การเลือกลักษณะสำหรับการแทนค่าข้อมูลสูญหายในการวัดประสิทธิภาพการผลิตปลานิลในจังหวัดสุพรรณบุรี

Feature Selection for Imputation of Missing Data to Measure the Efficiency of Nile Tilapia Production in Suphanburi Province

จารินี ศานติจรรยาพร [1*] สุชาดา กรเพชรปาณี [2] พัชรี วงษ์เกษม [3]

Jarinee Santijanyaporn [1*] Suchada Kornpetpanee [2] Patcharee Wongkasem [3]

[1] Faculty of Science and Technology, Suan Dusit University, Thailand

[2] College of Research Methodology and Cognitive Science, Burapha University, Thailand

[3] Faculty of Science, Burapha University, Thailand

บทคัดย่อ

การวิจัยนี้มีวัตถุประสงค์เพื่อ 1) พัฒนาวิธีการแทนค่าข้อมูลสูญหายแบบใหม่ (FSNNR) โดยการรวมวิธีการเลือกลักษณะ (Feature selection) กับการแทนค่าข้อมูลสูญหายด้วยวิธี Nearest Neighbor Regression Imputation 2) เปรียบเทียบประสิทธิภาพของวิธีการแทนค่าข้อมูลสูญหายแบบใหม่กับวิธีการแทนค่าข้อมูลสูญหายแบบเดิม 3 วิธี (RI, KNN และ NNR) ภายใต้ 36 สถานการณ์ จาก 3 เงื่อนไข ได้แก่ ขนาดตัวอย่าง ร้อยละของการสูญหายของข้อมูล และขนาดของส่วนเบี่ยงเบนมาตรฐานของความคลาดเคลื่อนของข้อมูล โดยใช้การจำลองสถานการณ์ด้วยวิธีมอนติคาร์โลทดลองซ้ำเป็นจำนวน 1,000 ครั้ง ในแต่ละสถานการณ์ และ 3) เพื่อนำวิธีการแทนค่าข้อมูลสูญหายที่พัฒนาขึ้นไปใช้แทนค่าในแบบจำลองการวัดประสิทธิภาพการผลิตปลานิลของจังหวัดสุพรรณบุรี ผลการศึกษาปรากฏว่า

1) วิธีการแทนค่าข้อมูลสูญหายแบบใหม่ (FSNNR) มีขั้นตอนดังนี้

ขั้นตอนที่ 1 เลือกลักษณะของข้อมูล 2 ลักษณะ ด้วยวิธี Nearest Neighbor โดยใช้สูตร Euclidean distance ขั้นตอนที่ 2 แทนค่าข้อมูลสูญหายด้วยวิธี K-Nearest Neighbor Imputation โดยกำหนดค่า k = 2 และขั้นตอนที่ 3 นำข้อมูลสมบูรณ์ที่ได้จากขั้นตอนที่ 2 มาแทนค่าข้อมูลสูญหายด้วยวิธี Regression Imputation

2) ประสิทธิภาพของการแทนค่าข้อมูลสูญหายแบบใหม่ ดีกว่าการแทนค่าข้อมูลสูญหายแบบเดิม จำนวน 33 สถานการณ์

3) ประสิทธิภาพทางเทคนิคของการผลิตปลานิลของเกษตรกรแต่ละราย เมื่อแทนค่าข้อมูลสูญหายด้วยวิธี FSNNR ปรากฏว่า ส่วนใหญ่ของเกษตรกรผู้เลี้ยงปลานิลเป็นกลุ่มที่มีค่าประสิทธิภาพทางเทคนิคระดับมากที่สุด ร้อยละ 82.76 และมีค่าประสิทธิภาพทางเทคนิคระดับมาก ร้อยละ 24.14

คำสำคัญ: การเลือกลักษณะ, การแทนค่าข้อมูลสูญหาย, ข้อมูลสูญหาย, ประสิทธิภาพการผลิต

**Corresponding author. E-mail: jarinee.s@gmail.com*

ABSTRACT

The objectives of this research were: 1) to development of a new method (FSNNR) for imputation of missing data that combined feature selection with the nearest neighbor regression imputation method, 2) to compare the efficiency of FSNNR and three imputations (RI, KNN and NNR) under 36 situations from these three conditions; the sample sizes; the missing percentages; and the data deviations. The data were simulated using the Monte Carlo technique; repeated 1,000 times for each situation, and 3) to measure the efficiency of Nile tilapia production in Suphanburi Province. The results were as followed:

1) The new method for imputation of missing data was as follows:

First, we select two features of data using the Nearest Neighbor. Next, we impute the missing value using the K-Nearest Neighbor Imputation (k=2). Finally, complete the data obtained from step 2 to impute missing data with Regression Imputation.

2) The FSNNR was better performed than the other imputations under 33 combinations of simulated conditions.

3) The technical efficiency of each Nile Tilapia farmer when replacing missing data with FSNNR method, it was found that the Nile Tilapia farmers have the highest level of technical efficiency 82.76% and high technical efficiency of 24.14%.

Keywords: feature selection, imputation, missing data, efficiency of production

ความนำ

จังหวัดสุพรรณบุรีได้กำหนดประเด็นยุทธศาสตร์เกี่ยวกับการเพิ่มขีดความสามารถด้านเกษตรเชื่อมโยงสู่เกษตรอุตสาหกรรมและพาณิชยกรรมเพื่อการบริโภคและการส่งออก โดยมีเป้าประสงค์เพิ่มมูลค่าผลิตภัณฑ์สินค้าเกษตร อุตสาหกรรมและเกษตรอุตสาหกรรม เพื่อการบริโภคและเพิ่มมูลค่าทางเศรษฐกิจจังหวัด จังหวัดสุพรรณบุรีเป็นจังหวัดที่มีการเลี้ยงปลานิลที่มีผลผลิตต่อฟาร์มค่อนข้างสูง หากต้องการเพิ่มขีดความสามารถในการผลิตปลานิล ก็ต้องเพิ่มประสิทธิภาพทางเทคนิคและผลิตภาพการผลิตสินค้าเกษตร ดังนั้น ก่อนที่จะวางแผนเพิ่มประสิทธิภาพทางเทคนิคและผลิตภาพการผลิตด้านการใช้ปัจจัยการผลิต ควรทราบว่า ประสิทธิภาพทางเทคนิคและผลิตภาพด้านการใช้ปัจจัยการผลิตของการผลิตปลานิลอยู่ในระดับใด การวัดประสิทธิภาพเชิงเทคนิคมีวิธีการวัดมากมาย แต่ที่ได้รับความนิยมมาก คือ การวัดประสิทธิภาพตามแนวคิดของ Farrell (1957) เป็นลักษณะการวัดประสิทธิภาพเชิงเปรียบเทียบ (Relative efficiency) โดยการประมาณค่าสมการพรมแดนหรือประมาณค่าเส้นพรมแดน (Frontier equation) แล้วพิจารณาดูระยะห่างระหว่าง ณ จุดที่กำลังพิจารณาอยู่นั้นกับเส้นพรมแดน จากการศึกษาการประมาณค่าเส้นสมการพรมแดน สามารถแบ่งวิธีการประมาณค่าเส้นพรมแดนออกเป็น 2 ประเภท ได้แก่ Data Envelopment Analysis (DEA) เป็นวิธีการคำนวณที่ใช้หลักการคณิตศาสตร์ที่เรียกว่า Linear Programming โดยแบบจำลองที่นำเสนอเป็นการพิจารณาทางด้านผลผลิต (Output orientation) ทั้งในด้านปัจจัยนำเข้าและด้านผลผลิต และ Stochastic Frontier Analysis (SFA) เป็นการคำนวณที่ใช้หลักการทางเศรษฐมิติที่ใช้วิธีการประมาณค่าพารามิเตอร์ (Aigner, Lovell, & Schmidt, 1977) โดยแบบจำลอง Stochastic Frontier กำหนดให้ความไม่มีประสิทธิภาพเชิงเทคนิค

(Technical inefficiency) เป็นส่วนประกอบหนึ่งของค่าความคลาดเคลื่อนสำหรับข้อมูลที่นำมาใช้ในการวิเคราะห์ประสิทธิภาพนั้น สามารถใช้ได้ทั้งข้อมูลปฐมภูมิและข้อมูลทุติยภูมิ ในการเก็บรวบรวมข้อมูลจากกลุ่มตัวอย่างที่เป็นผู้ประกอบการ ปัญหาที่พบในงานวิจัย คือ มีข้อมูลสูญหาย (Missing data) จำนวนมาก โดยผู้ตอบแบบสอบถามได้ละไว้ ไม่ตอบ ทำให้นักวิจัยต้องละทิ้งแบบสอบถามดังกล่าว การตัดข้อมูลทิ้งไปเป็นวิธีทั่วไปของโปรแกรมสำเร็จรูปทางสถิติ จากการทดลองปรากฏว่า ถ้าตัวแปรแต่ละตัวมีข้อมูลหายไปโดยสุ่มเพียง 10% จะมีผลให้ต้องตัดหน่วยวิเคราะห์ทิ้งถึง 59% (Ibrahim & Molenberghs, 2009) ถือว่าเป็นความสูญเสียในอัตราที่สูงมาก การละทิ้งแบบสอบถามไปเป็นสิ่งที่ทำได้ ถ้าข้อมูลหามาได้โดยง่าย ต้นทุนไม่มาก ไม่สิ้นเปลืองมาก สามารถหาข้อมูลสำรองเอาไว้มาก ๆ เผื่อการถูกตัดทิ้ง แต่โดยข้อเท็จจริงแล้วการได้มาซึ่งแบบสอบถามในแต่ละชุด ต้องเสียเวลา ค่าใช้จ่าย แรงงาน เงินและทรัพยากรอื่น ๆ เป็นจำนวนมาก เช่น การสำรวจโดยการสัมภาษณ์ผู้ประกอบการ ผู้สัมภาษณ์ต้องนัดหมายเพื่อเข้าพบผู้ประกอบการ อาจไม่ยอมให้พบหรือต้องนัดหมายผ่านบุคคลอื่น อาจไปแล้วไม่ได้พบ ต้องนัดหมายใหม่ พบแล้วแต่ไม่มีเวลาให้มากนัก ต้องรีบสัมภาษณ์หรือนัดพบใหม่เพื่อสัมภาษณ์ต่อ ในบางครั้งข้อมูลที่เกี่ยวกับสถานะทางการเงินของธุรกิจ ผู้ให้สัมภาษณ์อาจมองว่าเป็นความลับทางธุรกิจจึงไม่ให้ข้อมูล ทั้งนี้ในการวิเคราะห์ประสิทธิภาพต้องมีตัวแปรที่ครบถ้วนสมบูรณ์ และในแต่ละตัวแปรต้องไม่มีค่าข้อมูลสูญหาย (Islam, Tai, & Kusairi, 2016)

จากการศึกษางานวิจัยการแทนค่าข้อมูลสูญหายโดยใช้เทคนิคทางเหมืองข้อมูล โดยส่วนใหญ่วิธีการนี้ได้มาจากการหาความสัมพันธ์ระหว่างกลุ่มข้อมูลที่จะนำมาประมาณค่าที่สูญหาย (Beretta & Santaniello, 2016) เช่น การแทนค่าข้อมูลสูญหายด้วยวิธี Regression Imputation การแทนค่าข้อมูลสูญหายด้วยวิธี K-Nearest Neighbor Imputation เนื่องจากโดยปกติการแทนค่าข้อมูลสูญหายด้วยวิธี K-Nearest Neighbor Imputation เป็นวิธีที่นิยมใช้อย่างแพร่หลาย ง่ายและไม่ซับซ้อนในการใช้งาน สามารถทำงานได้ดีทั้งข้อมูลที่เป็นเชิงเส้นและไม่เป็นเชิงเส้น และยังทำงานได้ดี แม้ว่าจะมีจำนวนตัวอย่างที่น้อย (Troyanskaya et al., 2001) แต่ยังมีปัญหาเกี่ยวกับผลกระทบของค่านอกเกณฑ์ (Outlier) Chaimongkol and Suwattee (2004) ได้แนะนำการแทนค่าข้อมูลสูญหายด้วยวิธี Nearest Neighbor Regression Imputation (NNR) เพื่อแก้ปัญหาดังกล่าว โดยการนำการแทนค่าข้อมูลสูญหายด้วยวิธี Nearest Neighbor Imputation รวมกับการแทนค่าข้อมูลสูญหายด้วยวิธี Regression Imputation เนื่องจากเป็นการแทนค่าข้อมูลสูญหายโดยการคำนวณหาระยะห่างระหว่างจุดด้วยวิธี Euclidean distance ระหว่างข้อมูลกลุ่มที่สมบูรณ์กับกลุ่มที่สูญหาย แล้วแทนค่าข้อมูลสูญหายจนครบทุกชุดข้อมูล หลังจากนั้นนำข้อมูลที่สมบูรณ์ที่ได้มาสร้างสมการถดถอย เพื่อแทนค่าข้อมูลสูญหายอีกครั้งหนึ่ง แต่ถ้ามีลักษณะเด่นบางค่าไม่สัมพันธ์ (ไม่มีความใกล้เคียงกัน) กับค่าข้อมูลสูญหายถูกใช้ในการคำนวณ ผลลัพธ์ที่ได้จะถูกโน้มเอียงด้วยลักษณะที่ไม่เหมาะสม จึงส่งผลให้การประมาณค่าข้อมูลสูญหายที่ได้มีประสิทธิภาพลดลง

ในที่นี้ผู้วิจัยต้องการปรับปรุงประสิทธิภาพของการแทนค่าข้อมูลสูญหายด้วยวิธี Nearest Neighbor Regression Imputation โดยการแก้ปัญหาของข้อมูลที่ได้มาอาจถูกโน้มเอียงด้วยลักษณะเด่นที่ไม่เหมาะสม ซึ่งจะทำให้ได้ข้อมูลที่ไม่ใกล้เคียงกับค่าข้อมูลสูญหายที่ต้องการแทนค่าจริง ๆ ด้วยการนำวิธีเลือกลักษณะด้วยวิธี K-Nearest Neighbor ในการเลือกลักษณะของข้อมูลที่มีความเหมาะสมสำหรับการพยากรณ์ข้อมูลสูญหายก่อน เพื่อลดผลกระทบจากค่านอกเกณฑ์ การผสมผสานระหว่างวิธีการเลือกลักษณะด้วยวิธี K-Nearest Neighbor กับการแทนค่าข้อมูลสูญหายด้วยวิธี Nearest Neighbor Regression Imputation เมื่อแก้ปัญหาเรื่องข้อมูลสูญหายจากการเก็บรวบรวมข้อมูลการผลิตปลานิลในจังหวัดสุพรรณบุรีได้แล้ว ผู้วิจัยจะนำข้อมูลที่สมบูรณ์มาใช้ในการวิเคราะห์ประสิทธิภาพการดำเนินงานด้านการผลิตปลานิลของจังหวัดสุพรรณบุรีเพื่อให้การวัดประสิทธิภาพการผลิตมีความแม่นยำมากขึ้น

วัตถุประสงค์ของการวิจัย

1. เพื่อพัฒนาวิธีการแทนค่าข้อมูลสูญหายใหม่ (FSNNR) โดยการรวมวิธีการเลือกลักษณะ (Feature selection) กับการแทนค่าข้อมูลสูญหายด้วยวิธี Nearest Neighbor Regression Imputation

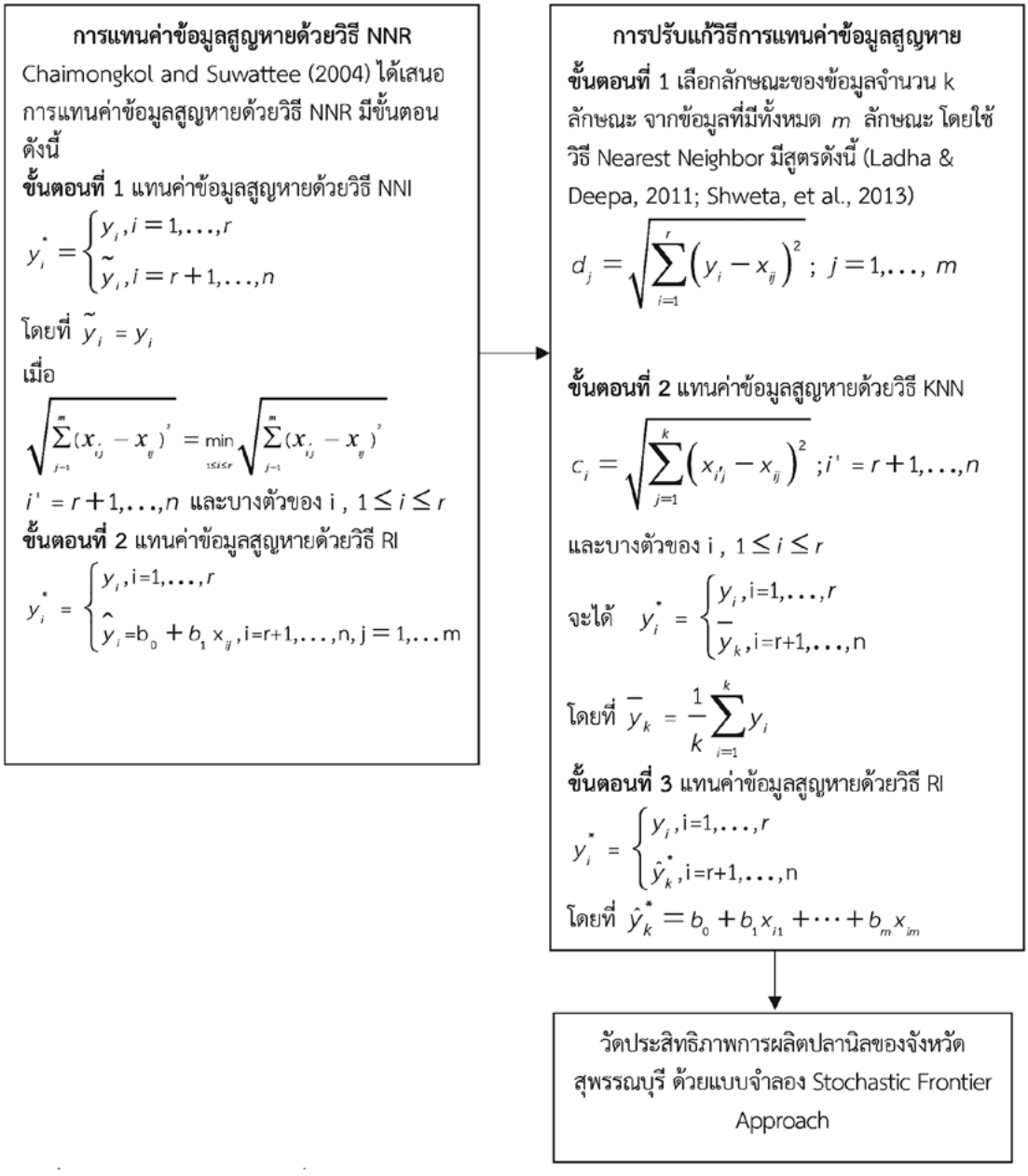
2. เพื่อเปรียบเทียบวิธีการแทนค่าข้อมูลสูญหายที่พัฒนาใหม่ กับการแทนค่าข้อมูลสูญหายด้วยวิธี Regression Imputation (RI) การแทนค่าข้อมูลสูญหายด้วยวิธี K-Nearest Neighbor Imputation (KNN) และการแทนค่าข้อมูลสูญหายด้วยวิธี Nearest Neighbor Regression Imputation (NNR) ภายใต้ 3 เงื่อนไข คือ ขนาดตัวอย่าง 3 สถานการณ์ (50, 100 และ 300) ร้อยละของการสูญหายของข้อมูล 3 สถานการณ์ (5% 10% และ 15%) และขนาดของส่วนเบี่ยงเบนมาตรฐานของความคลาดเคลื่อนของข้อมูล 4 สถานการณ์ (ความคลาดเคลื่อนของข้อมูลที่มีการแจกแจงแบบปกติ โดยมีพารามิเตอร์ μ = 0 และกำหนดให้ σ= 5, 10, 15 และ 20) การจำลองสถานการณ์ใช้วิธีมอนติคาร์โลด้วยโปรแกรม MATLAB ในแต่ละเงื่อนไขทดลองซ้ำเป็นจำนวน 1,000 ครั้ง

3. เพื่อนำวิธีการแทนค่าข้อมูลสูญหายที่พัฒนาขึ้นไปใช้แทนค่าในแบบจำลองการวัดประสิทธิภาพการผลิตปลานิลของจังหวัดสุพรรณบุรี

กรอบแนวคิดการวิจัย

การแทนค่าข้อมูลสูญหายด้วยวิธี K-Nearest Neighbor Imputation เป็นวิธีที่นิยมใช้อย่างแพร่หลาย ง่ายและไม่ซับซ้อนต่อการใช้งาน สามารถทำงานได้ดีทั้งข้อมูลที่เป็นเชิงเส้นและไม่เป็นเชิงเส้น และยังทำงานได้ดีแม้ว่าจะมีจำนวนตัวอย่างที่น้อย แต่ยังมีปัญหาเกี่ยวกับผลกระทบของค่านอกเกณฑ์ Chaimongkol and Suwattee (2004) ได้แนะนำการแทนค่าข้อมูลสูญหายด้วยวิธี Nearest Neighbor Regression Imputation เพื่อแก้ปัญหาดังกล่าว โดยการนำการแทนค่าข้อมูลสูญหายด้วยวิธี Nearest Neighbor Imputation รวมกับวิธี Regression Imputation เนื่องจากเป็นการแทนค่าข้อมูลสูญหายโดยการคำนวณหาระยะห่างระหว่างจุดด้วยวิธี Euclidean distance ระหว่างข้อมูลกลุ่มที่สมบูรณ์กับกลุ่มที่สูญหาย แล้วแทนค่าข้อมูลสูญหายจนครบทุกชุดข้อมูล แล้วนำข้อมูลที่สมบูรณ์ที่ได้มาสร้างสมการถดถอย เพื่อแทนค่าข้อมูลสูญหายอีกครั้งหนึ่ง ผลลัพธ์คือ ค่าที่ได้อาจถูกโน้มเอียงด้วยลักษณะเด่นที่ไม่เหมาะสม ซึ่งจะทำให้ได้ข้อมูลที่มีลักษณะไม่สัมพันธ์กับค่าสูญหายถูกนำมาใช้ในการคำนวณ ผลที่ตามมาคือ ข้อมูลที่ได้มาอาจถูกโน้มเอียงด้วยลักษณะเด่นที่ไม่เหมาะสม ซึ่งจะทำให้ได้ข้อมูลที่ไม่ใกล้เคียงกับค่าสูญหายที่ต้องการแทนค่าจริง ๆ จึงส่งผลให้ประสิทธิภาพของการแทนค่าข้อมูลสูญหายที่ได้ไม่ดีเท่าที่ควร

ดังนั้น วิธีที่เสนอใหม่เป็นการปรับการแทนค่าข้อมูลสูญหายด้วยวิธี Nearest Neighbor Imputation โดยการเพิ่มขั้นตอนการเลือกลักษณะโดยใช้วิธี Nearest Neighbor เนื่องจากเป็นวิธีการที่สามารถลดผลกระทบที่เกิดจากค่านอกเกณฑ์และแก้ปัญหาข้อมูลที่ได้มาอาจถูกโน้มเอียงด้วยลักษณะเด่นที่ไม่เหมาะสม โดยวิธีการนำเสนอนี้มีลักษณะเด่น ดังนี้ 1) สามารถวัดค่าได้และมีความเป็นอิสระในการจำแนกลักษณะ (Ladha & Deepa, 2011; Shweta, Nikita, & Madhvi, 2013) 2) ไม่มีข้อตกลงเบื้องต้นเกี่ยวกับข้อมูล 3) มีผลกระทบกับค่านอกเกณฑ์ในระดับต่ำ 4) มีผลกระทบต่อตัวอย่างที่มีขนาดเล็กในระดับต่ำ และ 5) มีความสามารถในการเรียนรู้เพิ่มขึ้นอย่างมีประสิทธิภาพ (Kumari & Swarnkar, 2011) นอกจากนี้ยังเป็นวิธีการที่มีความซับซ้อนน้อย และใช้เวลาค่อนข้างเร็วสำหรับการประมวลผล มีความยืดหยุ่นสูง และมีการวัดความคล้ายคลึงกันของข้อมูลในระดับตัวอย่าง เพื่อนำวิธีการดังกล่าวมาใช้ในการทดแทนค่าข้อมูลสูญหายของข้อมูลที่ได้จากการเก็บรวบรวมจากเกษตรกรผู้เลี้ยงปลานิล เพื่อใช้ในการวัดประสิทธิภาพการผลิตปลานิลในจังหวัดสุพรรณบุรีได้อย่างมีประสิทธิภาพมากขึ้น



ภาพที่ 1 กรอบแนวคิดการวิจัยเรื่อง การเลือกลักษณะสำหรับการแทนค่าข้อมูลสูญหายในการวัดประสิทธิภาพการผลิตปลานิลในจังหวัดสุพรรณบุรี

สมมติฐานการวิจัย

1. การแทนค่าข้อมูลสูญหายวิธีใหม่ที่พัฒนาขึ้นสามารถให้ค่าเฉลี่ยความคลาดเคลื่อนสมบูรณ์ น้อยกว่าการแทนค่าข้อมูลสูญหายด้วยวิธี Regression Imputation (RI) การแทนค่าข้อมูลสูญหายด้วยวิธี K-Nearest Neighbor Imputation (KNN) และการแทนค่าข้อมูลสูญหายด้วยวิธี Nearest Neighbor Regression Imputation (NNR)

2. ประสิทธิภาพการผลิตปลานิลของเกษตรกรแต่ละรายอยู่ในระดับมากขึ้นไป

ทฤษฎีที่เกี่ยวข้อง

1. การแทนค่าข้อมูลสูญหายด้วยวิธี Regression Imputation (RI)

กำหนดให้ ตัวอย่างมีขนาดเท่ากับ n หน่วย และ y เป็นตัวแปรตามที่ขึ้นอยู่กับตัวแปร $x_1, x_2,..., x_m$

โดยที่ $y_1, y_2,...,y_n$ เป็นค่าของตัวแปร y ที่สนใจศึกษาที่ไม่สูญหาย

$y_{r+1}, y_{r+2},...,y_n$ เป็นค่าของตัวแปร y ที่เป็นข้อมูลสูญหาย

x_{ij}; $i = 1, 2,..., n, j = 1, 2,...,m$ และ เป็นค่าของตัวแปรอิสระที่เก็บรวบรวมข้อมูลได้แบบไม่สูญหายในหน่วยที่ i ของตัวแปร x_j

วิธีการประมาณค่าสูญหายแบบ Regression Imputation (RI) หรือการประมาณสูญหายด้วยวิธีการถดถอยเชิงพหุเชิงเส้น เป็นการประมาณค่าสูญหายโดยการประมาณสมการถดถอยของตัวแปร y โดยใช้ข้อมูลไม่สูญหาย $(x_{ij};y_i)$, $i = 1, 2,..., r$ แล้วประมาณค่าข้อมูลสูญหายของตัวแปร y_i; $i = r, r+1,..., n$ โดยใช้สมการถดถอยที่หาค่าได้ คือ

$$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \hat{\beta}_2 x_{i2} + ... + \hat{\beta}_m x_{im}$$

แทนค่าสูญหาย ด้วยค่าประมาณจากการคำนวณดังนี้

$$y^*_i = \begin{cases} y_i, & i = 1, 2, ..., r \\ \hat{y}_i, & i = r+1, r+2,...,n \end{cases}$$

2. การแทนค่าข้อมูลสูญหายด้วยวิธี K-Nearest Neighbor Imputation (KNN)

วิธี K-Nearest Neighbor Imputation มีขั้นตอนดังนี้

1) กำหนดค่า k

2) คำนวณหาค่าระยะทางที่ใกล้ที่สุด จากระยะทางยุคลิด (Euclidean Distance) ดังนี้

$$d_{ik} = \sqrt{\sum_{j=1}^{m}(x_{ij} - x_{kj})^2} = \min_{1 \le k \le r} \sqrt{\sum_{j=1}^{m}(x_{ij} - x_{kj})^2}$$

สำหรับ $i = r+1, r+2,..., n$ และ $1 \le k \le r$

3) เรียงค่าระยะทางจากน้อยไปหามาก แล้วเลือกระยะทางที่ใกล้ที่สุดมาจำนวน k ตัว

4) หาค่าเฉลี่ยของตัวแปรตาม y_i ที่สอดคล้องกับระยะทางที่ใกล้ที่สุดมาจำนวน k ตัว ในข้อ 3) จะได้ว่า $y^*_i = \bar{y}_i$ โดยที่ $\bar{y}_i$ คือค่าเฉลี่ยของหน่วยตัวอย่างที่ทำให้ค่า d_{ik} ที่น้อยที่สุด จำนวน k ตัว

5) แทนค่าสูญหาย ด้วยค่าประมาณจากการคำนวณดังนี้

$$y^*_i = \begin{cases} y_i, & i = 1, 2, ...,r \\ \bar{y}_i, & i = r+1, r+2,...,n \end{cases}$$

3. การแทนค่าข้อมูลสูญหายด้วยวิธี Nearest Neighbor Regression Imputation (NNR)

ขั้นตอนของวิธีการ NNR มีดังนี้

1) คำนวณหาค่าระยะทางที่ใกล้ที่สุด จากระยะทางยุคลิด (Euclidean distance) ดังนี้

$$d_k = \sqrt{\sum_{j=1}^{m}(x_{ij} - x_{kj})^2} = \min_{1 \le k \le r} \sqrt{\sum_{j=1}^{m}(x_{ij} - x_{kj})^2}$$

สำหรับ $i = r+1, r+2,..., n$ และ $1 \le k \le r$

จะได้ว่า $y^*_i = y_k$ โดยที่ y_k คือค่าของหน่วยตัวอย่างที่ k ที่ทำให้ค่า d_k ที่น้อยที่สุด $(1 \le k \le r)$

2) แทนค่าสูญหาย ด้วยค่าประมาณจากการคำนวณดังนี้

$$y^*_i = \begin{cases} y_i, & i = 1, 2, ...,r \\ y_k, & i = r+1, r+2,...,n \end{cases}$$

3) เมื่อได้ข้อมูลที่สมบูรณ์จากข้อ 2) นำข้อมูลที่ได้มาสร้างสมการถดถอยเชิงเส้นพหุ ดังนี้

$$\hat{y}_i = b_0 + b_1 x_{i1} + b_2 x_{i2} + ... + b_m x_{im}$$

เมื่อ $\hat{y}_i$ เป็น ค่าประมาณของตัวแปร y ของหน่วยตัวอย่างที่ i

$b_1, b_2, ..., b_m$ เป็นค่าประมาณของสัมประสิทธิ์ของการถดถอย

$x_{i1}, x_{i2}, ..., x_{im}$ เป็นค่าของตัวแปร i ที่ไม่สูญหายของหน่วยตัวอย่างที่ i เมื่อ $i = r+1, r+2, ..., n$

4) แทนค่าสูญหาย ด้วยค่าประมาณจากการคำนวณดังนี้

$$y^*_i = \begin{cases} y_i, & i = 1, 2, \ldots, r \\ \hat{y}_i, & i = r+1, r+2, \ldots, n \end{cases}$$

วิธีดำเนินการวิจัย

การดำเนินงานวิจัยแบ่งเป็นขั้นตอน ได้ดังนี้

ขั้นตอนที่ 1 การพัฒนาวิธีการแทนค่าข้อมูลสูญหายด้วยวิธีใหม่ (FSNNR) แบ่งเป็น 3 ขั้นตอน ดังนี้

ขั้นตอนย่อยที่ 1 ศึกษาวรรณกรรมที่เกี่ยวข้องกับการเลือกลักษณะ (Feature selection) และการแทนค่าข้อมูลสูญหาย เช่น RI, KNN และ NNR

ขั้นตอนย่อยที่ 2 พัฒนาวิธีการแทนค่าข้อมูลสูญหายมีรายละเอียดดังนี้

1. เลือกลักษณะของข้อมูล โดยใช้วิธีสมาชิกที่ใกล้ที่สุด (K-Nearest Neighbor: KNN) เพื่อเลือกตัวแทนของลักษณะที่ใกล้ที่สุด สูตรการหาระยะห่างระหว่างคุณลักษณะ ดังนี้

$$d_j = \sqrt{\sum_{i=1}^{r} \left(y_i - x_{ij} \right)^2} \quad \text{โดยที่ } j = 1, \ldots, m$$

เมื่อ y_j เป็นข้อมูลสูญหาย เมื่อ j= r+1, r+2 ,..., n d_j เป็นระยะทางระหว่างตัวแปร y กับลักษณะที่ x_j โดยที่ j = 1,..., m

2. แทนค่าข้อมูลสูญหายด้วยวิธี Nearest Neighbor Regression Imputation (NNR) ดังนี้

2.1 คำนวณระยะห่างระหว่างตัวอย่างที่ประกอบด้วยข้อมูลสูญหายกับตัวอย่างทั้งหมดด้วยสมการดังนี้

$$c_i = \sqrt{\sum_{j=1}^{2} \left(x_{i'j} - x_{ij} \right)^2}$$

เมื่อ c_i เป็นระยะทางระหว่างหน่วยที่ i กับหน่วยที่ i^1 โดยที่ i = 1, 2, ..., r และ i^1 = r+1, r+2 , ... , n

2.2 นำข้อมูลจากตัวอย่างที่มีระยะห่างน้อยที่สุด จำนวน k ค่า มาประมาณค่าข้อมูลสูญหาย โดยใช้สูตร $\bar{y}_k = \frac{1}{k} \sum_{i=1}^{k} y_i$

2.3 ชุดข้อมูลที่ได้มาแทนประมาณค่าสูญหาย $\hat{y}_{r+1}$, ... , $\hat{y}_n$ โดยใช้วิธีการแทนค่าข้อมูลสูญหายด้วยสมการถดถอยพหุคูณ โดยอยู่ในรูปสมการ $y^*_i = b_0 + b_1 x_{i1}' + b_2 x_{i2}'$ เมื่อ i = r+1, r+2 , ... , n

ขั้นตอนที่ 2 การเปรียบเทียบประสิทธิภาพของวิธีการแทนค่าข้อมูลสูญหายวิธีใหม่ (FSNNR) กับการแทนค่าข้อมูลสูญหายด้วยวิธี Regression Imputation (RI) การแทนค่าข้อมูลสูญหายด้วยวิธี K-Nearest Neighbor Imputation (KNN) และการแทนค่าข้อมูลสูญหายด้วยวิธี Nearest Neighbor Regression Imputation (NNR)

ขั้นตอนที่ 3 การวิเคราะห์และวัดประสิทธิภาพการผลิตปลานิลในจังหวัดสุพรรณบุรี โดยใช้ข้อมูล 2 ชุด ดังนี้ 1) ข้อมูลทุติยภูมิของเกษตรกรผู้เลี้ยงปลานิล จากสำนักงานประมงอำเภอบางปลาม้า จังหวัดสุพรรณบุรี ที่มีข้อมูลสูญหายจำนวน 2 ราย จากทั้งหมด 31 ราย และ 2) ข้อมูลทุติยภูมิของเกษตรกรผู้เลี้ยงปลานิล จากสำนักงานประมงอำเภอบางปลาม้า จังหวัดสุพรรณบุรี ที่มีการแทนค่าข้อมูลสูญหายด้วยวิธี FSNNR จำนวน 31 ราย โดยการวิเคราะห์ประสิทธิภาพเชิงเทคนิคการผลิตโดยการประมาณค่าฟังก์ชั่นเส้นพรมแดนการผลิต และความมีประสิทธิภาพทางเทคนิคของการผลิตด้วยแบบจำลอง Stochastic Frontier Approach ในรูปแบบของฟังก์ชั่นการผลิต Cobb-Douglas โดยมีตัวแปรที่ใช้ในแบบจำลอง ดังนี้

1. แบบจำลองในการวิเคราะห์ประสิทธิภาพการผลิตปลานิลในจังหวัดสุพรรณบุรี มีตัวแปรดังนี้

ตัวแปรอิสระ เป็น พื้นที่ในการเลี้ยง (หน่วย: ตารางเมตร), ค่าแรงงาน (หน่วย: บาท/ปี), จำนวนอาหาร (หน่วย: กิโลกรัม/เดือน) และค่าเสื่อมสภาพของเครื่องจักร (หน่วย: บาท/ปี)

ตัวแปรตาม เป็น ผลผลิตปลานิล (หน่วย: กิโลกรัม)

2. แบบจำลองความไม่มีประสิทธิภาพทางเทคนิคของการผลิตปลานิลในจังหวัดสุพรรณบุรี

ในการศึกษาปัจจัยที่มีอิทธิพลต่อความไม่มีประสิทธิภาพทางเทคนิคการผลิตปลานิลในจังหวัด

สุพรรณบุรี (หน่วย: กิโลกรัม) ขึ้นอยู่กับปัจจัยต่อไปนี้ ความหนาแน่นของปลาในบ่อ/กระชัง (หน่วย: ตัว/ตารางเมตร) และระยะเวลาที่เลี้ยง (หน่วย: เดือน)

การวิเคราะห์ข้อมูล

การวิเคราะห์ข้อมูลใช้โปรแกรม Frontier version 4.1 เป็นโปรแกรมฟรีแวร์ในการวิเคราะห์ประสิทธิภาพเชิงเทคนิคการผลิต (Technical efficiency) ในรูปแบบของฟังก์ชั่นการผลิต Cobb-Douglas ตัวแบบอยู่ในรูปของสมการดังนี้

$$y_i = f(x_i, \boldsymbol{\beta}) \times TE_i$$

โดยที่ y_i คือ ผลผลิตของผู้ผลิต i โดยที่ i เท่ากับ 1 ถึง I

x_i คือ เวกเตอร์ของปัจจัยการผลิตจำนวน N ชนิด ที่ใช้โดยผู้ผลิต i

$f(x, \boldsymbol{\beta})$ คือ เส้นพรมแดนการผลิต (Production Frontier)

$\boldsymbol{\beta}$ คือ เวกเตอร์ของพารามิเตอร์ที่ต้องประมาณค่า

และคำนวณหาระดับประสิทธิภาพการผลิตของเกษตรกรแต่ละราย และหาค่าประสิทธิภาพการผลิตเฉลี่ยรวม โดยที่ประสิทธิภาพการผลิตเชิงเทคนิค เท่ากับ

$$TE_i = \frac{y_i}{f(x_i, \boldsymbol{\beta})}$$

ประสิทธิภาพการผลิตเชิงเทคนิคแสดงถึงอัตราส่วนของผลผลิตที่เป็นอยู่กับผลผลิตที่เป็นไปได้สูงสุด (เส้นพรมแดนการผลิต) ถ้า TE_i เท่ากับ 1 แสดงว่า y_i สามารถบรรลุระดับการผลิตที่เป็นไปได้สูงสุด ถ้า TE_i น้อยกว่า 1 แสดงว่า การให้ค่าการวัดของจำนวนผลผลิตที่ขาดเมื่อเทียบกับระดับการผลิตที่เป็นไปได้สูงสุด

ผลการวิจัย

1. ผลการเปรียบเทียบประสิทธิภาพของวิธีการแทนค่าข้อมูลสูญหายที่พัฒนาขึ้น (FSNNR) กับวิธีการอีก 3 แบบ คือ วิธี RI วิธี KNN และวิธี NNR โดยเปรียบเทียบค่าเฉลี่ยความคลาดเคลื่อนสมบูรณ์ (Mean Absolute Error: MAE) แสดงได้ดังตารางที่ 1

ตารางที่ 1 ค่า MAE ของวิธีการแทนค่าข้อมูลสูญหาย จำแนกตามขนาดตัวอย่าง ส่วนเบี่ยงเบนมาตรฐานของความคลาดเคลื่อนของข้อมูล และระดับการสูญหายของข้อมูล

ขนาดตัวอย่าง	ขนาดของส่วนเบี่ยงเบนมาตรฐานของความคลาดเคลื่อนของข้อมูล	ร้อยละของการสูญหายของข้อมูล	ค่า MAE ของวิธีการแทนค่าข้อมูลสูญหาย			
			RI	KNN	NNR	FSNNR
50	5	5%	0.9913	2.6059	2.9301	1.8771
		10%	0.9920	0.8921	1.5430	0.4804
		15%	0.9928	0.8444	1.4518	0.3346
	10	5%	0.9913	1.8554	3.0088	1.1067
		10%	0.9920	0.8921	1.5430	0.6944
		15%	0.9928	0.8444	1.4518	0.3346
	15	5%	1.0026	2.1053	2.6883	1.0189
		10%	0.9929	0.9203	1.5536	0.3988
		15%	0.9936	0.8471	1.4531	0.3450
	20	5%	0.9916	1.8980	2.8880	0.7398
		10%	0.9920	0.8872	1.5632	0.5176
		15%	0.9928	0.8422	1.4360	0.3504

ตารางที่ 1 (ต่อ)

ขนาดตัวอย่าง	ขนาดของส่วนเบี่ยงเบนมาตรฐานของความคลาดเคลื่อนของข้อมูล	ร้อยละของการสูญหายของข้อมูล	ค่า MAE ของวิธีการแทนค่าข้อมูลสูญหาย			
			RI	KNN	NNR	FSNNR
100	5	5%	0.9963	1.0968	1.3837	0.2607
		10%	0.9960	0.9736	1.2130	0.1929
		15%	0.9962	0.9262	1.1728	0.1777
	10	5%	0.9963	1.0968	1.3907	0.2607
		10%	0.9960	0.9736	1.2130	0.1929
		15%	0.9962	0.9262	1.1728	0.1777
	15	5%	0.9952	1.0831	1.3897	0.2721
		10%	0.9960	0.9551	1.2179	0.1821
		15%	0.9962	0.9206	1.1512	0.1801
	20	5%	0.9961	1.1239	1.4575	0.2492
		10%	0.9963	0.9485	1.1835	0.1917
		15%	0.9959	0.9200	1.1748	0.1845
300	5	5%	0.9996	0.8334	1.3098	0.2323
		10%	0.9999	0.8029	1.2936	0.2304
		15%	1.0000	0.7997	1.2844	0.2329
	10	5%	0.9996	0.8332	1.3108	0.2328
		10%	0.9999	0.8035	1.2935	0.2314
		15%	1.0000	0.7993	1.2835	0.2314
	15	5%	0.9997	0.8330	1.3065	0.2433
		10%	0.9997	0.8072	1.2852	0.2317
		15%	1.0000	0.8021	1.2841	0.2314
	20	5%	0.9998	0.8374	1.3119	0.2401
		10%	0.9996	0.8055	1.2882	0.2370
		15%	1.0000	0.7952	1.2892	0.2354

หมายเหตุ: MAE แทน Mean Absolute Error
RI แทน Regression Imputation
KNN แทน K-Nearest Neighbor Imputation
NNR แทน Nearest Neighbor Regression Imputation
FSNNR แทน Feature Selection Nearest Neighbor Regression Imputation

จากตารางที่ 1 ปรากฏว่า ค่า MAE ของการแทนค่าข้อมูลสูญหายด้วยวิธี FSNNR มีค่าต่ำกว่าการแทนค่าข้อมูลสูญหายด้วยวิธี KKN และ NNR ทุกกรณี และค่า MAE ของการแทนค่าข้อมูลสูญหายด้วยวิธี FSNNR มีค่าต่ำกว่าการแทนค่าข้อมูลสูญหายด้วยวิธี RI ทุกกรณี ยกเว้นกรณีที่ขนาดตัวอย่างเท่ากับ 50 ร้อยละของการสูญหายของข้อมูลเท่ากับ 5% และขนาดของส่วนเบี่ยงเบนมาตรฐานของความคลาดเคลื่อนของข้อมูลเท่ากับ 5, 10 และ 15 โดยสรุปวิธี FSNNR มีประสิทธิภาพดีกว่าวิธี RI, KNN และ NNR จำนวน 33 สถานการณ์

2. การวัดประสิทธิภาพเชิงเทคนิคการผลิต โดยการประมาณค่าฟังก์ชั่นเส้นพรมแดนการผลิต และความมีประสิทธิภาพทางเทคนิคของการผลิตด้วยแบบจำลอง Stochastic Frontier Approach ในรูปแบบของฟังก์ชั่นการผลิต Cobb-Douglas โดยมีตัวแปรดังนี้ ปริมาณการผลิตปลานิล (กิโลกรัมต่อปี) (Y_i) พื้นที่ในการเลี้ยง (ตารางเมตร) (x_1) ค่าแรงงาน (บาท/ปี) (x_2) จำนวนอาหาร (กิโลกรัม/เดือน) (x_3) และค่าเสื่อมสภาพของเครื่องจักร (บาท/ปี) (x_4) ในการวิเคราะห์พรมแดนการผลิตเชิงเฟ้นสุ่ม ในการวิจัยนี้เป็นการวิเคราะห์ข้อมูลที่ได้แทนค่าข้อมูล สูญหายด้วยวิธี FSNNR เพื่อให้ได้ข้อมูลที่สมบูรณ์ ก่อนทำการวิเคราะห์ต่อไป

2.1 ผลการประมาณค่าความมีประสิทธิภาพทางเทคนิคการผลิตปลานิล กรณีมีข้อมูลสูญหาย 2 ราย

ผลการประมาณค่าตามแบบจำลอง Production Frontier แบบ Cobb-Douglas ทำให้ได้แบบจำลองฟังก์ชั่นการผลิตดังนี้

$$\ln y = 23.8448 + 0.2636(\ln x_1)^{*} - 0.7339\ (\ln x_2) - 1.1666\ (\ln x_3) + 0.1568\ (\ln x_4)$$

จากผลการประมาณค่าความมีประสิทธิภาพทางเทคนิคของการผลิตปลานิล กรณีมีข้อมูลสูญหาย 2 ราย ปรากฏว่า ปัจจัยที่มีผลต่อผลผลิตปลานิลมีเพียงตัวแปรเดียว คือ พื้นที่ในการเลี้ยง (t-ratio = 1.8094) โดยมีความสัมพันธ์ต่อปริมาณการผลิตปลานิล ที่ระดับนัยสำคัญทางสถิติ .10 ซึ่งมีค่าสัมประสิทธิ์เท่ากับ 0.2636 หมายความว่า ถ้าตัวแปรอื่น ๆ คงที่ และเพิ่มปริมาณพื้นที่ในการเลี้ยงขึ้นร้อยละ 1 จะทำให้ปริมาณผลผลิตเพิ่มขึ้น ร้อยละ 0.2636

ผลการวิเคราะห์ความไม่มีประสิทธิภาพในการผลิตปลานิลในจังหวัดสุพรรณบุรี กรณีมีข้อมูลสูญหาย 2 ราย มีผลดังนี้

$$ui = - 45.7702 + 6.5977\ (\ln z_1)^{*} - 14.2949\ (\ln z_2)$$

จากผลการวิเคราะห์ความไม่มีประสิทธิภาพในการผลิตปลานิลในจังหวัดสุพรรณบุรี กรณีมีข้อมูลสูญหาย 2 ราย ปรากฏว่า ปัจจัยที่ส่งผลต่อความไม่มีประสิทธิภาพทางเทคนิคการผลิตปลานิลในจังหวัดสุพรรณบุรี คือ ระยะเวลาที่เลี้ยง (เดือน) (t-ratio = 1.7673) ระดับนัยสำคัญทางสถิติ .10 ซึ่งมีค่าสัมประสิทธิ์เท่ากับ 6.5977 หมายความว่า ถ้าเกษตรกรมีปัญหาด้านระยะเวลาที่เลี้ยงเพิ่มขึ้น 1 เดือน จะส่งผลทำให้ความไม่มีประสิทธิภาพเพิ่มขึ้นร้อยละ 6.5977

ผลการประมาณค่าประสิทธิภาพทางเทคนิคของการผลิตปลานิลในจังหวัดสุพรรณบุรี กรณีมีข้อมูลสูญหาย 2 ราย ปรากฏว่า ประสิทธิภาพเชิงเทคนิคของการผลิตปลานิลในจังหวัดสุพรรณบุรีเฉลี่ย เท่ากับ 0.4222 เมื่อพิจารณาประสิทธิภาพทางเทคนิคของเกษตรกรแต่ละราย ปรากฏว่า ส่วนใหญ่ของเกษตรกรผู้เลี้ยงปลานิลเป็นกลุ่มที่มีค่าประสิทธิภาพทางเทคนิคระดับปานกลาง (0.4001-0.6000) คิดเป็นร้อยละ 48.28 (ดังตารางที่ 2)

ตารางที่ 2 ประสิทธิภาพเชิงเทคนิคของการผลิตปลานิลในจังหวัดสุพรรณบุรี กรณีมีข้อมูลสูญหาย 2 ราย

ประสิทธิภาพทางเทคนิค	จำนวน (ราย)	ร้อยละ
มากที่สุด (0.8001 – 1.0000)	1	3.45
มาก (0.6001 – 0.8000)	3	10.34
ปานกลาง (0.4001 – 0.6000)	14	48.28
น้อย (0.2001 – 0.4000)	9	31.03
น้อยที่สุด (0.0000 – 0.2000)	2	6.90
ประสิทธิภาพเฉลี่ย เท่ากับ 0.4222		

2.2 ผลการประมาณค่าความมีประสิทธิภาพทางเทคนิคการผลิตปลานิล กรณีแทนค่าข้อมูลสูญหายด้วยวิธี FSNNR

ผลการประมาณค่าตามแบบจำลอง Production Frontier แบบ Cobb-Douglas ทำให้ได้แบบจำลองฟังก์ชั่นการผลิตดังนี้

$\ln y = 24.6477+0.1796(\ln x_1) - 0.1097 (\ln x_2)^* - 0.9421 (\ln x_3) + 0.2429 (\ln x_4)$

จากผลการประมาณค่าความมีประสิทธิภาพทางเทคนิคของการผลิตปลานิล กรณีแทนค่าข้อมูลสูญหายด้วยวิธี FSNNR ปรากฏว่า ปัจจัยที่มีผลต่อผลผลิตปลานิลมีเพียงตัวแปรเดียว คือ ค่าแรงงาน ($\ln x_2$) (t-ratio = -1.9224) โดยมีความสัมพันธ์ต่อปริมาณการผลิตปลานิล ที่ระดับนัยสำคัญทางสถิติ .10 ซึ่งมีค่าสัมประสิทธิ์เท่ากับ -0.1097 หมายความว่า ถ้าตัวแปรอื่น ๆ คงที่ และเพิ่มค่าแรงงานขึ้นร้อยละ 1 จะทำให้ปริมาณผลผลิตลดลงร้อยละ 0.1097

ผลการประมาณค่าประสิทธิภาพทางเทคนิคของการผลิตปลานิลในจังหวัดสุพรรณบุรี ปรากฏว่า ประสิทธิภาพเชิงเทคนิคของการผลิตปลานิลในจังหวัดสุพรรณบุรีเฉลี่ยเท่ากับ 0.8935 เมื่อพิจารณาประสิทธิภาพทางเทคนิคของเกษตรกรแต่ละราย ปรากฏว่า ส่วนใหญ่ของเกษตรกรผู้เลี้ยงปลานิลเป็นกลุ่มที่มีค่าประสิทธิภาพทางเทคนิคระดับมากที่สุด คิดเป็นร้อยละ 82.76 (ดังตารางที่ 3)

ตารางที่ 3 ประสิทธิภาพเชิงเทคนิคของการผลิตปลานิลในจังหวัดสุพรรณบุรี

ประสิทธิภาพทางเทคนิค	จำนวน (ราย)	ร้อยละ
มากที่สุด (0.8001 – 1.0000)	24	82.76
มาก (0.6001 – 0.8000)	7	24.14
ประสิทธิภาพเฉลี่ย เท่ากับ 0.8935		

3. ผลการเปรียบเทียบประสิทธิภาพของแบบจำลองที่เกิดจากข้อมูลที่มีการสูญหายและแบบจำลองที่เกิดจากข้อมูลที่แทนค่าข้อมูลสูญหายด้วยวิธี FSNNR

ผลการวิเคราะห์ปรากฏว่า ค่า log likelihood ของแบบจำลองที่เกิดจากข้อมูลที่มีการสูญหายมีค่าเท่ากับ -690.9153 ในขณะที่ค่า log likelihood ของแบบจำลองที่เกิดจากข้อมูลที่แทนค่าข้อมูลสูญหายด้วยวิธี FSNNR มีค่าเท่ากับ 372.2362 ซึ่งค่า log likelihood ของแบบจำลองที่เกิดจากข้อมูลที่แทนค่าข้อมูลสูญหายด้วยวิธี FSNNR มีค่าสูงกว่า แสดงให้เห็นว่า แบบจำลองที่เกิดจากข้อมูลที่แทนค่าข้อมูลสูญหายด้วยวิธี FSNNR สามารถอธิบายความสัมพันธ์ระหว่างตัวแปรได้ดีกว่า

การอภิปรายผล

การเลือกลักษณะ โดยใช้วิธี Nearest Neighbor วัดระยะทางด้วยวิธี Euclidean distance เป็นการลดจำนวนลักษณะ และเพื่อเพิ่มประสิทธิภาพในการแทนค่าข้อมูลสูญหายด้วยวิธี NNR ซึ่งสอดคล้องกับผลการศึกษาของ Li, Harner, and Adjeroh (2011) ได้พัฒนาวิธี Random KNN Feature Selection หรือ RKNN-FS โดยมีกระบวนการทำงานคือ ใช้วิธี KKN เพื่อหาระยะทางที่น้อยที่สุดระหว่างตัวแปร หลังจากนั้นเรียงความสำคัญของตัวแปรนำเข้า (Input variable) จึงคัดเลือกตัวแปรจำนวน K หน่วยมาใช้ในการวิเคราะห์ในส่วนอื่นๆ ต่อไป หลังจากการทดลองกับข้อมูลต่าง ๆ ปรากฏว่า วิธี RKNN-FS สามารถดำเนินการได้เร็วกว่าวิธีอื่น ๆ และมีความคงทนในการคัดเลือกลักษณะ และสำหรับการแทนค่าข้อมูลสูญหายด้วยวิธี NNR เป็นวิธีการแทนค่าข้อมูลสูญหายที่แก้ปัญหาการบิดเบือนการแจกแจงของตัวแปรอิสระและความสัมพันธ์ของตัวแปร และยังคงรักษาโครงสร้างของเมทริกซ์ความแปรปรวนความแปรปรวนร่วม (Variance-Covariance Matrix) ของตัวแปร Y เมื่อข้อมูลเกิดความสมบูรณ์ทุกหน่วยตัวอย่างแล้วจึงนำมาแทนค่าข้อมูลสูญหายด้วยวิธี RI อีกครั้ง เนื่องจากการประมาณค่าความสัมพันธ์ของตัวแปร เมื่อตัวแปรมีความสัมพันธ์กันสูง จะทำให้การประมาณค่ามีความแม่นยำสูง ซึ่งเป็นไปตามทฤษฎีของการถดถอยพหุเชิงเส้นและสามารถลดความคลาดเคลื่อนในการวิเคราะห์และสรุปผล (Eskelson et al., 2009)

การแทนค่าข้อมูลสูญหายด้วยวิธี FSNNR มีค่าเฉลี่ยความคลาดเคลื่อนสมบูรณ์ต่ำกว่าการแทนค่าวิธี RI, KNN และ NNR ทุกกรณี ผลการวิจัยชี้ให้เห็นว่า ในการแทนค่าข้อมูลสูญหายสามารถใช้วิธีการแทนค่าข้อมูลสูญหายด้วยวิธี FSNNR ภายใต้เงื่อนไขกรณีขนาดตัวอย่างมีขนาดใหญ่และระดับการสูญหายของข้อมูลทุกกรณี สอดคล้องกับผลการศึกษางานของ Li et al. (2011) และ Shweta et al. (2013) ที่เลือกลักษณะของข้อมูลโดยใช้ลักษณะเด่น (Feature) และทำการแทนค่าข้อมูลสูญหายพบว่า การเลือกลักษณะทำให้ได้ค่าของข้อมูลสูญหายใกล้เคียงกับข้อมูลจริง และทำให้ประสิทธิภาพของการแทนค่าข้อมูลสูญหายเพิ่มขึ้น และตรงกับผลงานวิจัยของ Beretta and Santaniello (2016) ที่พบว่าเมื่อขนาดของตัวอย่างมีขนาดกลาง (N = 100 – 400) และมีจำนวนตัวแปรไม่มากนัก การแทนค่าข้อมูลสูญหายด้วยวิธี K-Nearest Neighbor มีค่าความถูกต้องแม่นยำ ในขณะที่ค่า K มีขนาดเล็ก ยกเว้นกรณีที่ขนาดตัวอย่างเท่ากับ 50 ขนาดของส่วนเบี่ยงเบนมาตรฐานของความคลาดเคลื่อนของข้อมูล (μ = 0 และ σ = 5, 10 และ15) และระดับข้อมูลสูญหายเท่ากับ 5%

การวัดประสิทธิภาพของการผลิตปลานิลในจังหวัดสุพรรณบุรี หากชุดข้อมูลที่มีข้อมูลสูญหายมักทำให้ได้โมเดลที่มีประสิทธิภาพต่ำ โดยการทำนายของโมเดลอาจเกิดข้อผิดพลาดในการทำนายได้ ส่วนใหญ่ข้อมูลสูญหายมักจะเกิดมาจากการเก็บข้อมูลที่ไม่สมบูรณ์หรือการกรอกข้อมูลในระหว่างการจัดเก็บชุดข้อมูลเกิดผิดพลาด ดังนั้น การแทนค่าข้อมูลสูญหายด้วยวิธีการที่เหมาะสมก่อนการวิเคราะห์ข้อมูลจะพบว่า สถิติทดสอบและสถิติอื่น ๆ มีอำนาจทดสอบ (Power of test) สูงขึ้น การแทนค่าข้อมูลสูญหาย ทำให้ประสิทธิภาพของการประมาณค่าและการสรุปผลการวิจัยสูงขึ้น (Raymond, 2016) การวัดประสิทธิภาพของการผลิตปลานิลในจังหวัดสุพรรณบุรีด้วยแบบจำลอง Production Frontier Approach ในรูปแบบของฟังก์ชั่นการผลิต Cobb-Douglas ที่วิเคราะห์ได้ค่าสัมประสิทธิ์ของตัวแปรอิสระมีเครื่องหมายเป็นลบ 2 ตัวแปร คือ ค่าแรงงานและค่าเสื่อมสภาพของเครื่องจักร ส่วนพื้นที่ในการเลี้ยงและจำนวนอาหาร มีเครื่องหมายเป็นบวก มีเพียงตัวแปรพื้นที่ในการเลี้ยงเพียงตัวแปรเดียวที่มีผลต่อผลผลิตปลานิล ซึ่งสอดคล้องกับผลการศึกษา ของ Alawode and Jinad (2014) เพื่อศึกษาประสิทธิภาพทางเทคนิคของเกษตรกรผู้เลี้ยงปลาดุก ด้วยวิธี Stochastic Frontier Production Analysis Function ปรากฏว่า ปัจจัยนำเข้า ได้แก่ ค่าแรงงาน อาหารและขนาดของปลามีความสัมพันธ์ในทางลบ ในขณะที่จำนวนบ่อลูกปลามีความสัมพันธ์ในทางบวกกับผลผลิตปลาดุก และงานของ Alam, Khan, and Huq (2012) ที่ปรากฏว่า ปัจจัยนำเข้าได้แก่ อาหาร แรงงาน พลังงานเชื้อเพลิง มีความสัมพันธ์

ในทางบวก ในขณะที่ค่าบำรุงรักษาเครื่องจักร และค่าใช้จ่ายในการดำเนินการมีความสัมพันธ์ในทางลบกับผลผลิตของปลาต่อรอบการผลิต สำหรับตัวแปรพื้นที่ในการเลี้ยงเพียงตัวแปรเดียวที่มีผลต่อผลผลิตปลา

ประสิทธิภาพทางเทคนิคของการผลิตปลานิลของเกษตรกรแต่ละราย ปรากฏว่า ส่วนใหญ่ของเกษตรกรผู้เลี้ยงปลานิลเป็นกลุ่มที่มีค่าประสิทธิภาพทางเทคนิคระดับมากที่สุด ร้อยละ 82.76 และระดับมาก ร้อยละ 24.14 ทั้งนี้เนื่องจากกระทรวงเกษตรและสหกรณ์ได้เห็นถึงความสำคัญและศักยภาพของการผลิตสินค้าปลานิลของเกษตรกรไทย และแนวโน้มการตลาดที่มีอัตราการขยายตัวอย่างต่อเนื่องทั้งภายในและภายนอกประเทศ จึงได้มอบหมายให้กรมประมงจัดทำยุทธศาสตร์การพัฒนาปลานิล ปี พ.ศ. 2553 – 2557 มีเป้าหมายเพื่อให้ประเทศไทยเป็นผู้นำในการผลิตสินค้าปลานิลที่มีคุณภาพและได้มาตรฐาน โดยให้ความสำคัญกับทุกปัจจัยที่มีผลต่อการผลิตปลานิล เริ่มจากการควบคุมการผลิตตั้งแต่ต้นน้ำถึงปลายน้ำเพื่อให้ได้สินค้าที่มีความปลอดภัยและมีคุณภาพตามมาตรฐานสากล การวิจัยและพัฒนาด้านสายพันธุ์ การวิจัยและพัฒนาเพื่อลดต้นทุนการเลี้ยง การส่งเสริมให้เกิดการรวมกลุ่ม รวมถึงการส่งเสริมด้านการตลาด (กรมประมง, 2553) จึงส่งผลให้ประสิทธิภาพทางเทคนิคของการผลิตปลานิลของเกษตรกรอยู่ในระดับมากถึงมากที่สุด

นักวิจัยสามารถนำการแทนค่าข้อมูลสูญหายด้วยวิธี FSNNR ไปใช้ในการแทนค่าข้อมูลสูญหายในการวัดประสิทธิภาพการผลิตของผลผลิตทางการเกษตรประเภทอื่น ๆ เช่น การผลิตถั่วเหลือง การผลิตปลาสวยงาม

เอกสารอ้างอิง

กรมประมง. (2553). *ยุทธศาสตร์การพัฒนาปลานิล (พ.ศ. 2553–2557)*. กรุงเทพฯ: กรมประมง กระทรวงเกษตรและสหกรณ์.

Aigner, D., Lovell, C. A. K., & Schmidt, P. (1977). Formulation and estimation of stochastic frontier production function models. *Journal of Econometrics*, *6*(1), 21–37. https://doi.org/10.1016/0304-4076(77)90052-5

Alam, M. F., Khan, M. A., & Huq, A. A. Anwarul. (2012). Technical efficiency in Tilapia farming of Bangladesh: a stochastic frontier production approach. *Aquaculture International*, *20*(4), 619–634. https://doi.org/10.1007/s10499-011-9491-3

Alawode, O. O., & Jinad, A. O. (2014). Evaluation of technical efficiency of catfish produ ction in oyo state: a case study of Ibadan Metropolis. *Journal of Emerging Trends in Educational Research and Policy Studies*, *5*(2), 223–231.

Beretta, L., & Santaniello, A. (2016). Nearest neighbor imputation algorithms: a critical evaluation. *BMC Medical Informatics and Decision Making*, *16*(S3), 198-208.

Chaimongkol, W., & Suwattee, P. (2004). Nearest Neighbor-Regression Imputation (Vol. 5). *Presented at the Applied Statistics Conference 2004*, Chaingmai: Chaingmai University.

Eskelson, B. N. I., Temesgen, H., Lemay, V., Barrett, T. M., Crookston, N. L., & Hudak, A. T. (2009). The roles of nearest neighbor methods in imputing missing data in forest inventory and monitoring databases. *Scandinavian Journal of Forest Research*, *24*(3), 235–246. https://doi.org/10.1080/02827580902870490

Farrell, M. J. (1957). The measurement of productive efficiency. *Journal of the Royal Statistical Society. Series A (General)*, *120*(3), 253-290. https://doi.org/10.2307/2343100

Kumari, B., & Swarnkar, T. (2011). Filter versus wrapper feature subset selection in large dimensionality microarray : A Review. *International Journal of Computer Science and Information Technologies. 2*(3), 1048–1053.

Ibrahim, J. G., & Molenberghs, G. (2009). Missing data methods in longitudinal studies: a review. *TEST*, *18*(1), 1–43. https://doi.org/10.1007/s11749-009-0138-x

Islam, G. M. N., Tai, S. Y., & Kusairi, M. N. (2016). A stochastic frontier analysis of technical efficiency of fish cage culture in Peninsular Malaysia. *SpringerPlus*, *5*(1), 1–11. https://doi.org/10.1186/s40064-016-2775-3

Ladha, L., & Deepa, T. (2011). Feature selection methods and algorithms. *International Journal on Computer Science and Engineering (IJCSE)*, *3*(5), 1787–1797.

Li, S., Harner, E. J., & Adjeroh, D. A. (2011). Random KNN feature selection - a fast and stable alternative to Random Forests. *BMC Bioinformatics*, *12*(450), 1-11. https://doi.org/10.1186/1471-2105-12-450, 1–11.

Raymond, M. R. (2016). Missing data in evaluation research. *Evaluation & the Health Professions, 4*(9), 395-420. doi:10.1177/016327878600900401

Shweta, S., Nikita, J., & Madhvi, G. (2013). A Review paper on feature selection methodologies and their applications. *International Journal of Engineering Research and Development*, *7*(6), 57–61.

Troyanskaya, O., Cantor, M., Sherlock, G., Brown, P., Hastie, T., Tibshirani, R., & Altman, R. B. (2001). Missing value estimation methods for DNA microarrays. *Bioinformatics (Oxford, England)*, *17*(6), 520–525.