการเก็บข้อมูลซ้ำ ทุกร้อยละการสูญหาย และทุกระดับอัตตสหสัมพันธ์

5.2.1.5 ในการวิจัยครั้งนี้จะเห็นว่า โดยส่วนใหญ่แล้ววิธี Multiple Imputation จะเป็นวิธีประมาณค่าสูญหายที่ดี แต่จะเห็นว่าในบางกรณี การประมาณค่าวิธี Last Observation Carried Forward และวิธี Previous Row Mean จะให้ผลไม่แตกต่างกับวิธี Multiple Imputation มากนัก ดังนั้นในทางปฏิบัติสามารถที่จะใช้วิธี Last Observation Carried Forward หรือวิธี Previous Row Mean นี้แทนได้ เพราะว่าจะสะดวกและง่ายต่อการทำ ซึ่งเราไม่จำเป็นต้องทราบฟังก์ชั่นการแจกแจงของข้อมูลก็สามารถหาค่าประมาณได้

5.2.2 ด้านการศึกษาวิจัย

5.2.2.1 สำหรับงานวิจัยครั้งนี้ผู้วิจัยได้ศึกษาเฉพาะกรณีตัวแปรอิสระมีการแจกแจงปกติ ที่ตัวแปรอิสระบางตัวมีอัตตสหสัมพันธ์กันตามเวลา และตัวแปรอิสระทุกตัวไม่สัมพันธ์กัน สำหรับงานวิจัยครั้งต่อไปอาจทำการศึกษากรณีที่ตัวแปรอิสระมีการแจกแจงแบบอื่น ๆ หรือกรณีที่ตัวแปรอิสระสัมพันธ์กัน

5.2.2.2 สำหรับงานวิจัยครั้งนี้ผู้วิจัยได้ศึกษาเฉพาะข้อมูลระยะยาวที่เกิด อัตตสหสัมพันธ์ในรูปแบบอัตตสหสัมพันธ์อันดับที่หนึ่ง (AR(1)) สำหรับงานวิจัยครั้งต่อไปอาจทำการศึกษากรณีที่ข้อมูลระยะยาวเกิดอัตตสหสัมพันธ์ในรูปแบบอื่น ๆ

5.2.2.3 สำหรับงานวิจัยครั้งนี้ผู้วิจัยได้ศึกษาเฉพาะข้อมูลระยะยาวที่นำมาวิเคราะห์ในตัวแบบ Generalized Estimating Equations สำหรับงานวิจัยครั้งต่อไปอาจทำการศึกษา กรณีที่นำไปวิเคราะห์ในตัวแบบอื่นๆ ต่อไป

5.2.2.4 สำหรับงานวิจัยครั้งนี้ผู้วิจัยได้ศึกษาเฉพาะข้อมูลที่มีระยะเวลาในการเก็บซ้ำที่ 3 และ 5 คาบเวลา สำหรับงานวิจัยครั้งต่อไปอาจทำการศึกษากรณีที่ระยะเวลาในการเก็บซ้ำมากขึ้นกว่าเดิม

5.2.2.5 สำหรับงานวิจัยครั้งนี้ผู้วิจัยได้ทำการศึกษาเมื่อการสูญหายเกิดขึ้นอย่างสุ่ม โดยการไม่มีรูปแบบที่แน่นอน ซึ่งการสูญหายบางครั้งอาจจะเกิดการสูญหายในหน่วยตัวอย่างเดียวกัน ที่คนละคาบเวลา หรือเกิดการสูญหายที่คนละหน่วยตัวอย่าง คาบเวลาเดียวกัน หรืออาจจะเกิดการสูญหายจากคนละหน่วยตัวอย่าง ที่คนละคาบเวลาก็ได้ ดังนั้น ในงานวิจัยครั้งต่อไปควรทำการศึกษาเมื่อการสูญหายเกิดขึ้นอย่างเป็นรูปแบบ

รายการอ้างอิง

ภาษาไทย

พัชรีภรณ์ พรหมหมัด. การประมาณค่าพารามิเตอร์ในสมการถดถอยเชิงเส้นอย่างง่ายของข้อมูลระยะยาว. วิทยานิพนธ์ปริญญามหาบัณฑิต ภาควิชาสถิติ คณะพาณิชยศาสตร์และการบัญชี จุฬาลงกรณ์มหาวิทยาลัย, 2547.

ศิริญญา ธีระอนันต์ชัย และ ลี่ลี อิงศรีสว่าง. ตัวแบบเชิงเส้นวางนัยทั่วไปสำหรับการศึกษาติดตามระยะยาวของจำนวนการเรียกค่าสินไหมทดแทนการประกันภัยรถยนต์ในกรุงเทพมหานคร. วารสารวิทยาศาสตร์ มศว 25 : 31-46.

ศุภลักษณ์ กรรณิกา. การเปรียบเทียบวิธีการประมาณค่าสูญหายในการวางแผนการทดลองแบบจัตุรัสละติน. วิทยานิพนธ์ปริญญามหาบัณฑิต ภาควิชาสถิติ คณะพาณิชยศาสตร์และการบัญชี จุฬาลงกรณ์มหาวิทยาลัย, 2549.

ภาษาอังกฤษ

Jos W.R. Twisk. Applied Longitudinal Data Analysis for Epidemiology. New York : Cambridge University Press, 2003.

Jean Mundahl Engels, Paula Diehr. Imputation of missing longitudinal data : a comparison of methods. Journal of Clinical Epedemiology 56(2003) : 968-976.

A. Plaia, and A.L. Bondi. Single imputation method of missing values in environmental pollution data sets. Journal of Atmospheric Environment 40(2006) : 7316-7330.

Donald Hedeker, Robert D. Gibbons. LONGITUDINAL DATA ANALYSIS. New York : John Wiley and Sons.

Peter J. Diggle, Patrick J. Heagerty, Kung-Yee Liang, Scott L. Zeger. Analysis of Longitudinal Data. Second edition. New York : Oxford University Press, 2002.

ภาคผนวก

ภาคผนวก

โปรแกรมที่ใช้ในการวิจัย

ตัวอย่างโปรแกรม

```
n=60
t=3
rho=0.2
mean=20
sd=3
beta0=0.5
beta1=0.5
beta2=0.5
beta3=0.5
beta4=0.5
Mis=0.10

h=1
s=1
l=1

##Parameter##

beta<-rbind(beta0,beta1,beta2,beta3,beta4)

##Gen x##

x0<-array(,dim=c(n*t,1))
  for(i in 1:(n*t)){
          x0[i,]=1}
a<-array(,dim=c(t,1))
```

```
      for(i in 1:t){
          a[i,]=i}
b<-rep(a,n)
x1<-array(b,dim=c(n*t,1))
x2<-array(rnorm(n*t,mean,sd),dim=c(n*t,1))
x3<-array(rnorm(n*t,mean,sd),dim=c(n*t,1))
ex<-array(rnorm(n*t,mean,sd),dim=c(n*t,1))
x4<-array(,dim=c(n*t,1))
m=0
for(i in 1:n)
  for(j in 1:t){
    m=m+1
          if(j==1)
            x4[m]=ex[m]
          else
            x4[m]=(rho*x4[m-1])+ex[m] }
x.matrix<-cbind(x0,x1,x2,x3,x4)

##Gen Error##

u<-array(rnorm(n*t,0,sd),dim=c(n*t,1))
e<-array(,dim=c(n*t,1))
m1=0
for(i in 1:n)
  for(j in 1:t){
    m1=m1+1
          if(j==1)
            e[m1]=u[m1]
          else
            e[m1]=(rho*e[m1-1])+u[m1]
  }
```

```
##y real##

y<-array(,dim=c(n*t,1))
y<-((x.matrix%*%beta)+e)

##Missing of y##

yMis<-y
g<-Mis*n*t
rnd<-array(0,dim=c(g,1))

for(i in 1:g){

    repeat{
        p<-runif(1,1,n*t)
        q<-trunc(p)
        r<-q%%t
        if(r!=1){
         w=0
          for(j in 1:g){
                if(q==rnd[j]){
                w=1
                break}
      }
          if(w!=1){
              rnd[i]=q
                break}
      }
    }
}
```

```
for(i in 1:g){
   f<-rnd[i]
   yMis[f]=0}

library(car)
yMiss<-recode(yMis,"0 = NA")

##LOCF##

y_1<-yMis
for(i in 1:g){
index = rnd[i]-1
repeat
{
        aa <- y_1[index]
         if(aa!=0){
           y_1[rnd[i]]=y_1[index]
           break}
         if(aa==0){
           index=index-1}
}
}

##Quasi-Likelihood##
##Locf##

library(gee)
ID <- rep(1:n,each=t)
Data1<-data.frame(x1,x2,x3,x4,y_1)
Nf_1<-gee(y_1 ~ x1+x2+x3+x4, id = ID, data = Data1, family = gaussian, corstr = "AR-
M", Mv=1)
```

```
Para_1<-array(coefficients(Nf_1),dim=c(5,1))

y.estimate1<-array(,dim=c(n*t,1))
y.estimate1<-(x.matrix%*%Para_1)+e

MAPE_1<-array(,dim=c(1,1))
dis_1<-abs((y-y.estimate1)/y)
MAPE_1[h]<-(sum(dis_1)/(n*t))*100

##Previous row mean##

y_2<-yMis
bb<-array(,dim=c(g,1))
bb<-sort(rnd)
for(i in 1:g){
    bbb = bb[i]
      if(y_2[bbb]==0){
        pos = bbb%%t
        if(pos==2){
          y_2[bbb]<- y_2[bbb-1]}
        else if(pos==0){
            y_2[bbb]<-(sum(y_2[bbb-2],y_2[bbb-1]))/2}
      }
}

##Quasi-Likelihood##
##Previous row mean##

Data2<-data.frame(x1,x2,x3,x4,y_2)
Nf_2<-gee(y_2 ~ x1+x2+x3+x4, id = ID, data = Data2, family = gaussian, corstr = "AR-
M", Mv=1)
```

```
Para_2<-array(coefficients(Nf_2),dim=c(5,1))

y.estimate2<-array(,dim=c(n*t,1))
y.estimate2<-(x.matrix%*%Para_2)+e

MAPE_2<-array(,dim=c(1,1))
dis_2<-abs((y-y.estimate2)/y)
MAPE_2[s]<-(sum(dis_2)/(n*t))*100

##MI##

y_3<-yMis
library(mice)
cri<-data.frame(x1,yMiss)
imp<-mice(cri,m=3,maxit=1)
impute.mi<-array(,dim=c(n*t,2))
impute.mi<-complete(imp)
y_3[bb]<-impute.mi[bb,2]

##Quasi-Likelihood##
##MI##

Data3<-data.frame(x1,x2,x3,x4,y_3)
Nf_3<-gee(y_3 ~ x1+x2+x3+x4, id = ID, data = Data3, family = gaussian, corstr = "AR-
M", Mv=1)
Para_3<-array(coefficients(Nf_3),dim=c(5,1))

y.estimate3<-array(,dim=c(n*t,1))
y.estimate3<-(x.matrix%*%Para_3)+e
MAPE_3<-array(,dim=c(1,1))
dis_3<-abs((y-y.estimate3)/y)
```

```
MAPE_3[l]<-(sum(dis_3)/(n*t))*100

checkMAPE1 = 0
checkMAPE2 = 0
checkMAPE3 = 0
numCheckMAPE = 0
finish1 = 0
finish2 = 0
finish3 = 0

repeat{

##Gen x##

x0<-array(,dim=c(n*t,1))
  for(i in 1:(n*t)){
          x0[i,]=1}

a<-array(,dim=c(t,1))
      for(i in 1:t){
          a[i,]=i}
b<-rep(a,n)
x1<-array(b,dim=c(n*t,1))
x2<-array(rnorm(n*t,mean,sd),dim=c(n*t,1))
x3<-array(rnorm(n*t,mean,sd),dim=c(n*t,1))
ex<-array(rnorm(n*t,mean,sd),dim=c(n*t,1))
x4<-array(,dim=c(n*t,1))
m=0

for(i in 1:n)
  for(j in 1:t){
```

```
   m=m+1
          if(j==1)
            x4[m]=ex[m]
          else
            x4[m]=(rho*x4[m-1])+ex[m] }
x.matrix<-cbind(x0,x1,x2,x3,x4)

##Gen Error##

u<-array(rnorm(n*t,0,sd),dim=c(n*t,1))
e<-array(,dim=c(n*t,1))
m1=0
for(i in 1:n)
 for(j in 1:t){
   m1=m1+1
          if(j==1)
            e[m1]=u[m1]
          else
            e[m1]=(rho*e[m1-1])+u[m1]
 }

##y real##

y<-array(,dim=c(n*t,1))
y<-((x.matrix%*%beta)+e)

##Missing of y##

yMis<-y
for(i in 1:g){
   f<-rnd[i]
```

```
   yMis[f]=0}

library(car)
yMiss<-recode(yMis,"0 = NA")

##LOCF##

if(checkMAPE1 == 1 & finish1 == 0){
          numCheckMAPE = numCheckMAPE+1
          finish1 = 1
}

else{
y_1<-yMis
for(i in 1:g){
index = rnd[i]-1
repeat
{
        aa <- y_1[index]
         if(aa!=0){
           y_1[rnd[i]] = y_1[index]
           break}
         if(aa==0){
           index=index-1}
}
}

h=h+1
library(gee)
ID <- rep(1:n,each=t)
Data1<-data.frame(x1,x2,x3,x4,y_1)
```

```
Nf_1<-gee(y_1 ~ x1+x2+x3+x4, id = ID, data = Data1, family = gaussian, corstr = "AR-
M", Mv=1)
Para_1<-array(coefficients(Nf_1),dim=c(5,1))

y.estimate1<-(x.matrix%*%Para_1)+e

dis_1<-abs((y-y.estimate1)/y)
MAPE_1[h]<-(sum(dis_1)/(n*t))*100
diffMAPE1<-abs(MAPE_1[h-1]-MAPE_1[h])

          if(diffMAPE1<0.001){
            checkMAPE1 = 1}

}  ##End loop else

##End LoCF##

##Prev##

if(checkMAPE2 == 1 & finish2 == 0){
          numCheckMAPE = numCheckMAPE+1
          finish2 = 1
}

else{
y_2<-yMis
bb<-array(,dim=c(g,1))
bb<-sort(rnd)
for(i in 1:g){
    bbb = bb[i]
      if(y_2[bbb]==0){
```

```
        pos = bbb%%t
        if(pos==2){
          y_2[bbb]<- y_2[bbb-1]}
        else if(pos==0){
            y_2[bbb]<-(sum(y_2[bbb-2],y_2[bbb-1]))/2}
      }
}

s=s+1
Data2<-data.frame(x1,x2,x3,x4,y_2)
Nf_2<-gee(y_2 ~ x1+x2+x3+x4, id = ID, data = Data2, family = gaussian, corstr = "AR-
M", Mv=1)
Para_2<-array(coefficients(Nf_2),dim=c(5,1))

y.estimate2<-(x.matrix%*%Para_2)+e

dis_2<-abs((y-y.estimate2)/y)
MAPE_2[s]<-(sum(dis_2)/(n*t))*100
diffMAPE2 = abs(MAPE_2[s-1]-MAPE_2[s])

          if(diffMAPE2<0.001){
            checkMAPE2 = 1}

}  ##End loop else

##END Prev##

##MI##

if(checkMAPE3 == 1 & finish3 == 0){
          numCheckMAPE = numCheckMAPE+1
```

```
        finish3 = 1
}

else{
y_3<-yMis
library(mice)
cri<-data.frame(x1,yMiss)
imp<-mice(cri,m=3,maxit=1)
impute.mi<-array(,dim=c(n*t,2))
impute.mi<-complete(imp)
y_3[bb]<-impute.mi[bb,2]

l=l+1
Data3<-data.frame(x1,x2,x3,x4,y_3)
Nf_3<-gee(y_3 ~ x1+x2+x3+x4, id = ID, data = Data3, family = gaussian, corstr = "AR-
M", Mv=1)
Para_3<-array(coefficients(Nf_3),dim=c(5,1))

y.estimate3<-(x.matrix%*%Para_3)+e

dis_3<-abs((y-y.estimate3)/y)
MAPE_3[l]<-(sum(dis_3)/(n*t))*100
diffMAPE3 = abs(MAPE_3[l-1]-MAPE_3[l])

        if(diffMAPE3<0.001){
          checkMAPE3 = 1}

}  ##End loop else

##END MI##
```

```
if(numCheckMAPE == 3)
break

} ## End Repeat

MAPE.LOCF<-(sum(MAPE_1[1]:MAPE_1[h])/h)
MAPE.Prev<-(sum(MAPE_2[1]:MAPE_2[s])/s)
MAPE.MI<-(sum(MAPE_3[1]:MAPE_3[l])/l)

ComPMAPE<-array(0,dim=c(3,1))

ComPMAPE[1]<-MAPE.LOCF
ComPMAPE[2]<-MAPE.Prev
ComPMAPE[3]<-MAPE.MI
print(ComPMAPE)
```

ประวัติผู้เขียนวิทยานิพนธ์

นางสาวนฤมล คุ้มปิยะผล เกิดวันที่ 26 พฤษภาคม พ.ศ. 2529 ที่จังหวัดตรัง สำเร็จการศึกษาปริญญาตรีวิทยาศาสตรบัณฑิต สาขาวิชาคณิตศาสตร์ มหาวิทยาลัยธรรมศาสตร์ ปีการศึกษา 2550 และเข้าศึกษาต่อในหลักสูตรสถิติศาสตรมหาบัณฑิตที่จุฬาลงกรณ์ มหาวิทยาลัย เมื่อ พ.ศ. 2551