

การทำนายผลผลิตข้าวหอมมะลิ ภาคตะวันออกเฉียงเหนือ กรณีที่มีการสูญหาย
ของข้อมูลโดยการประมาณค่าสูญหายจากการปรับแก้
วิธีเคเนียร์เรสเนเบอร์โดยใช้ค่าเฉลี่ยเดไซล์

Predicting Jasmine Rice Yield in Northeastern Region in Missing Data
Imputation from Adjusted K Nearest Neighbor Method
Based on Decile Mean

พัทชนา สุวรรณแสน,¹ ภัทราวดี มากมี¹ และอาฟีฟี ลาเตะ²

Patchana Suwannasaen,¹ Patrawadee Makmee¹ and Afifi Lateh²

¹วิทยาลัยวิทยาการวิจัยและวิทยาการปัญญา มหาวิทยาลัยบูรพา

¹College of Research Methodology and Cognitive Science, Burapha University

²คณะศึกษาศาสตร์ มหาวิทยาลัยสงขลานครินทร์ วิทยาเขตปัตตานี

²Faculty of Education, Prince of Songkla University, Pattani Campus

Received: December 27, 2019

Revised: February 12, 2020

Accepted: February 19, 2020

บทคัดย่อ

การวิจัยนี้มีวัตถุประสงค์เพื่อพัฒนาวิธีการประมาณค่าข้อมูลสูญหายแบบใหม่ Decile Mean K Nearest Neighbor Bhattacharyya Imputation--DKNN-BH และเพื่อเปรียบเทียบประสิทธิภาพของวิธีการประมาณค่าข้อมูลสูญหายแบบใหม่ DKNN-BH กับวิธีการประมาณค่าข้อมูลสูญหายค่าเฉลี่ยเลขคณิต วิธีการประมาณค่าข้อมูลสูญหาย K Nearest Neighbor Imputation--KNN และวิธีการประมาณค่าข้อมูลสูญหาย Decile Mean K Nearest Neighbor Imputation--DKNN จากข้อมูลผลผลิตข้าวหอมมะลิและข้อมูลที่เกี่ยวข้อง ทำการทดลองซ้ำจำนวน 500 ครั้ง และทำนายผลผลิตข้าวหอมมะลิ ภาคตะวันออกเฉียงเหนือ เมื่อกำหนดให้มีข้อมูลสูญหาย โดยใช้วิธีการประมาณค่าข้อมูลสูญหายแบบใหม่ DKNN-BH ที่พัฒนาขึ้น ผลการวิจัยสรุปได้ดังนี้ การพัฒนาวิธีการประมาณค่าข้อมูลสูญหายแบบใหม่ DKNN-BH เกิดจากการปรับแก้วิธีการประมาณค่าข้อมูลสูญหายด้วยวิธี KNN โดยปรับแก้ด้วยการใช้ค่าเฉลี่ยเดไซล์และการหาระยะทางแบบ Bhattacharyya เมื่อนำมาใช้กับข้อมูลการทำนายผลผลิตข้าวหอมมะลิ ภาคตะวันออกเฉียงเหนือ ซึ่งกำหนดให้มีข้อมูลสูญหาย โดยเปรียบเทียบวิธีการประมาณค่าข้อมูลสูญหายแบบใหม่ DKNN-BH ที่พัฒนาขึ้นกับ 3 วิธีข้างต้น พบว่า ข้อมูลผลผลิตข้าวหอมมะลิ ภาคตะวันออกเฉียงเหนือที่ใช้วิธีการประมาณค่าข้อมูลสูญหายแบบใหม่ DKNN-BH มีค่าเฉลี่ยความคลาดเคลื่อนกำลังสองเฉลี่ยต่ำที่สุดในทุกกรณี โดยมีค่าเฉลี่ยความคลาดเคลื่อนกำลังสองเฉลี่ยต่ำที่สุด ที่ระดับการสูญหายร้อยละ 5 และค่า k เท่ากับ 11 และเมื่อนำไปทำนายผลผลิตข้าวหอมมะลิ ภาคตะวันออกเฉียงเหนือ ด้วยวิธีการวิเคราะห์ถดถอยพหุ โดยใช้วิธี Stepwise ในการคัดเลือกตัวแปร และเปรียบเทียบกับข้อมูลจริงจะมีค่าความคลาดเคลื่อนร้อยละสัมบูรณ์เฉลี่ย เท่ากับ 3.155

คำสำคัญ: ข้อมูลสูญหาย, ผลผลิตข้าวหอมมะลิ, การปรับแก้วิธีเคเนียร์เรสเนเบอร์

Abstract

This research aimed to develop a new method for calculating missing data by using Decile Mean K Nearest Neighbor Bhattacharyya Imputation method--DKNN-BH and compare the efficiency of the new missing data imputation method, DKNN-BH, with Mean Imputation, K Nearest Neighbor--KNN and Decile Mean K Nearest Neighbor Imputation--DKNN methods. This data was collected from the production of Jasmine rice and related data. In order to forecast the Jasmine rice yield prediction in Northeastern Thailand, the missing data was implemented to the DKNN-BH method and each situation was replicated 500 times. The results concluded that the newly developed missing data estimation method, DKNN-BH method, was derived from the fine tuning of KNN method, using decile mean and Bhattacharyya distance. After imputing the Jasmine rice yield in Northeastern Thailand data to these methods and comparing them with the 3 missing data imputation methods, the results showed that the estimated yield of Jasmine rice in Northeastern from DKNN-BH method had the lowest Mean Square Error in all cases at the percentage of missing data at 5% and k equal 11. This method was utilized to predict the Jasmine rice yield in Northeastern using multiple regression method. After comparing the results with real data, the Mean Absolute Percentage Error values were 3.155

Keywords: missing data, Jasmine rice yield, adjusted K Nearest Neighbor method



บทนำ

ข้อมูลที่มีความถูกต้องสมบูรณ์เป็นสิ่งที่นักวิจัย นักวิเคราะห์ข้อมูล หรือนักจัดการข้อมูลปรารถนา นำไปสู่ การใช้ข้อมูลในการคาดคะเน ทำนายหรือพยากรณ์สิ่งที่จะ เกิดขึ้นในอนาคต เพื่อให้เกิดประโยชน์ในการวางแผนล่วงหน้าได้อย่างใกล้เคียงความเป็นจริง ข้อมูลสูญหาย (missing data) เป็นปัญหาที่นักวิจัยพบในการนำข้อมูลไปวิเคราะห์ด้วย เทคนิคต่าง ๆ ในทางปฏิบัติเป็นไปได้ยากที่ข้อมูลจากการเก็บ รวบรวมจะมีความสมบูรณ์ถูกต้อง อาจเนื่องมาจากขั้นตอน ในการเก็บข้อมูลไม่รัดกุม ผู้ให้ข้อมูลไม่ใส่ใจในรายละเอียด ของการตอบ หรือเกิดเหตุการณ์ที่ไม่สามารถเก็บข้อมูล ได้ตามแผนเหตุการณ์ข้างต้นจึงทำให้เกิดข้อมูลสูญหายซึ่ง ปัญหาเหล่านี้อาจทำให้เกิดผลกระทบต่อประสิทธิภาพใน การวิเคราะห์ข้อมูล

ปัญหาข้อมูลสูญหายอาจไม่ใช่ปัญหาที่รุนแรง ถ้า การวิเคราะห์ข้อมูลเป็นการวิเคราะห์ตัวแปรเดียว (univariate data) เช่น การหาค่าเฉลี่ย ร้อยละ หรือสถิติพรรณนาอื่น ๆ

แต่ถ้าข้อมูลจำเป็นต้องใช้การวิเคราะห์ตัวแปรพหุ (multivariate data) เช่น การวิเคราะห์ถดถอยพหุ การวิเคราะห์ปัจจัย การวิเคราะห์กลุ่ม การวิเคราะห์จำแนก เป็นต้น ในกรณีนี้ การสูญหายของข้อมูลจะส่งผลกระทบต่อรุนแรง เพราะถ้าหน่วย วิเคราะห์ใดมีตัวแปรขาดหายไปเพียงหนึ่งตัวแปรก็จะตัด หน่วยวิเคราะห์นั้นทิ้งทั้งหน่วยโดยไม่สนใจว่าตัวแปรอื่นมี ข้อมูลครบถ้วนหรือไม่ ซึ่งการตัดข้อมูลทิ้ง Kim and Curry (1977) ได้ทำการศึกษาโดยการทดลองพบว่า ถ้าตัวแปร แต่ละตัวหายไปโดยสุ่มเพียงร้อยละ 10 จะมีผลทำให้ต้อง ตัดข้อมูลทิ้งถึงร้อยละ 59 ซึ่งเป็นการสูญเสียข้อมูลที่มีอัตรา สูงมาก และข้อมูลที่เหลือหลังจากการตัดค่าสังเกตที่ไม่ สมบูรณ์ทิ้งจะทำให้การวิเคราะห์ข้อมูล เกิดความเอนเอียง และไม่ถูกต้อง ดังนั้นควรพิจารณาหาวิธีการประมาณค่า ข้อมูลสูญหายด้วยวิธีที่เหมาะสมก่อนการวิเคราะห์ข้อมูล

วิธีการจัดการข้อมูลสูญหาย แบ่งเป็น 2 ประเภท หลัก คือ การตัดกลุ่มข้อมูลที่สูญหายทิ้งไป (ignoring and discarding data) และการประมาณค่าข้อมูลสูญหายด้วย

ค่าสูญหายจากการคำนวณ (imputation) โดยมีผู้ศึกษาและวิจัยเพื่อพัฒนาวิธีการจัดการค่าสูญหายด้วยวิธีการต่าง ๆ ที่มีประสิทธิภาพที่สุดมาใช้ในการประมาณค่าสูญหายของข้อมูลให้เหมาะสมกับประเภทของข้อมูลที่เกิดการสูญหายเริ่มตั้งแต่วิธีพื้นฐาน ได้แก่ การหาค่าเฉลี่ย (mean) และค่าฐานนิยม (mode) ซึ่งเป็นวิธีที่ง่ายในการประมาณค่าข้อมูลสูญหาย แต่วิธีการเหล่านี้อาจจะทำให้เกิดการเบี่ยงเบนของข้อมูลและเหมาะสมกับข้อมูลบางประเภทเท่านั้น ไปจนถึงวิธีที่มีความยุ่งยากซับซ้อนมาก ๆ เช่น Expectation Maximization หรือ Maximum Likelihood Estimation เป็นต้น การประมาณค่าข้อมูลสูญหายโดยใช้เทคนิคทางเหมืองข้อมูล เป็นอีกวิธีการหนึ่งที่มีความสนใจ ตัวอย่างการประมาณค่าข้อมูลสูญหายโดยเทคนิคทางเหมืองข้อมูล เช่น วิธีโครงข่ายประสาทเทียม หรือ Artificial Neural Networks (Bishop, 1995) วิธี Singular Value Decomposition--SVD แต่ทั้งสองวิธีนี้เป็นวิธีการประมาณค่าที่มีความซับซ้อน เข้าใจยาก อีกวิธีการหนึ่งในการประมาณค่าด้วยเทคนิคเหมืองข้อมูลที่มีความนิยม คือ วิธีเคเนียร์เรสเนเบอร์ หรือ K Nearest Neighbors--KNN (Troyanskaya et al., 2001) ซึ่งเป็นวิธีที่มีความยืดหยุ่นสูง ง่ายต่อการใช้งาน และมีประสิทธิภาพ แต่ยังมีจุดด้อยในการถูกโจมตีโดยค่านอกเกณฑ์ ผู้วิจัยจึงมีความสนใจในการพัฒนาวิธีการประมาณค่าสูญหายจากจุดด้อยของวิธี KNN ในส่วนนี้ และเมื่อมีการพัฒนาวิธีการประมาณค่าสูญหายจำเป็นอย่างยิ่งที่จะต้องสามารถนำไปใช้ประโยชน์ได้จริง จึงได้ค้นคว้าข้อมูลในงานวิจัยด้านต่าง ๆ ที่จะสามารถนำวิธีการประมาณค่าสูญหายไปใช้หนึ่งนั้น คือ ข้อมูลทางภาคการเกษตร

ภาคเกษตรกรรมมีบทบาทสำคัญอย่างมากต่อประเทศไทย เนื่องจากสภาพภูมิประเทศสามารถมีการเพาะปลูกพืชหลากหลายชนิดที่ให้ผลผลิตได้ตลอดปี ดังนั้นผลผลิตการเกษตรจึงมีผลต่อการเจริญเติบโตทางเศรษฐกิจของประเทศ (Office of Agricultural Economics, 2019) โดยพืชไร่นานานเป็นพืชในกลุ่มพืชอาหารที่สามารถนำส่วนต่าง ๆ ของพืชมาใช้ประกอบอาหารได้ ตัวอย่างพืชอาหารที่สำคัญ คือ ข้าว การปลูกข้าวขึ้นอยู่กับปัจจัยหลายอย่าง จากข้อมูลผลผลิตข้าวนานปีในปี 2560 ของสำนักงานเศรษฐกิจการเกษตรพบว่า ผลผลิตข้าวนานปี ของภาคตะวันออกเฉียงเหนือมีปริมาณมากที่สุดในประเทศ

ซึ่งมีปริมาณผลผลิตสูงถึงประมาณ 12.2 ล้านตัน นั่นคือภาคตะวันออกเฉียงเหนือเป็นแหล่งผลิตข้าวขนาดใหญ่ของประเทศ และข้าวหอมมะลิ ยังคงเป็นพันธุ์ข้าวที่ได้รับความนิยม มีปริมาณผลผลิตมาก สามารถขายและส่งออกสร้างรายได้ให้กับเกษตรกรและประเทศได้เป็นอย่างดี การทราบถึงปริมาณผลผลิตข้าวของประเทศไทยมีความสำคัญเป็นอย่างมาก สำหรับผู้บริหาร นักวิชาการเกษตร รวมทั้งเกษตรกร เพื่อนำไปใช้เป็นข้อมูลประกอบการตัดสินใจ การลงทุน การวางแผนการผลิต การกำหนดราคาข้าว และการกำหนดพื้นที่เพาะปลูก เป็นการส่งเสริมให้ปริมาณผลผลิตข้าวต่อไร่สูงขึ้น และยังส่งผลให้มูลค่าการส่งออกข้าวของประเทศไทยเพิ่มสูงขึ้นตามไปด้วย ดังนั้นการทราบถึงข้อมูลที่เหมาะสมต่าง ๆ ที่เกี่ยวข้องจะมีผลต่อการทำนายผลผลิตข้าว เพื่อช่วยในการสนับสนุนหรือให้ข้อมูลประกอบการตัดสินใจเพื่อวางแผนการผลิตที่ดี จากงานวิจัยของชนิษฐากุลนาวัน ปานจิตร หองประดิษฐ์, ชมัยพร เจริญพร และภัทรสินี ภัทรโกศล (Kulnawin, Longpradit, Chareanporn, & Bhattarakosol, 2014). ซึ่งศึกษาเปรียบเทียบวิธีทางเหมืองข้อมูลในการทำนายผลผลิตพืชไร่ทางการเกษตร มีขั้นตอนกระบวนการของการทำความเข้าใจข้อมูลก่อนจึงจะสามารถนำไปวิเคราะห์ข้อมูลด้วยเทคนิคทางเหมืองข้อมูลได้และพบว่า มีการนำเทคนิคในการประมาณค่าข้อมูลสูญหายเข้ามาช่วยทำให้ข้อมูลมีความสมบูรณ์จะทำให้การทำนายผลผลิต มีความถูกต้องแม่นยำและมีประสิทธิภาพมากยิ่งขึ้น

ดังนั้นในการศึกษาวิจัยในครั้งนี้ผู้วิจัยจึงสนใจจะพัฒนาวิธีการประมาณค่าข้อมูลสูญหายโดยการปรับแก้การประมาณค่าสูญหายด้วย KNN โดยใช้วิธีค่าเฉลี่ยเคิลล์ และการหาระยะทางแบบ Bhattacharyya มาช่วยในการปรับแก้เพื่อให้ได้วิธีการประมาณค่าสูญหายที่มีประสิทธิภาพในการประมาณค่ามากยิ่งขึ้น เรียกว่า วิธีการประมาณค่าสูญหาย Decile Mean K Nearest Neighbors Bhattacharyya--DKNN-BH เพื่อแก้ปัญหาค่านอกเกณฑ์ของข้อมูลในการประมาณค่าสูญหาย และประยุกต์นำไปใช้กับการทำนายผลผลิตข้าวหอมมะลิ ให้การทำนายผลผลิตข้าวหอมมะลิมีประสิทธิภาพ ถูกต้องและใกล้เคียงกับผลผลิตจริงมากที่สุด ซึ่งจะเป็นประโยชน์สำหรับนักวิชาการเกษตรและเกษตรกรในการเตรียมการให้ความรู้ เพื่อการเพิ่มผลผลิตที่มีคุณภาพต่อไป

วัตถุประสงค์ของการวิจัย

1. เพื่อพัฒนาวิธีการประมาณค่าข้อมูลสูญหายแบบใหม่ DKNN-BH ที่เกิดจากการปรับแก้ วิธีการประมาณค่าข้อมูลสูญหายจากวิธีเคเนียร์เรสเนเบอร์ด้วยค่าเฉลี่ยเดไซล์ และการหาระยะทาง Bhattacharyya

2. เพื่อเปรียบเทียบประสิทธิภาพของวิธีการประมาณค่าข้อมูลสูญหายแบบใหม่ DKNN-BH กับวิธีการประมาณค่าข้อมูลสูญหายค่าเฉลี่ยเลขคณิต (MI) วิธีการประมาณค่าข้อมูลสูญหาย KNN และวิธีการประมาณค่าข้อมูลสูญหาย DKNN โดยใช้ข้อมูลผลผลิตข้าวหอมมะลิ ภาคตะวันออกเฉียงเหนือ เป็นตัวแปรเกณฑ์ 1 ตัวแปร และข้อมูลที่เกี่ยวข้อง จำนวนตัวแปรทำนาย 5 ตัวแปร ประกอบด้วย พื้นที่เพาะปลูก (หน่วยเป็น แสนไร่) ปริมาณน้ำฝนเฉลี่ย (หน่วยเป็น มิลลิเมตร) ความชื้นสัมพัทธ์เฉลี่ย (หน่วยเป็น ชั่วโมง) อุณหภูมิเฉลี่ย (หน่วยเป็น องศาเซลเซียส) และราคาข้าวหอมมะลิเฉลี่ย (หน่วยเป็น บาท) โดยพิจารณา 2 เงื่อนไข คือ ระดับการสูญหายของข้อมูล 5 ระดับ คือ ร้อยละ 5 10 20 30 และ 40 ของขนาดข้อมูล และค่าคงที่ k 5 ระดับ คือ 11 13 15 17 และ 19

3. เพื่อทำนายผลผลิตข้าวหอมมะลิ ภาคตะวันออกเฉียงเหนือ เมื่อกำหนดให้มีข้อมูลสูญหาย โดยใช้วิธีการประมาณค่าข้อมูลสูญหายแบบใหม่ DKNN-BH วิธีการประมาณค่าข้อมูลสูญหายจากวิธีเคเนียร์เรสเนเบอร์ ด้วยค่าเฉลี่ยเดไซล์และการหาระยะทาง Bhattacharyya ที่พัฒนาขึ้น

กรอบแนวคิดการวิจัย

การประมาณค่าสูญหายของข้อมูลสามารถกระทำได้หลายวิธีไม่ว่าจะเป็นเทคนิคทางสถิติ เช่น การประมาณค่าด้วยศูนย์ การแทนที่ด้วยค่าเฉลี่ย การแทนที่ด้วยค่าฐานนิยม (Kim & Curry, 1977; Little, 1992; Kaiser, 2014) หรือโดยการใช้เทคนิคทางเหมืองข้อมูล (Data Mining) เช่น วิธีการแทนที่ด้วยวิธี KNN เป็นต้น จากงานวิจัยหลายงานวิจัยพบว่าวิธีการประมาณค่าสูญหายของข้อมูลด้วยเทคนิคทางเหมืองข้อมูลจากข้อมูลจริงโดยวิธี KNN เป็นวิธีการที่เหมาะสมในการประมาณค่าข้อมูล (Jerez et al., 2010) ดังนั้นผู้วิจัยจึงมีความสนใจในการประมาณค่าข้อมูลสูญหายสำหรับข้อมูลที่มีหลายตัวแปรโดยการพัฒนาเทคนิควิธีการ KNN และจากการศึกษาพบว่าวิธี KNN มี

ข้อดี คือ มีความยืดหยุ่นสูง ซึ่งมีการวัดความคล้ายคลึงกันของข้อมูล ทนต่อข้อมูลที่สูญหาย และง่ายต่อการใช้งาน แต่มีข้อเสียคือ ค่าตอบที่ได้อาจถูกโน้มเอียงโดยค่านอกเกณฑ์ (Meesad & Hengpraprom, 2008) วิธี KNN มีขั้นตอนการดำเนินการ คือ เริ่มที่กำหนด ค่า k จากนั้นเป็นการคำนวณหาระยะทางระหว่างข้อมูลตัวอย่างที่สนใจกับข้อมูลอื่น ๆ ทุกตัวด้วยการหาระยะทาง Euclidian เพื่อดูความคล้ายคลึงกันของข้อมูล และเลือกค่าข้อมูลที่มีค่าระยะทางน้อยที่สุด k ตัว เพื่อนำมาพิจารณาหาคำตอบด้วยการหาค่าเฉลี่ยจากค่า k ตัวที่ได้

จากวิธีการประมาณค่าสูญหายของ KNN พบว่า ในขั้นตอนของการประมาณค่าทดแทนนั้นคำตอบที่ได้มาจากหาค่าเฉลี่ยจากสมาชิก k ตัว นั่นคือ การประมาณค่าสูญหายจากค่าเฉลี่ยของสมาชิก k ตัว นั่นเอง เป็นสาเหตุทำให้เกิดข้อเสียของวิธี KNN ที่ยังคงถูกโน้มเอียงโดยค่านอกเกณฑ์ได้ จากการศึกษาของงานวิจัยของ Rana (Rana, Siraj-Ud-Doulah, Midi, & Imon, 2012) ศึกษาการวัดแนวโน้มเข้าสู่ส่วนกลางแบบใหม่ที่มีความแกร่ง และพบว่าวิธีค่าเฉลี่ยเดไซล์ เป็นวิธีที่มีความแกร่งและมีความไวต่อค่านอกเกณฑ์ นอกจากนี้ในขั้นตอนของการหาระยะทางซึ่งเดิมใช้การหาระยะทาง Euclidian แต่จากการศึกษาพบว่าสถิติ Bhattacharyya มีความไวต่อค่านอกเกณฑ์ (Pasunon & Nilakorn, 2007) เช่นกันจึงนำมาซึ่งการศึกษาเกี่ยวกับการหาระยะแบบ Bhattacharyya นอกจากนี้ในขั้นตอนของการหาความคล้ายคลึงกันของข้อมูลโดยใช้การหาระยะทางนั้น จากงานวิจัยของ Bhattacharyya (1943) พบวิธีการหาระยะทาง Bhattacharyya จากคุณสมบัติข้างต้นผู้วิจัยจึงได้นำวิธีค่าเฉลี่ยเดไซล์และการหาระยะทาง Bhattacharyya มาช่วยในการปรับแก้การประมาณค่าข้อมูลสูญหายร่วมกับวิธี KNN ซึ่งจะได้กรอบแนวคิดของการปรับแก้วิธีการประมาณค่าข้อมูลสูญหายแบบใหม่

ในการศึกษาเอกสารและงานวิจัยที่เกี่ยวข้อง พบว่าในการประมาณค่าสูญหายนั้นจำนวนการสูญหายก็ส่งผลต่อการประมาณค่า โดย Little and Rubin (1987) ได้กล่าวถึงจำนวนของการสูญหายไว้ว่า ถ้าข้อมูลสูญหายน้อยกว่าร้อยละ 5 การตัดข้อมูลทิ้งอาจไม่มีปัญหา แต่เมื่อข้อมูลสูญหายมีจำนวนระหว่างร้อยละ 15–20 การจัดการข้อมูลสูญหายจะมีความจำเป็น และจะมีความสำคัญมากที่สุดเมื่อมีจำนวนสูญหาย ร้อยละ 30–40 ดังนั้นการสูญหาย

ในช่วงระหว่างร้อยละ 5-40 ในการสูญหายของข้อมูลจะมีความสำคัญ ผู้วิจัยจึงกำหนดปริมาณการสูญหายอยู่ในช่วงดังกล่าว และเกณฑ์ที่ใช้ในการทดสอบวิธีการประมาณค่าสูญหายของข้อมูลในที่นี่ใช้วิธีค่าคลาดเคลื่อนกำลังสองเฉลี่ย (MSE) ซึ่งเป็นวิธีที่มีความนิยมและมีความน่าเชื่อถือในการตรวจสอบประสิทธิภาพของการประมาณค่าสูญหายของแต่ละวิธี

วิธีดำเนินการวิจัย

การทำนายผลผลิตข้าวหอมมะลิ ภาคตะวันออกเฉียงเหนือ เมื่อกำหนดให้มีข้อมูลสูญหาย มีวิธีดำเนินการวิจัย ขั้นตอนรายละเอียด ดังนี้

ตัวแปรที่ใช้ศึกษา

1. ตัวแปรทำนาย มี 5 ตัวแปร คือ พื้นที่เพาะปลูก (หน่วยเป็น แสนไร่) ปริมาณน้ำฝนเฉลี่ย (หน่วยเป็น มิลลิเมตร) ความชื้นสัมพัทธ์เฉลี่ย (หน่วยเป็น ชั่วโมง) อุณหภูมิเฉลี่ย (หน่วยเป็น องศาเซลเซียส) และราคาข้าวหอมมะลิเฉลี่ย (หน่วยเป็น บาท)

2. ตัวแปรเกณฑ์ มี 1 ตัวแปร คือ ค่าทำนายผลผลิตข้าวหอมมะลิ ภาคตะวันออกเฉียงเหนือ (หน่วยเป็น แสนตัน)

ตัวอย่าง

ตัวอย่าง คือ ข้อมูลผลผลิตข้าวหอมมะลิ ในภาคตะวันออกเฉียงเหนือของประเทศไทย และข้อมูลที่เกี่ยวข้องกับผลผลิตการเกษตร ซึ่งเป็นปัจจัยทางธรรมชาติจากสิ่งแวดล้อมที่มีผลต่อผลผลิตข้าวหอมมะลิ ซึ่งเป็นข้อมูล ตั้งแต่ปี พ.ศ. 2538 ถึง 2560

การเก็บรวบรวมข้อมูล

ผู้วิจัยขอความอนุเคราะห์ให้ข้อมูลทุติยภูมิ (secondary data) ซึ่งเป็นข้อมูลผลผลิตข้าวหอมมะลิ ในภาคตะวันออกเฉียงเหนือของประเทศไทย และข้อมูลที่เกี่ยวข้องกับการเกษตร จากสำนักงานเศรษฐกิจการเกษตร และกรมอุตุนิยมวิทยา

การวิเคราะห์ข้อมูล

การวิเคราะห์ข้อมูลประกอบด้วย 2 ส่วน ได้แก่ ส่วนที่ 1 การประมาณค่าข้อมูลสูญหาย เนื่องจากในการ

พัฒนาวิธีการประมาณค่าข้อมูลสูญหายแบบใหม่จำเป็นต้องมีการนำไปใช้กับข้อมูลจริงได้จริง ผู้วิจัยจึงสนใจศึกษาข้อมูลที่เกี่ยวข้องกับการทำนายผลผลิตข้าวหอมมะลิ ภาคตะวันออกเฉียงเหนือเพื่อเปรียบเทียบวิธีการประมาณค่าข้อมูลสูญหายแบบใหม่ DKNN-BH ให้ผลการประมาณค่าได้ใกล้เคียงกับค่าจริง และส่วนที่ 2 นำข้อมูลที่มีการประมาณค่าข้อมูลสูญหายด้วยวิธีการประมาณค่าสูญหายแบบใหม่ ไปทำการทำนายผลผลิตข้าวหอมมะลิ ภาคตะวันออกเฉียงเหนือ

1. การประมาณค่าผลผลิตข้าวหอมมะลิ ภาคตะวันออกเฉียงเหนือ เมื่อมีการกำหนดให้มีข้อมูลสูญหาย

1.1 รวบรวมข้อมูลผลผลิตข้าวหอมมะลิ ในภาคตะวันออกเฉียงเหนือของประเทศไทย ตั้งแต่ปี พ.ศ. 2538 ถึง 2560

1.2 ศึกษาลักษณะของข้อมูล เพื่อพิจารณาตัวแปรที่เกี่ยวข้องเหมาะสมกับการทำนายผลผลิตข้าวหอมมะลิ ในภาคตะวันออกเฉียงเหนือของประเทศไทย

1.3 ประมาณค่าข้อมูลสูญหายจากการกำหนดให้มีการสูญหาย ด้วยวิธีการประมาณค่าข้อมูลสูญหายแบบใหม่ DKNN-BH ที่พัฒนาขึ้นเปรียบเทียบกับวิธีการประมาณค่าสูญหายโดยค่าเฉลี่ยเลขคณิต วิธีการประมาณค่าข้อมูลสูญหาย KNN และวิธีการประมาณค่าข้อมูลสูญหาย DKNN โดยประมาณค่าสูญหายภายใต้ระดับการสูญหายของข้อมูล 5 ระดับ คือ ร้อยละการสูญหายของข้อมูลเป็น ร้อยละ 5 10 20 30 และ 40 และ ค่าคงที่ k 5 ระดับ คือ 11 13 15 17 และ 19 ซึ่งเป็นค่าคงที่ในวิธีการประมาณค่าข้อมูลสูญหายแบบใหม่ DKNN-BH ซึ่งใช้ข้อมูลจริง ใน 1.1 และ 1.2 โดยกำหนดให้มีการสูญหายในผลผลิตข้าวหอมมะลิ ภาคตะวันออกเฉียงเหนือ เพียงตัวแปรเดียว

1.4 ตรวจสอบความถูกต้องในการประมาณค่าข้อมูลสูญหายด้วยวิธีการประมาณค่าข้อมูลสูญหายแบบใหม่ DKNN-BH และวิธีทั้ง 3 วิธี กับค่าจริงของผลผลิตข้าวหอมมะลิ ภาคตะวันออกเฉียงเหนือ โดยเปรียบเทียบค่าทดแทนจากวิธีการประมาณค่าทั้ง 4 วิธีกับผลผลิตของข้อมูลจริง เพื่อตรวจสอบความถูกต้องว่าจะให้ค่าความแม่นยำในการประมาณค่าสูญหายได้ใกล้เคียงกับค่าจริงเพียงใด โดยพิจารณาจากค่าความคลาดเคลื่อนกำลังสองเฉลี่ย (Mean Square Error--MSE) จากการทำซ้ำ 500 รอบ โดยมีสูตรในการคำนวณ ดังนี้

$$MSE = \frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2 \quad (1)$$

1.5 เลือกข้อมูลจากวิธีที่ให้ค่า MSE น้อยที่สุด เพื่อนำไปทำนายผลผลิตข้าวหอมมะลิ ภาคตะวันออกเฉียงเหนือ ในขั้นตอนต่อไป

2. การทำนายผลผลิตข้าวหอมมะลิ ภาคตะวันออกเฉียงเหนือ ที่ได้จากการข้อมูลในขั้นตอนก่อนหน้า การวิเคราะห์ข้อมูลในขั้นตอนนี้เป็นการสร้างตัวแบบของการทำนายผลผลิตข้าวหอมมะลิ ภาคตะวันออกเฉียงเหนือ ที่ได้จากการประมาณค่าข้อมูลสูญหาย ซึ่งใช้วิธีการประมาณค่าข้อมูลสูญหายวิธีใหม่ DKNN-BH โดยสร้างตัวแบบสำหรับการทำนายผลผลิตข้าวด้วยวิธีการวิเคราะห์ถดถอยพหุ หรือ Multiple Regression Analysis มีวิธีการ ดังนี้

2.1 สร้างตัวแบบสำหรับการทำนายผลผลิตข้าวด้วยวิธีการวิเคราะห์ถดถอยพหุ หรือ Multiple Regression Analysis จากข้อมูลสมบูรณ์ที่แทนค่าแล้วจากวิธีใหม่ DKNN-BH ที่พัฒนาขึ้นโดยตัวแปรที่ศึกษาในครั้งนี้ได้จากการสังเคราะห์งานวิจัยที่เกี่ยวข้องกับการทำนายผลผลิตข้าว และผลผลิตทางการเกษตร ซึ่งตัวแปรที่ใช้ศึกษาประกอบด้วย ผลผลิตข้าวหอมมะลิ พื้นที่เพาะปลูกข้าว ปริมาณน้ำฝนเฉลี่ย ความชื้นสัมพัทธ์เฉลี่ย อุณหภูมิเฉลี่ย และราคาข้าวหอมมะลิเฉลี่ย โดยสร้างตัวแบบจากข้อมูลตั้งแต่ปี พ.ศ. 2538 ถึง 2560

2.2 วิเคราะห์ผลรูปแบบจำลอง โดยเริ่มจากการหาความสัมพันธ์ระหว่างตัวแปร แล้วนำมาสร้างรูปแบบความสัมพันธ์ โดยใช้วิธี Stepwise ในการคัดเลือกตัวแปร

2.3 ตรวจสอบตัวแบบการทำนายผลผลิตข้าวหอมมะลิ ภาคตะวันออกเฉียงเหนือ ของปี พ.ศ. 2538 ถึง 2560 โดยเปรียบเทียบผลการทำนายกับผลผลิตของข้อมูลปีเดียวกันจากข้อมูลจริง เพื่อตรวจสอบค่าทำนายว่าจะให้ค่าความแม่นยำในการทำนายใกล้เคียงกับค่าจริงเพียงใด โดยพิจารณาจากค่าความคลาดเคลื่อนร้อยละสมบูรณ์เฉลี่ย (Mean Absolute Percentage Error--MAPE) โดยค่า MAPE มีสูตรการคำนวณ ดังนี้

$$MAPE = \frac{100\%}{n} \sum_{i=1}^n \left| \frac{y_i - \hat{y}_i}{y_i} \right| \quad (2)$$

เมื่อ y_i คือ ค่าผลผลิตจริงของข้อมูล

$\hat{y}_i$ คือ ค่าผลผลิตทดแทนของข้อมูลที่มี

การสูญหาย

n คือ จำนวนข้อมูล

ผลการวิจัย

การพัฒนาวิธีการประมาณค่าข้อมูลสูญหายด้วยวิธีใหม่เกิดจากการปรับแก้วิธีการประมาณค่าสูญหายวิธี KNN ด้วยการใช้ค่าเฉลี่ยเดไซ์และการหาระยะทางแบบ Bhattacharyya เพื่อประมาณค่าข้อมูลที่สูญหาย เรียกว่า การประมาณค่าข้อมูลสูญหายวิธี Decile Mean K Nearest Neighbor Bhattacharyya Imputation--DKNN-BH มีขั้นตอนการดำเนินการ ดังนี้

ส่วนที่ 1 คำนวณระยะทาง

ขั้นที่ 1 กำหนดจำนวน k โดยค่า k ที่เหมาะสมมีค่าอยู่ระหว่าง 10 ถึง 20 (Cartwright, Shepperd & Song, 2003; Ladha & Deepa, 2011)

ขั้นที่ 2 คำนวณระยะห่างระหว่างตัวอย่างที่ประกอบด้วยข้อมูลสูญหายกับตัวอย่างทั้งหมด โดยกำหนดให้ เป็นข้อมูลสูญหาย เมื่อ y ที่ $i = 1, 2, \dots, r$ $j = 1, 2, \dots, m$ และ d_i เป็นค่าระยะห่างระหว่างข้อมูลตำแหน่งที่ i ที่มีการสูญหายกับข้อมูลอื่น ๆ ที่มีค่าสมบูรณ์ โดยคำนวณได้จากสูตร ดังนี้

$$d_i = Bhattacharyya(y_{i'j}, y_{ij}) = -\log \left[\frac{\sqrt{y_{i'j} y_{ij}}}{\sqrt{\sum y_{i'j} \sum y_{ij}}} \right] \quad (3)$$

ขั้นที่ 3 เรียงระยะห่างที่คำนวณได้ทั้งหมดจากน้อยไปหามากและเลือกค่าที่มีระยะห่างกับค่าสูญหายน้อยที่สุด k ตัว

ส่วนที่ 2 การประมาณค่าสูญหาย

ขั้นที่ 4 นำค่าผลผลิตข้าวในตำแหน่งจากขั้นที่ 3 มาคำนวณหาค่าเฉลี่ยในทุตำแหน่ง จากข้อมูล k ตัว ซึ่งจะได้ค่าเฉลี่ยที่ 1 ถึง 9

ขั้นที่ 5 ประมวลค่าสูญหายโดยนำค่าจากขั้นตอนที่ 4 มาหาค่าเฉลี่ยเฉลี่ย และทำการประมวลค่าข้อมูลสูญหายจากสูตร

$$\text{ค่าสูญหาย} = \hat{y}_{ij} = \frac{\sum_{r=1}^9 D_r}{9} \quad (4)$$

เมื่อ D_r คือ ค่าเฉลี่ยของข้อมูล ในตำแหน่งที่ r

r คือ ตำแหน่งเฉลี่ยของข้อมูล มีค่าเท่ากับ 1, 2, 3, ..., 9

โดยค่าที่ได้จะเปลี่ยนแปลงไปขึ้นกับตำแหน่งของค่าที่สูญหาย

ผลการทำนายผลผลิตข้าวหอมมะลิ ภาคตะวันออก เชียงเหนือ เมื่อกำหนดให้มีข้อมูลสูญหาย มีรายละเอียด ดังนี้

ตาราง 1

ค่าสัมประสิทธิ์สหสัมพันธ์เพียร์สันระหว่างตัวแปรทำนายกับผลผลิตข้าวหอมมะลิ

ข้อมูลตัวแปร	ค่าสัมประสิทธิ์สหสัมพันธ์เพียร์สัน					
	ผลผลิตข้าว	พื้นที่เพาะปลูก	ปริมาณน้ำฝนเฉลี่ย	ความชื้นสัมพัทธ์เฉลี่ย	อุณหภูมิเฉลี่ย	ราคาข้าวเฉลี่ย
1. ผลผลิตข้าว	1.000	.987**	-.225**	0.071	.214**	.231**
2. พื้นที่เพาะปลูก		1.000	-.271**	0.044	.194**	.187**
3. ปริมาณน้ำฝนเฉลี่ย			1.000	.245**	-.335**	.228**
4. ความชื้นสัมพัทธ์เฉลี่ย				1.000	-.293**	.347**
5. อุณหภูมิเฉลี่ย					1.000	-.151**
6. ราคาข้าวเฉลี่ย						1.000

หมายเหตุ ** หมายถึง มีนัยสำคัญทางสถิติที่ระดับ .01

ข้อมูลตัวแปรที่ใช้ในการศึกษา

การศึกษาครั้งนี้ใช้ข้อมูลทุติยภูมิจากแหล่งข้อมูล 2 แหล่ง คือ สำนักงานเศรษฐกิจการเกษตรและกรมอุตุนิยมวิทยา ซึ่งเป็นข้อมูลที่เกี่ยวข้องทางการเกษตร ประกอบด้วย ผลผลิต พื้นที่เพาะปลูก และราคาข้าวหอมมะลิ ส่วนข้อมูลที่เกี่ยวข้องกับปัจจัยทางธรรมชาติจากสิ่งแวดล้อม ซึ่งได้แก่ ปริมาณน้ำฝน อุณหภูมิเฉลี่ยและความชื้นสัมพัทธ์ จากข้อมูล 17 จังหวัด สามารถสรุปข้อมูลตัวแปรต่าง ๆ ที่ใช้ในการศึกษาซึ่งเป็นการเตรียมข้อมูลก่อนการทำนายในขั้นตอนต่อไป จากข้อมูลจำนวน 330 ข้อมูล ซึ่งขั้นตอนนี้เป็นส่วนของการเตรียมข้อมูล ได้ผลความสัมพันธ์รายคู่ระหว่างผลผลิตข้าวหอมมะลิกับตัวแปรอื่น ๆ โดยพิจารณาจากค่าสัมประสิทธิ์สหสัมพันธ์เพียร์สันระหว่างตัวแปรทำนายทั้ง 5 ตัวแปรกับผลผลิตข้าวหอมมะลิได้ดังแสดงในตาราง 1

จากตาราง 1 พบว่า ตัวแปรที่มีความสัมพันธ์กับผลผลิตข้าวหอมมะลิ ได้แก่ พื้นที่เพาะปลูก ปริมาณน้ำฝนเฉลี่ย อุณหภูมิเฉลี่ยและราคาข้าวเฉลี่ย โดยตัวแปรพื้นที่เพาะปลูกมีความสัมพันธ์กับผลผลิตข้าวหอมมะลิสุงสุด มีค่าสัมประสิทธิ์สหสัมพันธ์เพียร์สัน เท่ากับ 0.987 และรองลงมาคือ ราคาข้าวเฉลี่ย มีค่าสัมประสิทธิ์สหสัมพันธ์เพียร์สัน เท่ากับ 0.231 ถึงแม้ว่าตัวแปรพื้นที่เพาะปลูก ปริมาณน้ำฝนเฉลี่ย ความชื้นสัมพัทธ์เฉลี่ย อุณหภูมิเฉลี่ยและราคาข้าวเฉลี่ยจะมีความสัมพันธ์กันเองดังตาราง 1 แต่ไม่เกิน 0.8 และยังมีค่าสัมประสิทธิ์สหสัมพันธ์เพียร์สันน้อยกว่าความสัมพันธ์ระหว่างตัวแปรเหล่านี้กับผลผลิตข้าวหอมมะลิ จึงเป็นไปตามเกณฑ์การป้องกันการเกิดตัวแปรอิสระมีความสัมพันธ์กันเอง หรือปัญหาที่เรียกว่าปัญหา Multicollinearity (Stevens, 1992) ดังนั้น งานวิจัยนี้จึงสามารถใช้ตัวแปรพื้นที่เพาะปลูก ปริมาณน้ำฝนเฉลี่ย ความชื้นสัมพัทธ์เฉลี่ย อุณหภูมิเฉลี่ยและราคาข้าวเฉลี่ย เป็นตัวแปรทำนายในการทำนายผลผลิตข้าวหอมมะลิ

2. การประมาณค่าผลผลิตข้าวหอมมะลิ ภาคตะวันออกเฉียงเหนือ เมื่อมีการกำหนดให้มีข้อมูลสูญหาย

นำข้อมูลผลผลิตข้าวหอมมะลิ ภาคตะวันออกเฉียงเหนือ และข้อมูลที่เกี่ยวข้องจำนวน 330 ข้อมูล มาทำการประมาณค่าสูญหาย ด้วยวิธีการประมาณค่าข้อมูล

สูญหายแบบใหม่ DKNN-BH ที่พัฒนาขึ้นเปรียบเทียบกับวิธีการประมาณค่าสูญหายโดยค่าเฉลี่ยเลขคณิต (Mean Imputation--MI) วิธีการประมาณค่าข้อมูลสูญหาย KNN และวิธีการประมาณค่าข้อมูลสูญหาย DKNN โดยประมาณค่าสูญหายภายใต้ระดับการสูญหายของข้อมูล 5 ระดับ คือ ร้อยละการสูญหายของข้อมูลเป็นร้อยละ 5 10 20 30 และ 40 ค่าคงที่ k 5 ระดับ คือ 11 13 15 17 และ 19 ซึ่งเป็นค่าคงที่ในวิธีการประมาณค่าข้อมูลสูญหายแบบใหม่ DKNN-BH โดยดำเนินการทำซ้ำ 500 รอบ โดยใช้เกณฑ์ในการพิจารณาเปรียบเทียบประสิทธิภาพ คือ ค่าความคลาดเคลื่อนกำลังสองเฉลี่ย (Mean Square Error--MSE) เพื่อสรุปว่าวิธีการประมาณค่าข้อมูลสูญหายวิธีการใดเป็นวิธีที่เหมาะสมที่สุดในแต่ละ สถานการณ์ โดยวิธีใดที่ให้ค่า MSE ต่ำกว่าจะเป็นวิธีที่มีประสิทธิภาพมากกว่า ผลการวิจัยจำแนกตามเกณฑ์ในการพิจารณาเปรียบเทียบประสิทธิภาพ จากข้อมูลจริงผลผลิตข้าวหอมมะลิและตัวแปรทำนายที่เกี่ยวข้อง ได้ผลการเปรียบเทียบ ดังนี้

ตาราง 2

ค่าความคลาดเคลื่อนกำลังสองเฉลี่ยของการประมาณค่าข้อมูลสูญหาย เมื่อพิจารณาจากข้อมูลผลผลิต ข้าวหอมมะลิ ภาคตะวันออกเฉียงเหนือ ปี พ.ศ. 2538-2560

ค่า k	ระดับการสูญหาย	วิธีการประมาณค่าข้อมูลสูญหาย			
		MI	KNN	DKNN	DKNN-BH
11	5%	9.67129	8.62614	8.63815	<u>1.18846</u>
	10%	9.49366	8.46427	8.47704	<u>1.28645</u>
	20%	9.64372	8.55995	8.57417	<u>1.42470</u>
	30%	9.51111	8.68911	8.70751	<u>1.58806</u>
	40%	9.54945	8.75110	8.76950	<u>1.79034</u>

ตาราง 2 (ต่อ)

ค่า k	ระดับการสูญหาย	วิธีการประมาณค่าข้อมูลสูญหาย			
		MI	KNN	DKNN	DKNN-BH
13	5%	9.67129	8.49097	8.48553	<u>1.35486</u>
	10%	9.49366	8.45232	8.45477	<u>1.43492</u>
	20%	9.64372	8.56014	8.57020	<u>1.59476</u>
	30%	9.51111	8.66847	8.68169	<u>1.75509</u>
	40%	9.54945	8.71615	8.72927	<u>1.97623</u>
15	5%	9.67129	8.56536	8.58382	<u>1.50133</u>
	10%	9.49366	8.48176	8.50119	<u>1.55625</u>
	20%	9.64372	8.57510	8.59170	<u>1.73018</u>
	30%	9.51111	8.63238	8.64498	<u>1.91168</u>
	40%	9.54945	8.69899	8.70967	<u>2.15811</u>
17	5%	9.67129	8.58818	8.59802	<u>1.63195</u>
	10%	9.49366	8.48843	8.49943	<u>1.67443</u>
	20%	9.64372	8.55074	8.55957	<u>1.87009</u>
	30%	9.51111	8.60608	8.61596	<u>2.06537</u>
	40%	9.54945	8.68975	8.69904	<u>2.35018</u>
19	5%	9.67129	8.58584	8.60514	<u>1.75730</u>
	10%	9.49366	8.43291	8.44629	<u>1.80735</u>
	20%	9.64372	8.53720	8.54819	<u>2.00483</u>
	30%	9.51111	8.62224	8.63181	<u>2.22799</u>
	40%	9.54945	8.67743	8.68218	<u>2.56390</u>

จากตาราง 2 เมื่อพิจารณาความคลาดเคลื่อนกำลังสองเฉลี่ย จากข้อมูลผลผลิตข้าวหอมมะลิและตัวแปรทำนายที่เกี่ยวข้องเมื่อข้อมูลมีระดับการสูญหาย เท่ากับ ร้อยละ 5 10 20 30 และ 40 และค่า k เท่ากับ 11 13 15 17 และ 19 พบว่า วิธีการประมาณค่าข้อมูลสูญหายแบบ DKNN-BH มีค่าความคลาดเคลื่อนกำลังสองเฉลี่ยต่ำที่สุดหรือมีความแม่นยำสูงสุดในทุกค่ากรณี และพบว่า เมื่อข้อมูลมีระดับการสูญหาย ร้อยละ 5 และค่า k เท่ากับ 11 ในการรันรอบที่ 327 ให้ค่าความคลาดเคลื่อนกำลังสองเฉลี่ย เท่ากับ 0.18500 ซึ่งเป็นค่าต่ำที่สุด ดังนั้นจึงนำข้อมูลนี้ไปสร้างตัวแบบในการทำนายผลผลิตข้าวหอมมะลิ ภาคตะวันออกเฉียงเหนือ ในขั้นตอนต่อไป

3. การสร้างตัวแบบเพื่อการทำนายผลผลิตข้าวหอมมะลิ ภาคตะวันออกเฉียงเหนือ

นำข้อมูลจากขั้นตอนข้างต้นมาสร้างตัวแบบการทำนายผลผลิตข้าวหอมมะลิภาคตะวันออกเฉียงเหนือ และนำไปทำนายผลผลิตข้าวหอมมะลิ โดยข้อมูลที่จะทำนายผลผลิตข้าวหอมมะลิ ภาคตะวันออกเฉียงเหนือ เป็นข้อมูลรายภาค รายปีของปี พ.ศ. 2538–2560 จำนวน 22 ปี

3.1 ข้อมูลเบื้องต้นของข้อมูลผลผลิตข้าวหอมมะลิ ภาคตะวันออกเฉียงเหนือ จากการประมาณค่าข้อมูลสูญหายที่ระดับการสูญหาย ร้อยละ 5 และค่า k เท่ากับ 11 ในการทำซ้ำรอบที่ 327 ด้วยวิธีประมาณค่า

ข้อมูลสูญหายแบบใหม่ DKNN-BH มีลักษณะข้อมูลดังแสดงในตาราง 3

จากตาราง 3 พบว่า ผลผลิตข้าวหอมมะลิ ภาคตะวันออกเฉียงเหนือ มีค่าเฉลี่ย 46.004 แสตัน ค่าเบี่ยงเบนมาตรฐาน 12.619 พื้นที่เพาะปลูก ภาคตะวันออกเฉียงเหนือ มีค่าเฉลี่ย 145.380 แสนไร่ ปริมาณน้ำฝนเฉลี่ย มีค่าเฉลี่ย 1212.739 มิลลิเมตร ความชื้นสัมพัทธ์เฉลี่ย มีค่าเฉลี่ย 72.735 เปอร์เซ็นต์ อุณหภูมิเฉลี่ย มีค่าเฉลี่ย 26.923°C และราคาข้าวเฉลี่ย มีค่าเฉลี่ย 9,847.649 บาท และมีค่าสูงสุด ต่ำสุด ส่วนเบี่ยงเบนมาตรฐานดังแสดงในตารางข้างต้น

3.2 ตัวแบบการทำนายผลผลิตข้าวหอมมะลิ ภาคตะวันออกเฉียงเหนือ

นำข้อมูลข้างต้นมาสร้างตัวแบบสำหรับการทำนายผลผลิตข้าวหอมมะลิ ภาคตะวันออกเฉียงเหนือ ด้วยวิธีการวิเคราะห์ถดถอยพหุ โดยมีตัวแปรที่นำมาพิจารณาดังต่อไปนี้

ตัวแปรเกณฑ์ คือ

y = ปริมาณผลผลิตข้าวหอมมะลิภาคตะวันออกเฉียงเหนือ
ตัวแปรทำนาย คือ

x_{Cul_area} = พื้นที่เพาะปลูกภาคตะวันออกเฉียงเหนือ

x_{RT} = ปริมาณน้ำฝนเฉลี่ย

x_{RH} = ความชื้นสัมพัทธ์เฉลี่ย

x_{Temp} = อุณหภูมิเฉลี่ย

x_{Cost} = ราคาข้าวหอมมะลิเฉลี่ย

ตาราง 3

ลักษณะข้อมูลเบื้องต้นของข้อมูลผลผลิตข้าวหอมมะลิ ภาคตะวันออกเฉียงเหนือ รายปี ของปี พ.ศ. 2538–2560 ที่ประมาณค่าสูญหายด้วยวิธีการประมาณค่าข้อมูลสูญหายแบบใหม่ DKNN-BH

ข้อมูลตัวแปร	ลักษณะข้อมูลเบื้องต้น			
	Min.	Max.	Mean	SD
1. ผลผลิตข้าว (แสดตัน)	24.591	63.909	46.004	12.619
2. พื้นที่เพาะปลูก (แสนไร่)	100.102	189.779	145.380	29.940
3. ปริมาณน้ำฝนเฉลี่ย (มิลลิเมตร)	788.592	1764.993	1212.739	261.816
4. ความชื้นสัมพัทธ์เฉลี่ย (%)	69.573	75.407	72.735	1.594
5. อุณหภูมิเฉลี่ย (°C)	25.813	27.573	26.923	.467
6. ราคาข้าวเฉลี่ย (บาท)	4484.800	15552.600	9847.649	3626.675

เมื่อดำเนินการวิเคราะห์การถดถอยพหุคูณโดยใช้วิธี Stepwise ในการคัดเลือกตัวแปร ได้ผลการวิเคราะห์ข้อมูล ดังแสดงในตาราง 4 หลังจากดำเนินการวิเคราะห์การถดถอยพหุคูณ โดยใช้วิธี Stepwise ในการคัดเลือกตัวแปรแล้ว ได้ตัวแบบสำหรับการทำนายผลผลิตข้าวหอมมะลิ ภาคตะวันออกเฉียงเหนือ ดังนี้

$$\hat{y} = b_0 + b_1x_{Cul_area} + b_2x_{RT} \quad (5)$$

$$= -17.532 + 0.393x_{Cul_area} + 0.005x_{RT}$$

การเปรียบเทียบค่าความคลาดเคลื่อนจากการวิเคราะห์การถดถอยพหุคูณ โดยใช้วิธี Stepwise ในการคัดเลือกตัวแปร นำผลการทำนายผลผลิตข้าวหอมมะลิที่ได้จากวิธีดังกล่าวมาเปรียบเทียบกับข้อมูลจริงของผลผลิตข้าวหอมมะลิภาคตะวันออกเฉียงเหนือ จากข้อมูลของสำนักงานเศรษฐกิจการเกษตร โดยพิจารณาจากค่าความคลาดเคลื่อนร้อยละสัมบูรณ์เฉลี่ย (Mean Absolute Percentage Error--MAPE) ในช่วงเวลาเดียวกัน โดยแสดงรายละเอียด ดังนี้

ตาราง 4

ผลการวิเคราะห์การถดถอยพหุคูณ โดยใช้วิธี Stepwise ในการคัดเลือกตัวแปร ได้ปัจจัยที่ส่งผลต่อผลผลิตข้าวหอมมะลิ ภาคตะวันออกเฉียงเหนือ

ตัวแปร	สัมประสิทธิ์การถดถอยจาก คะแนนดิบ ไม่ปรับมาตรฐาน (b)	สัมประสิทธิ์การถดถอยจาก คะแนนมาตรฐาน (β)	t	p
ค่าคงที่	-17.532		-8.045*	.000
พื้นที่เพาะปลูก	0.393	0.932	26.222*	.000
ปริมาณน้ำฝนเฉลี่ย	0.005	0.110	3.090*	.006

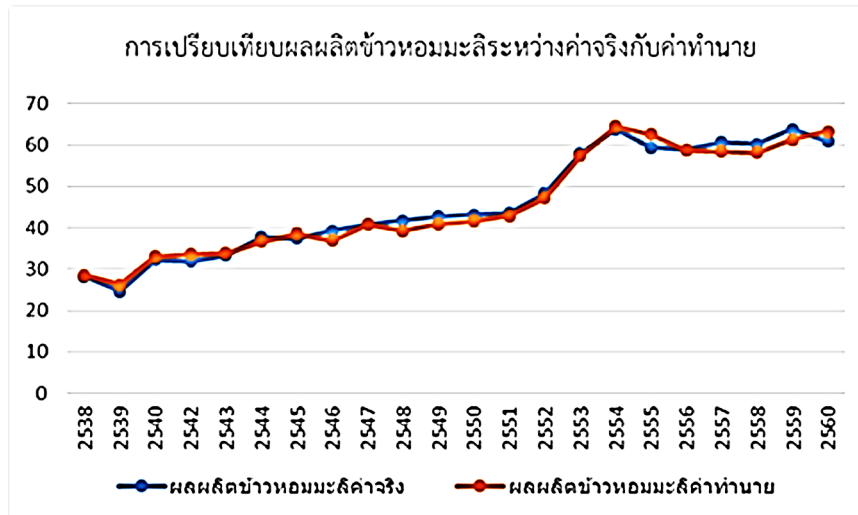
*p<.05, R= 0.991, R²= 0.982, p=.000

ตาราง 5

แสดงค่าทำนายผลผลิตข้าวหอมมะลิโดย วิธีการวิเคราะห์การถดถอยพหุคูณ โดยใช้วิธี Stepwise ในการคัดเลือกตัวแปร

ปี	ผลผลิตข้าวหอมมะลิค่าจริง	ผลผลิตข้าวหอมมะลิค่าทำนาย
2538	28.241	28.491
2539	24.591	26.287
2540	32.269	33.117
2542	31.963	33.664
2543	33.284	33.866
2544	37.795	36.612
2545	37.438	38.576
2546	39.329	36.881
2547	40.751	40.745
2548	41.782	39.320
2549	42.732	40.890
2550	43.156	41.594
2551	43.522	42.932
2552	48.234	47.126
2553	57.931	57.410
2554	63.884	64.560
2555	59.483	62.556
2556	58.838	58.789
2557	60.681	58.461
2558	60.194	58.170
2559	63.909	61.341
2560	60.949	63.267

MAPE เฉลี่ย 22 ปี เท่ากับ 3.155



ภาพ 1 การเปรียบเทียบผลผลิตข้าวหอมมะลิระหว่างค่าจริงกับค่าทำนาย

จากตาราง 5 เมื่อพิจารณาค่า MAPE ของตัวแบบการวิเคราะห์การถดถอยพหุคูณ โดยใช้วิธี Stepwise ในการคัดเลือกตัวแปร พบว่าค่า MAPE เท่ากับ 3.155 และเมื่อเปรียบเทียบผลผลิตข้าวหอมมะลิระหว่างค่าจริงกับค่าทำนาย ในการวิเคราะห์การถดถอยพหุคูณ โดยใช้วิธี Stepwise ในการคัดเลือกตัวแปร วิธีประมาณค่าสูญหายแบบใหม่ DKNN-BH สามารถช่วยประมาณค่าสูญหายในการเตรียมข้อมูลให้มีความสมบูรณ์เพื่อทำนายผลผลิตข้าวหอมมะลิ ภาคตะวันออกเฉียงเหนือ ได้ผลการทำนายใกล้เคียงกับข้อมูลจริง โดยสามารถแสดงได้ดังภาพ 1

การอภิปรายผล

ปัญหาการสูญหายของข้อมูลเป็นปัญหาหนึ่งในขั้นตอนการเตรียมพร้อมของข้อมูลก่อนนำข้อมูลนั้น ๆ ไปวิเคราะห์ การทราบถึงวิธีการในการประมาณค่าสูญหายของข้อมูลที่เหมาะสมจะทำให้ได้ข้อมูลที่มีความสมบูรณ์เหมาะกับการนำไปวิเคราะห์ข้อมูลในขั้นตอนต่อไป

ในการวิจัยในครั้งนี้จึงยึดแนวทางในการวิจัยที่จะพัฒนาวิธีการประมาณค่าข้อมูลสูญหายให้ได้วิธีการที่เหมาะสมสอดคล้องกับข้อมูลที่มี การพัฒนาวิธีการประมาณค่าข้อมูลสูญหายแบบใหม่ DKNN-BH เกิดจากการปรับแก้วิธีการประมาณค่าสูญหายวิธี KNN ด้วยการใช้ค่าเฉลี่ยเดโชล์และการหาระยะทางแบบ Bhattacharyya เพื่อเพิ่มประสิทธิภาพในการประมาณค่าข้อมูลสูญหายด้วยวิธี

KNN ซึ่งผู้วิจัยได้ยึดเป็นแนวทางในการวิจัยในครั้งนี้ตามแนวคิดของ Troyanskaya et al. (2001) โดยยังคงจุดเด่นในขั้นตอนของการประมาณค่าข้อมูลสูญหายด้วยวิธี KNN คือ ขั้นตอนการหาระยะทางเพื่อคงความคล้ายคลึงกันของข้อมูลและแทนค่าในขั้นตอนการแทนค่าซึ่งช่วยในการแก้ปัญหาค่านอกเกณฑ์ ซึ่งสอดคล้องกับผลการศึกษางานของ Malarvizhi and Thanamani (2012) ที่ใช้วิธี KNN ในการหาระยะทางที่ต่ำที่สุด เพื่อหาค่าข้อมูลที่ใกล้เคียงกับค่าสูญหายมากที่สุดซึ่งเป็นการหาความคล้ายคลึงกันของข้อมูล

จากการเปรียบเทียบประสิทธิภาพของวิธีการประมาณค่าข้อมูลสูญหายแบบใหม่ DKNN-BH กับวิธีการประมาณค่าข้อมูลสูญหายอีก 3 วิธี คือ วิธีประมาณค่าด้วยค่าเฉลี่ยเลขคณิต วิธี KNN และวิธี DKNN พบว่า วิธีการประมาณค่าข้อมูลสูญหายแบบใหม่ DKNN-BH มีค่า MSE ต่ำที่สุดหรือมีความแม่นยำสูงสุด ผลการวิจัยชี้ให้เห็นว่า ในการประมาณค่าข้อมูลสูญหายสามารถใช้วิธีการประมาณค่าข้อมูลสูญหายด้วยวิธีแบบใหม่ DKNN-BH กับข้อมูลที่เหมาะสมและมีระดับของการสูญหายอย่างเหมาะสมเพื่อช่วยลดความคลาดเคลื่อนและเพิ่มความแม่นยำในการประมาณค่าข้อมูลสูญหาย (Beretta & Santaniello, 2016) และวิธีการ DKNN-BH ที่นำเสนอได้ผลดีกว่าทุกวิธีอาจเนื่องมาจากจุดเด่นของการหาค่ากลางโดยค่าเฉลี่ยเดโชล์ตามการวิจัยของ Rana (Rana, Siraj-Ud-Doula, 2016)

Midi, & Imon, 2012) ที่มีความแกร่งในข้อมูลที่มีค่านอกเกณฑ์และไม่มีค่านอกเกณฑ์

ข้อเสนอแนะ

ข้อเสนอแนะในการนำผลการวิจัยไปใช้

1. สามารถนำผลการพยากรณ์ที่ได้ไปใช้ในการกำหนดการผลิตข้าวหรือผลผลิตที่เกี่ยวข้อง เพื่อให้มีการจัดการด้านการผลิตให้มีความน่าเชื่อถือมากขึ้น
2. ผู้เกี่ยวข้องกับการใช้ข้อมูลทางการเกษตรควรนำการประมาณค่าที่พัฒนาขึ้นไปประมาณค่าข้อมูลสูญหายเพื่อใช้ในการพยากรณ์ให้เกิดความสมบูรณ์มากยิ่งขึ้น นอกจากนี้หน่วยงานที่ใช้ เนื่องจากปัจจุบันข้อมูลมีขนาดใหญ่ขึ้นหรือที่เรียกว่า Big data ซึ่งรัฐบาลได้มีนโยบายขับเคลื่อนการดำเนินการในเรื่องการจัดการข้อมูลขนาดใหญ่ (big data) ซึ่งเป็นส่วนหนึ่งของแผนพัฒนาดิจิทัลเพื่อเศรษฐกิจและสังคม และแผน Thailand 4.0 การจัดการข้อมูลเหล่านี้จำเป็นอย่างยิ่งที่จะต้องได้ข้อมูลซึ่งมีความแม่นยำและถูกต้อง การประมาณค่าสูญหายก็เป็นวิธีการหนึ่งที่ช่วยในการจัดเตรียมข้อมูลก่อนการนำไปวิเคราะห์เพื่อช่วยในการตัดสินใจในขั้นตอนต่อไป

ข้อเสนอแนะในการทำวิจัยต่อไป

1. ควรศึกษาวิธีการประมาณค่าข้อมูลสูญหายที่จะนำมาประยุกต์ใช้ในการปรับปรุงให้เป็นวิธีการประมาณค่าข้อมูลสูญหายแบบผสมผสานกับวิธีอื่น ๆ เช่น การใช้ค่าเฉลี่ยถ่วงน้ำหนักในแบบต่าง ๆ มาร่วมในการประมาณค่าสูญหายที่พัฒนาขึ้น เป็นต้น

2. ควรศึกษาเพิ่มเติมในการประมาณค่าข้อมูลสูญหายด้วยวิธี DKNN-BH ไปใช้ในกรณีที่มีข้อมูลสูญหายในตัวแปรอิสระ เพื่อเปรียบเทียบประสิทธิภาพของการประมาณค่าข้อมูลสูญหายกับผลที่ได้จากการการศึกษาในครั้งนี้

3. ควรศึกษาเพิ่มเติมตัวแปรหรือปัจจัยที่น่าสนใจอื่น ๆ เช่น สภาพพื้นที่เพาะปลูก สภาพอากาศ มลภาวะ ความกดอากาศ การใส่ปุ๋ย การพรวนดิน เข้ามาพิจารณาเพิ่มสำหรับการสร้างโมเดลการทำนาย โดยใช้กับวิธีการประมาณค่าสูญหายที่พัฒนาขึ้น

กิตติกรรมประกาศ

ขอขอบคุณ มหาวิทยาลัยราชภัฏนครราชสีมา และสำนักงานคณะกรรมการการอุดมศึกษา ที่สนับสนุนทุนการศึกษา และขอขอบคุณสำนักงานเศรษฐกิจการเกษตร และกรมอุตุนิยมวิทยา ที่ให้ความอนุเคราะห์และอำนวยความสะดวกสำหรับข้อมูลที่สำคัญและเป็นประโยชน์ต่องานวิจัยนี้



References

- Beretta, L., & Santaniello, A. (2016). Nearest neighbor imputation algorithms: A critical evaluation. *BMC Medical Informatics and Decision Making*, 16(S3), 74. <https://doi.org/10.1186/s12911-016-0318-z>
- Bhattacharyya, A. (1943). On a measure of divergence between two statistical populations defined by their probability distributions. *Bulletin of the Calcutta Mathematical Society*, 35, 99-109.
- Bishop, C. M. (1995). *Neural networks for pattern recognition*. London: Oxford university press.
- Cartwright, M. H., Shepperd, M. J., & Song, Q. (2003). Dealing with missing software project data. *In Proceedings of the 9th IEEE International Software Metrics Symposium (METRICS'03)* (pp. 1-12). Sydney: IEEE Computer Society.

- Jerez, J. M., Molina, I., García-Laencina, P. J., Alba, E., Ribelles, N., Martín, M., & Franco, L. (2010). Missing data imputation using statistical and machine learning methods in a real breast cancer problem. *Artificial Intelligence in Medicine*, 50(2), 105–115. <https://doi.org/10.1016/j.artmed.2010.05.002>
- Kaiser, J. (2014). Dealing with missing values in data. *Journal of Systems Integration*, 5(1), 42-51. doi: 10.20470/JSI.V5I1.178
- Kim, J. O., & Curry, J. (1977). The treatment of missing data in multivariate analysis. *Sociological Methods & Research*, 6(2), 215–240. <https://doi.org/10.1177/004912417700600206>
- Kulnawin, K., Longpradit, P., Chareanporn, C., & Bhattarakosol, P. (2014). A comparative study of data mining techniques to predict agricultural production: A case study in Thai rice. *KKU Research Journal*, 19(1), 31-43. (in Thai)
- Ladha, L., & Deepa, T. (2011). Feature selection methods and algorithms. *International Journal on Computer Science and Engineering (IJCSE)*, 3(5), 1787-1797. Retrieved from <http://www.enggjournals.com/ijcse/doc/IJCSE11-03-05-051.pdf>
- Little, R. J. A. (1992). Regression with Missing X's: A review. *Journal of the American Statistical Association*, 87(420), 1227-1237, doi: 10.1080/01621459.1992.10476282
- Little, R. J. A., & Rubin, D. B. (1987). *Statistical analysis with missing data*. Hoboken, NJ: Wiley.
- Malarvizhi, M. R., & Thanamani, A. S. (2012). K-Nearest Neighbor in missing data imputation. *International Journal of Engineering Research and Development*, 5(1), 5-7. Retrieved from <http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.375.925&rep=rep1&type=pdf>
- Meesad, P., & Hengprapohm, K. (2008). Combination of KNN-Based feature selection and KNN-Based missing-value imputation of microarray data. *2008 3rd International Conference on Innovative Computing Information and Control, Dalian, Liaoning* (pp. 341-341). Dalian, Liaoning, China: IEEE. doi: 10.1109/ICICIC.2008.635.
- Office of Agricultural Economics. (2019). *Agricultural statistics of Thailand 2018*. Retrieved from <http://www.oae.go.th> (in Thai)
- Pasunon, P., & Nilakorn, P. (2007). Outliers detection in regression analysis by Bhattacharyya statistics. *The Proceeding of 45th Kasetsart University Annual Conference* (pp. 11-18). Bangkok: Kasetsart University. (in Thai)
- Rana, S., Siraj-Ud-Doula, M., Midi, H., & Imon, A. H. M. R. (2012). Decile mean: A new robust measure of central tendency. *Chiang Mai Journal of Science*, 39(3), 478-485. Retrieved from <http://www.thaiscience.info/journals/Article/CMJS/10905266.pdf>. (in Thai)
- Stevens, J. (1992). *Applied multivariate statistics for the social sciences* (2nd ed.). Hillsdale, NJ: Erlbaum.
- Troyanskaya, O., Cantor, M., Sherlock, G., Brown, P., Hastie, T., Tibshirani, R., Botstein, D., & Altman, R. B. (2001). Missing value estimation methods for DNA microarrays. *Bioinformatics (Oxford, England)*, 17(6), 520–525. <https://doi.org/10.1093/bioinformatics/17.6.520>

