

RICE PRODUCTION FORECASTING: A MACHINE LEARNING APPROACH

Surapap Rayanakorn Itsarawadee Hema Apitai Kanchanasuwon Supawee Rattanajitdamrong Sumate Pruekruedee

Bank of Thailand

**The opinions expressed in this work are those of the authors and should not be attributed to the Bank of Thailand or the Puey Ungphakorn Institute for Economic Research.*

ที่มาและความสำคัญ

2

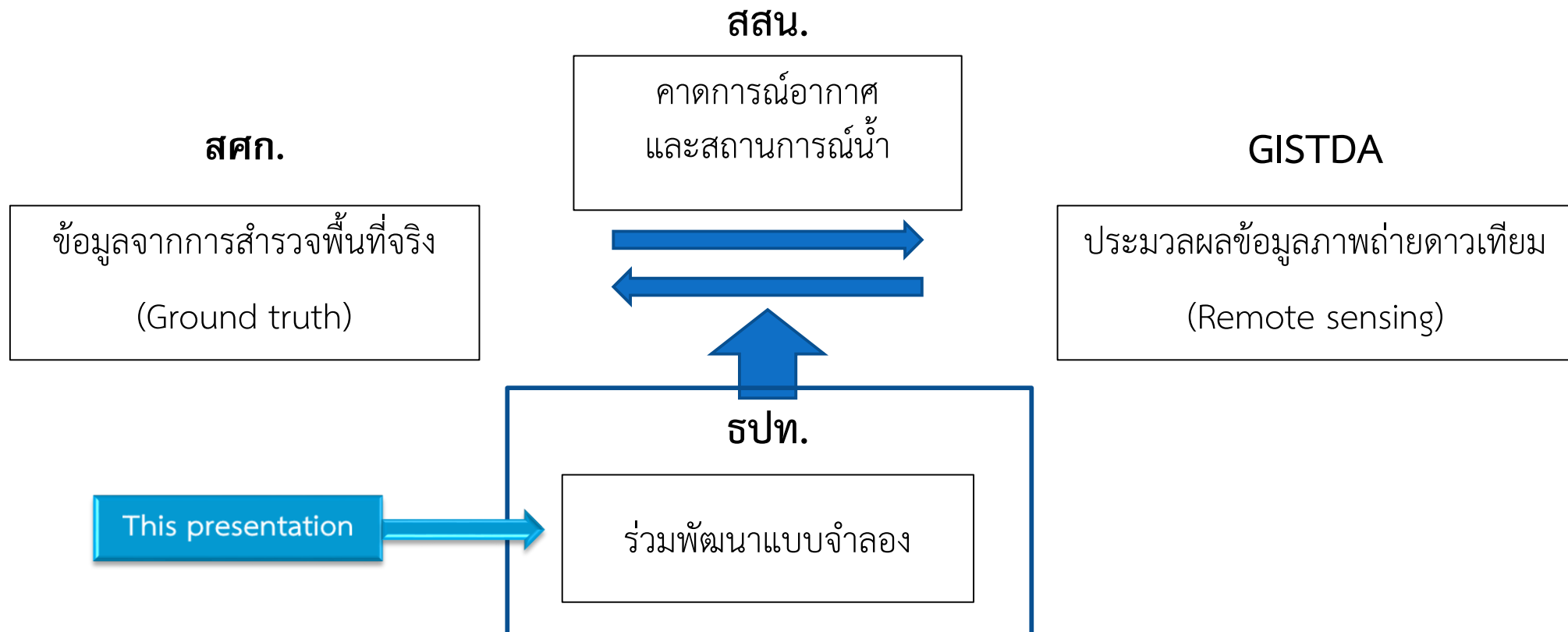
1. ประเมินการผลิตข้าวของทางการเป็นตัวหลักที่ใช้อ้างอิงในการตัดสินใจของภาครัฐและเอกชน
2. ที่ผ่านมา ทางการเผยแพร่ประมาณการผลิตข้าวในรูปแบบภาพรวมประเทศ
3. การมีประมาณการผลิตข้าวในระดับจังหวัดจะช่วย:
 - เพิ่มประสิทธิภาพการบริหารจัดการในระดับพื้นที่
 - สามารถสอบทานตัวเลขสถานการณ์จริงในพื้นที่
4. ปัจจุบัน หน่วยงาน อาทิ สสน. และ GISTDA ได้พัฒนาเครื่องชี้ต่างๆ ในระดับจังหวัดที่อาจช่วยประมาณการผลิตข้าว เช่น พยากรณ์อากาศระยะกลาง ภาพถ่ายดาวเทียมพื้นที่เพาะปลูกและความสมบูรณ์ของพืช
5. การศึกษานี้ทดลองใช้ Machine Learning ประมาณการผลิตข้าวรายจังหวัด โดยใช้ประโยชน์จากเครื่องชี้ที่กล่าวมาข้างต้น

แผนงานบูรณาการระบบงานประมาณการและติดตามภาวะการผลิตข้าว

3

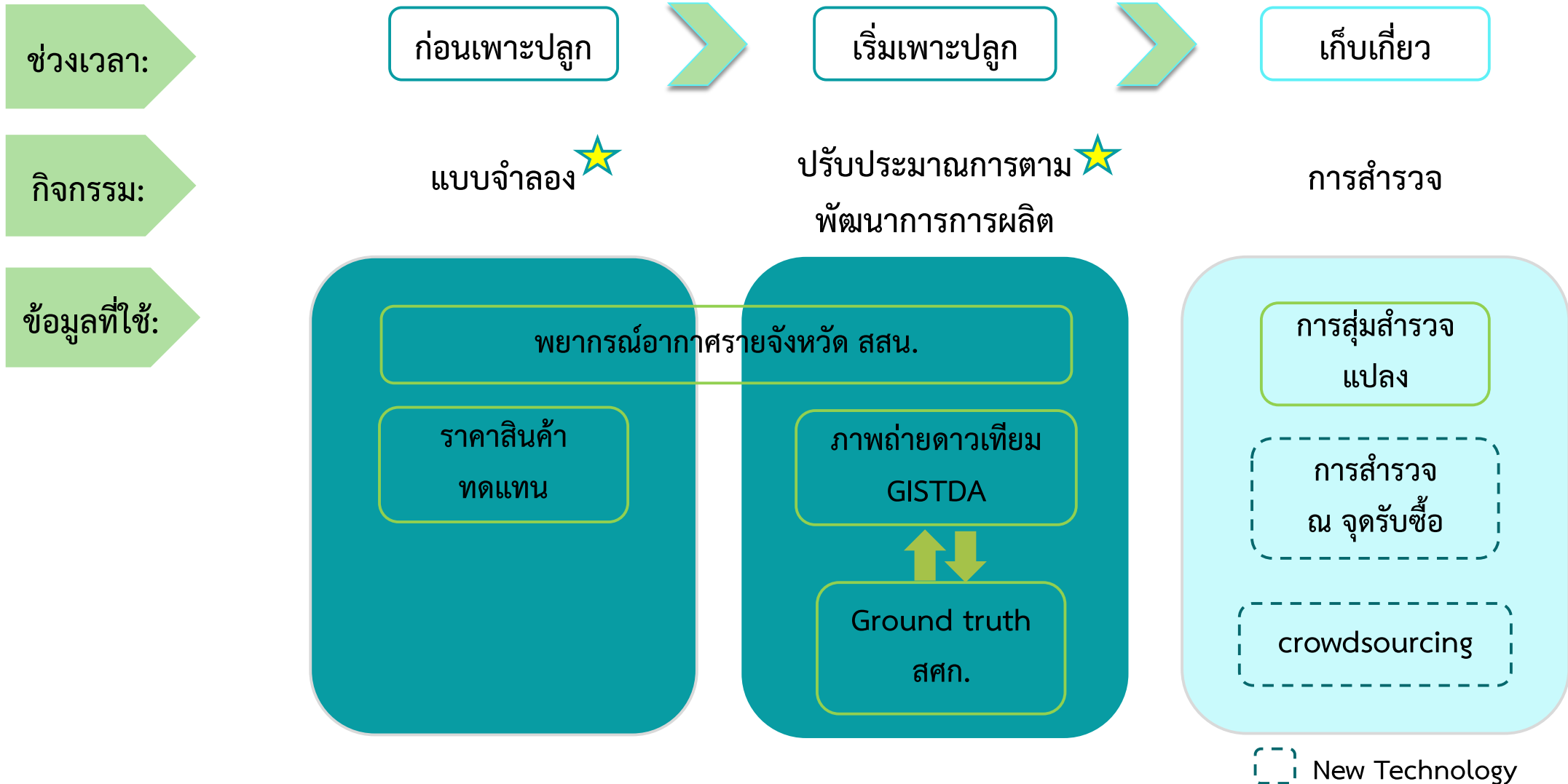
วัตถุประสงค์: เชื่อมโยงระบบงาน สศก. GISTDA และ สสน. เพื่อประมาณการและติดตามผลผลิตข้าว

ประโยชน์: 1. หน่วยงานที่เกี่ยวข้องสามารถนำประมาณการผลผลิตข้าวไปใช้วางแผนอย่างมีประสิทธิภาพ
2. เพิ่มประสิทธิภาพ supply chain ข้าว



แนวทางการเพิ่มประสิทธิภาพการประมาณการและติดตามภาวะการผลิตข้าว

4



Outline

5

- Background and Introduction
- **Research Questions / Literature Review**
- **Overview of Modeling and Data**
- Data Preparation and Feature Engineering
- Methodology
 - ▣ Data Augmentation
 - ▣ Forecasting Models
- Results and Discussion
- Conclusion

Research Questions / Contributions

6

- How do we forecast rice production?
 - ▣ Use traditional as well as alternative data.
 - ▣ Explore econometric and machine learning models → develop most suitable one(s).
- Given optimal forecasting models, what are their economic implications?
 - ▣ What are driving factors of rice production?
 - ▣ Applications of these models: BOT and other agencies' use.

Motivations and Literature Review

7

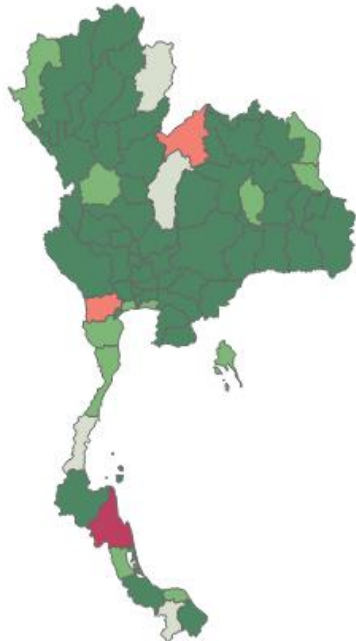
- Crop production economics & modeling
 - ▣ **Agricultural production & risk / Production function estimation**
 - Just and Pope (1979), Moschini and Hennessy (2001), and Chavas and Holt (1996)
 - ▣ **Factors related to crop production**
 - Zhang et al. (2006), Naylor et al. (2001), and others.
- Alternative data, data augmentation, and machine learning
 - ▣ **Spatial data / NDVI, EVI, and other vegetation indices**
 - Huang et al. (2013), Mkhabela et al. (2011), Sun et al. (2017), Yun (2003), and Kouadio et al. (2014)
 - ▣ **Data augmentation**
 - Iwana and Uchida (2020), Rashid and Louis (2019), and Wen et al. (2020)
 - ▣ **Machine Learning: Ensemble Learning, Gradient Boosting**
 - Ideas and algorithms: Breiman (1997), Friedman (1999, 2002), etc.
 - Applications: Varian (2014)

Project Overview

8

□ Modeling in a Nutshell

- Province-level forecasting model
- Data period: 2001-2019
- 69 provinces (out of 74 provinces with off-season rice production)



□ Rice

- **Off-season rice** (ข้าวนาปรัง)
- **In-season rice** (ข้าวนาปี)

□ Forecasting horizons

- **Short-term** (2-4 months before harvesting)
- **Long-term** (one-year ahead)

Horizon	Crops	
	Off-season Rice	In-season Rice
Short-term	1	3
Long-term	2	4

This presentation

↓

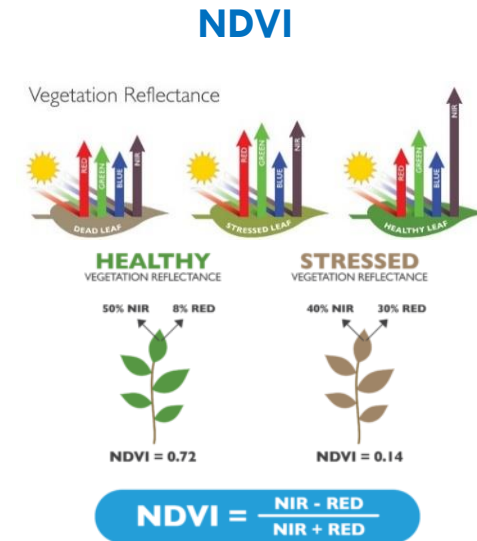
1, 2, 3, 4 are in boxes. A blue rounded rectangle highlights boxes 1 and 2. A green rounded rectangle highlights box 2.

Data Landscape

9

- Both standard and alternative data are helpful in production modeling.

Data <i>*Courtesy of OAE, GISTDA, and HII</i>	Short-term Forecasting Model	Long-term Forecasting Model
Climate data (actual & forecast) - Temperature - Rainfall - Oceanic Nino Index (ONI)	✓	✓
Reservoir (น้ำในเขื่อน)	✓	✓
Plant area	✓	✓ (t-1)
Normalized difference vegetation index (NDVI)	✓	
Prices of alternative crops		✓



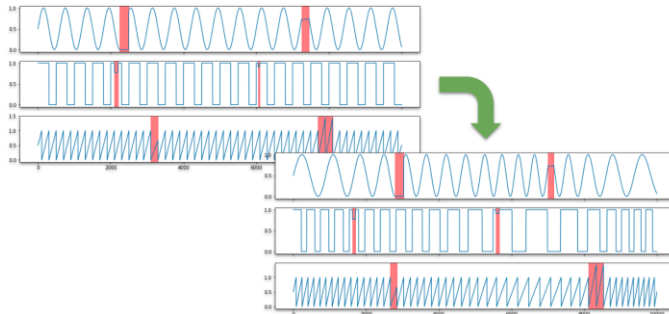
Source: Antognelli (2018)

Modeling Framework

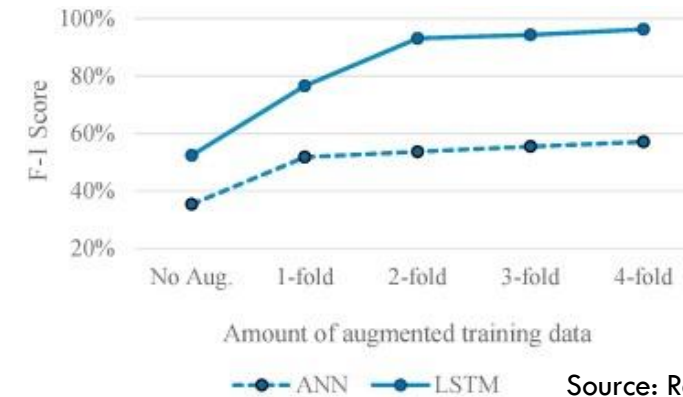
10

Challenge 1: There are few data points (19 points during 2001-2019).

- Overcome small sample data problem with **Data Augmentation**.

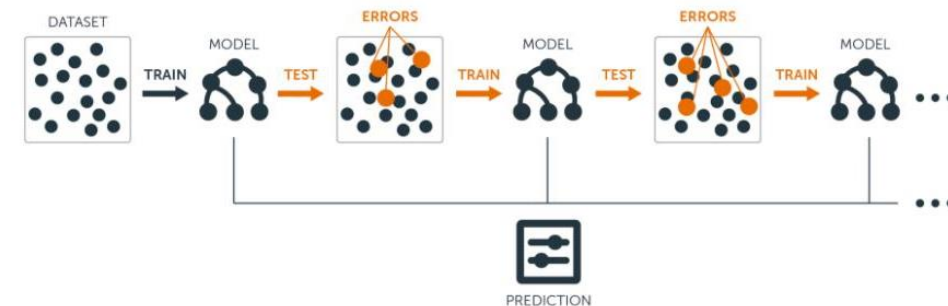


Source: Arundo (2020)



Challenge 2: Identify the best model from a few candidate econometric/machine learning models

- Econometrics: OLS and others.
- Ensemble learning:
 - DRF - regression trees
 - Gradient Boosting**



Source: Kobran and Banyas (2020)

Outline

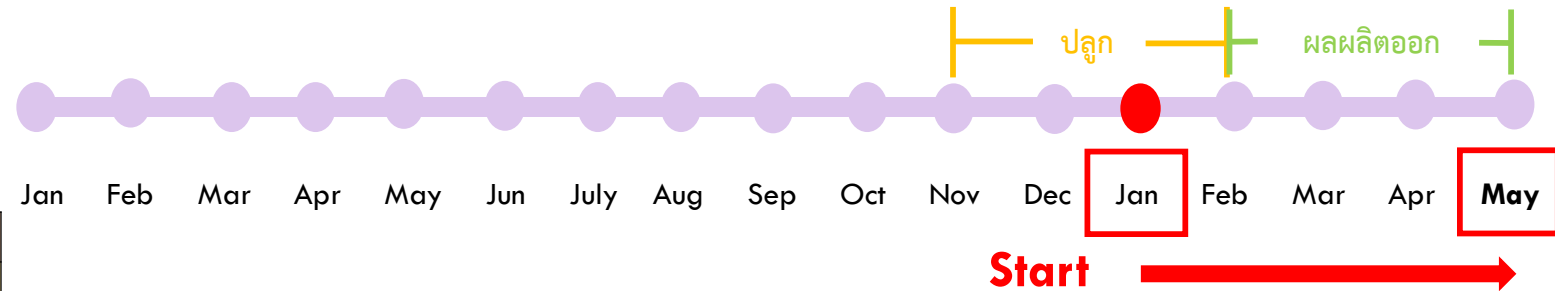
11

- Background and Introduction
- Research Questions / Literature Review
- Overview of Modeling and Data
- **Data Preparation and Feature Engineering**
- **Methodology**
 - **Data Augmentation**
 - **Forecasting Models**
- Results and Discussion
- Conclusion

Feature Selection/Engineering – Short-Term Model

12

Off-Season Rice Calendar



Short-term Model		
	Variables	Period
Climate	- Actual Rainfall & SD	Nov - Jan
	- Forecasted Rainfall & SD	Feb - May
	- Actual Temperature & SD	Nov - Jan
	- Forecasted Temperature & SD	Feb - May
	- ONI Index	Jan
	- Dry Spell	Nov - Apr
	- Days with Rainfall > 35 mm	
Reservoir	- Reservoir	Oct - Jan
Crop Health Index	- NDVI	Nov - Jan
Plant Area	- Plant Area	Most Recent
	- Ratio of Plant Areas	

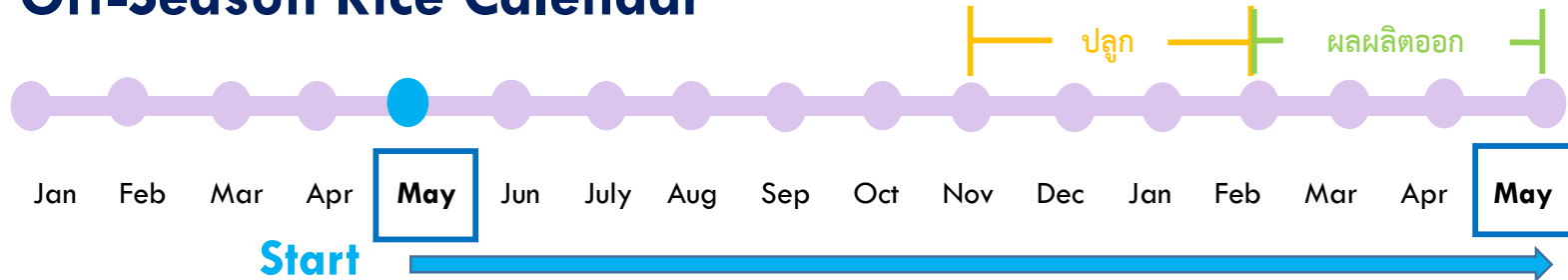
Short Model																
T-1										T						
Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec	Jan	Feb	Mar	Apr	May
										Plant Area _t						
										NDVI						
										Reservoir						
										Actual Rainfall & SD			Forecasted Rainfall & SD			
										Actual Temp & SD			Forecasted Temp & SD			
													Actual ONI			
										Dry Spell						
										Actual RM35mm			Forecasted RM35mm			

*Data Courtesy of OAE, GISTDA, and HII

Feature Selection/Engineering – Long-Term Model

13

Off-Season Rice Calendar



Long-term Model		
Variables		Period
Climate	- Forecast Rainfall & SD	Nov - May
	- Forecast Temperature & SD	
	- Forecast ONI Index	Jan
	- Dry Spell	Nov - Apr
	- Days with Rainfall > 35 mm	
Reservoir	- Reservoir	May
	- Sum of Rainfall Forecast	Jun - Oct
Price	- Rice Price - Price of Alternative Crops (Sugarcane, Cassava, Maize)	Feb - May
Plant Area	- Plant Area	Lagged 1 year

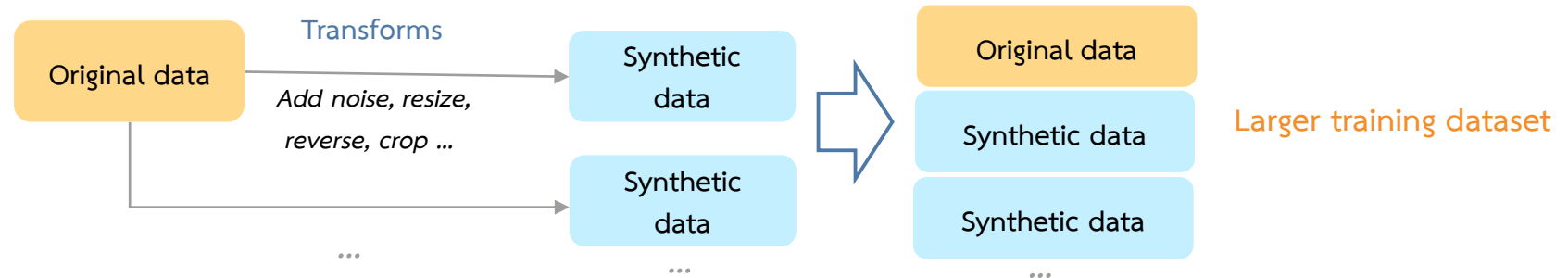
Long Model																
T-1												T				
Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec	Jan	Feb	Mar	Apr	May
Plant Area _{t-1}																
	Prices of Alternative Crops															
				Reservoir	Sum of Rain Forecast											
										Rainfall Forecast & SD						
										Temperature Forecast & SD						
												ONI Forecast				
										Dry Spell Forecast						
										RM35mm Forecast						

Data Augmentation

14

Data Augmentation คือ

เทคนิคในการเพิ่มปริมาณของข้อมูล โดยปรับเปลี่ยนหรือแปรรูปข้อมูลต้นฉบับ (original data) เพื่อสร้างชุดข้อมูลใหม่ (synthetic data)



Objective

- เพิ่มปริมาณและคุณภาพของ training data / Overcome ปัญหาจาก small dataset

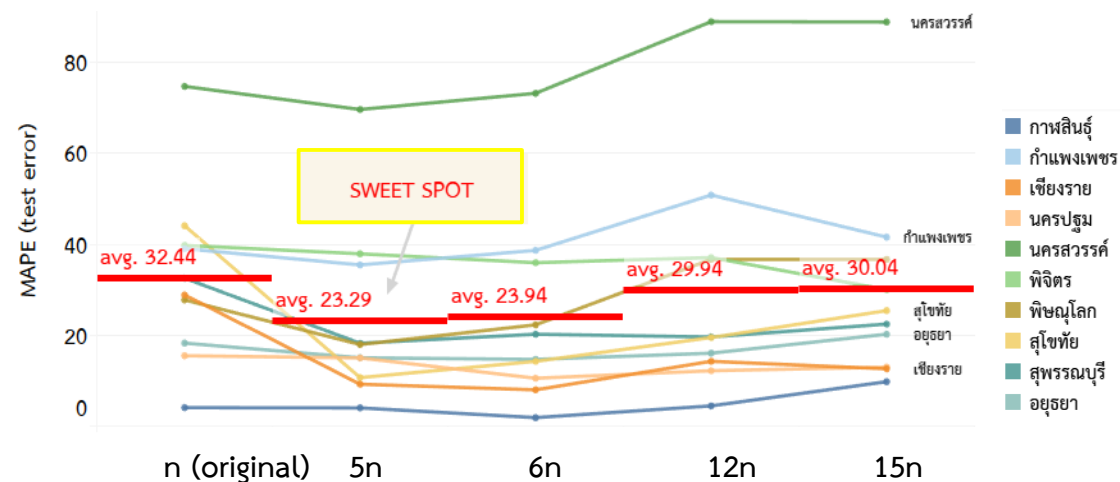
Benefits

- Increasing generalization performance / reduce error

Methodology

1. ใช้วิธี augmentation ที่สามารถรักษา characteristics ของข้อมูลไว้ได้
2. Transforms เฉพาะ training data
3. ทำ pair-wise transformation สำหรับทุกคู่ X และ Y variables
4. ทดลองหา Best setting (sweet spot)

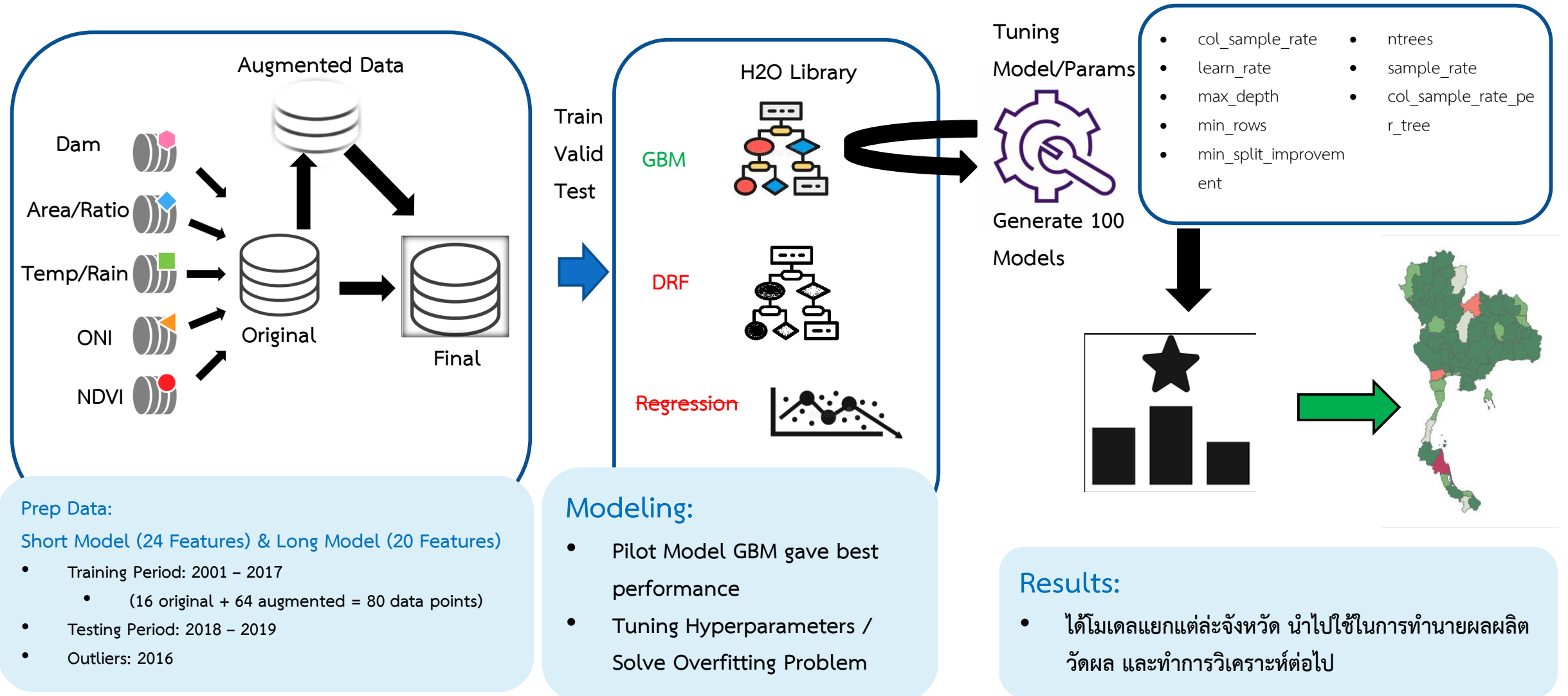
เปรียบเทียบค่า MAPE และปริมาณ augmented training data



* การทดลองใช้ model ที่ยังไม่ผ่านการ tune

Modeling Workflow

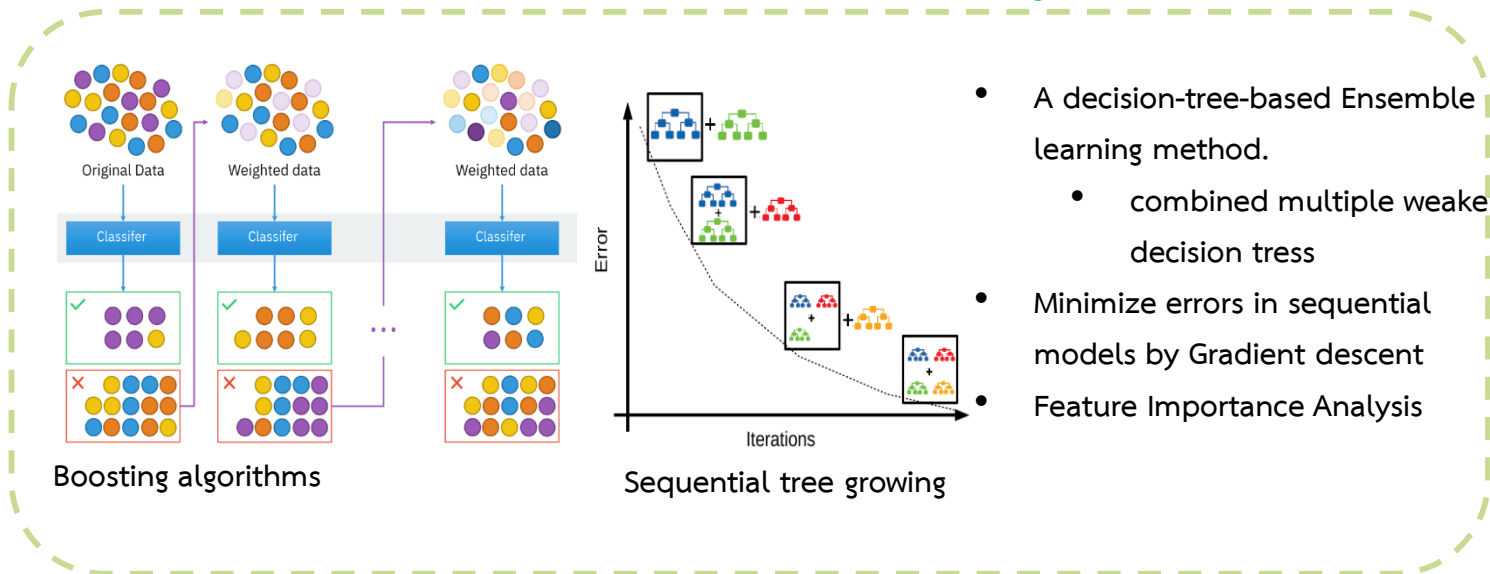
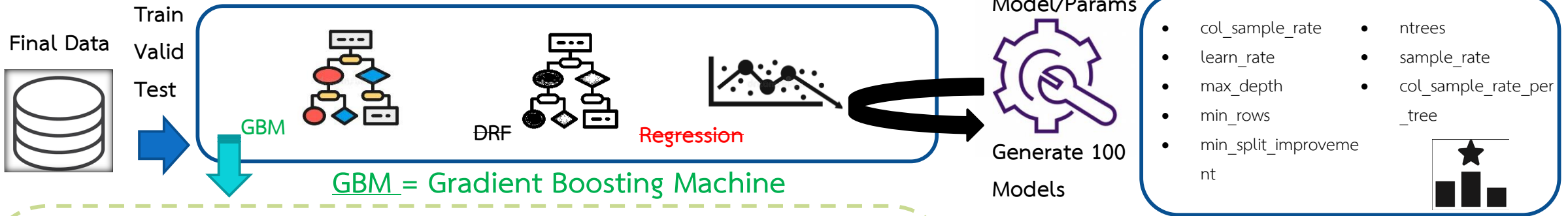
15



Modeling Workflow

16

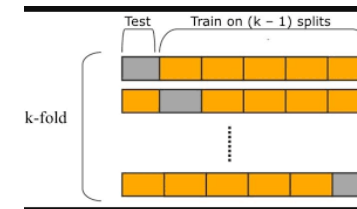
Tuning



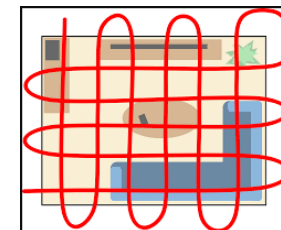
Modeling:

- Pilot Model GBM gave best performance
- Tuning Hyperparameters / Solve Overfitting Problem

Techniques



Cross-validation



Grid Search

Outline

17

- Background and Introduction
- Research Questions / Literature Review
- Overview of Modeling and Data
- Data Preparation and Feature Engineering
- Methodology
 - ▣ Data Augmentation
 - ▣ Forecasting Models
- **Results and Discussion**
- Conclusion

Overview of Short-term Model Outcome

Analysis on error statistics and distribution

18

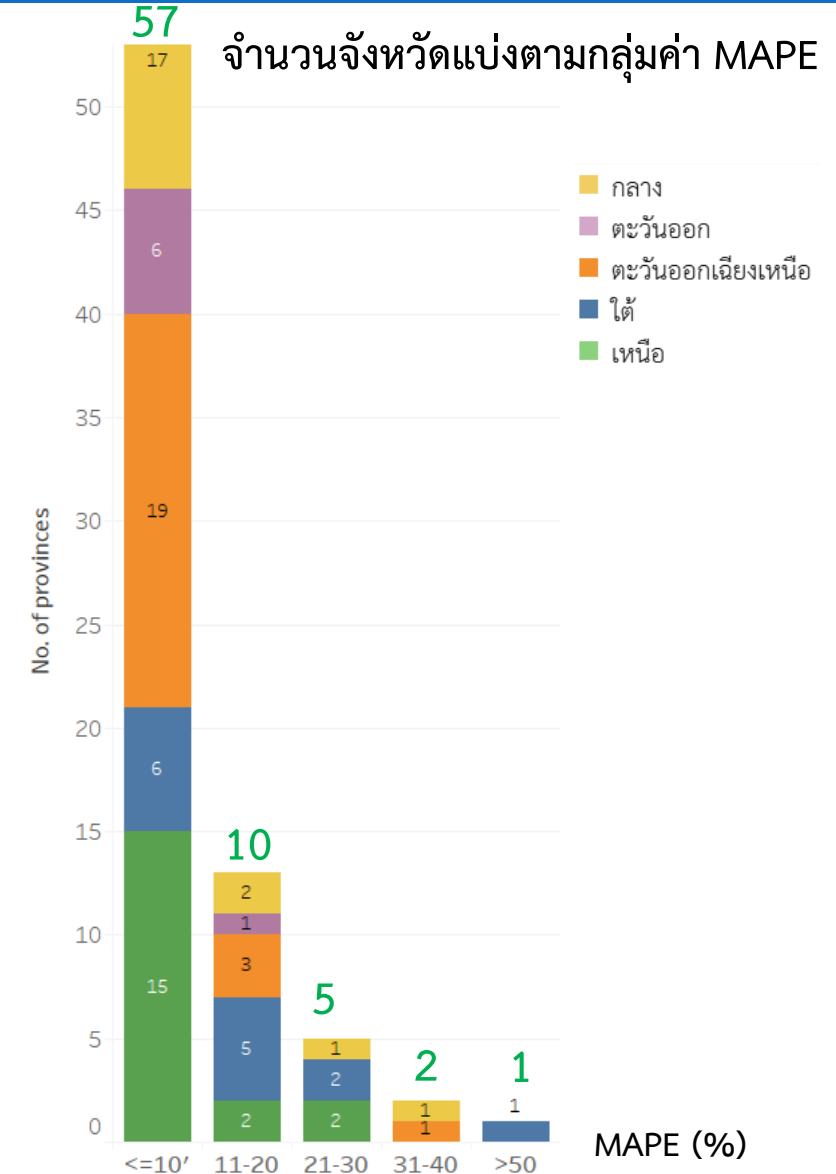
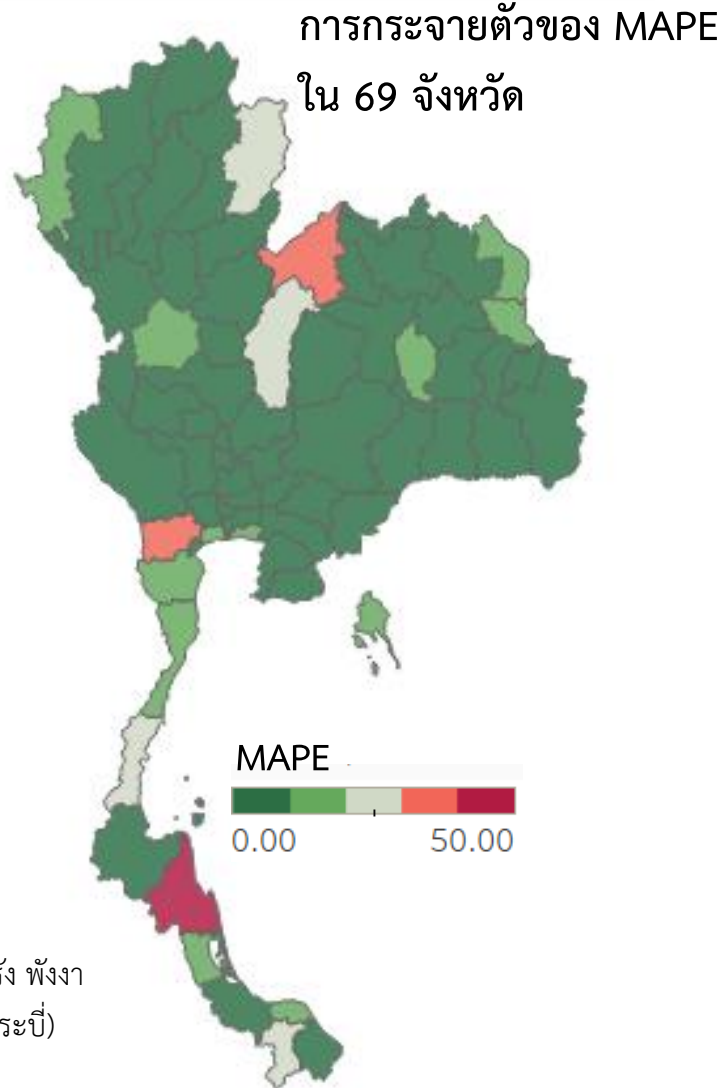
MAPE distribution

No. of provinces: 69

3rd quartile: 7.1%

Median: 3.3%

1st quartile: 2.2%



Short-term Model Outcome

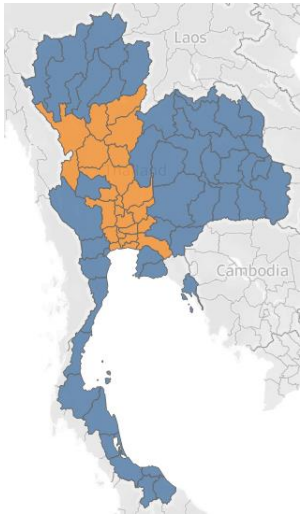
Analysis on error with respect to the level of rice production contribution

19

การกระจายตัวของปริมาณผลผลิตข้าวนาปรังเฉลี่ยปี 2018-2019 และค่า MAPE

จังหวัดที่ติดลุ่มน้ำ

(มีการเพาะปลูกมาก – 22 จังหวัด)



MAPE 3.7%

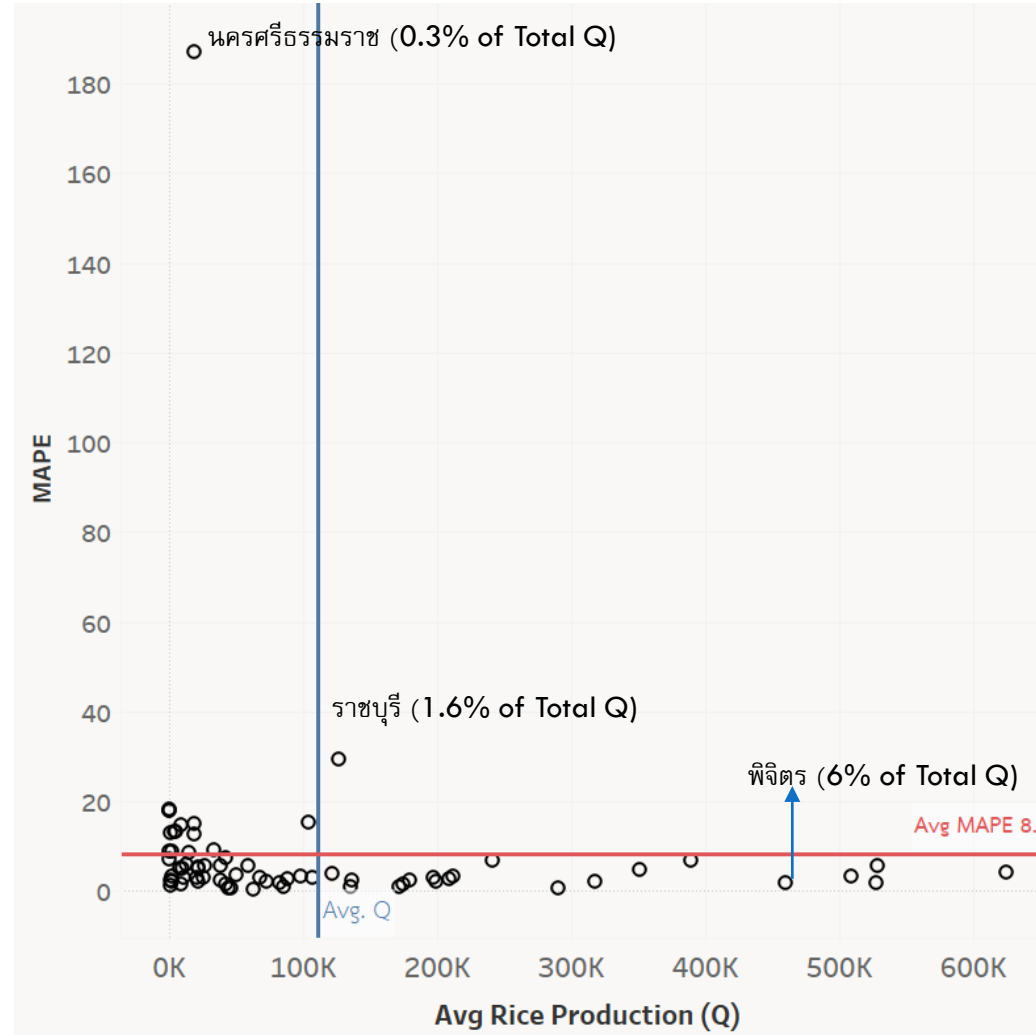
ผลผลิต (Q) 72.8%

MAPE 10.5%

ผลผลิต (Q) 27.2%

จังหวัดที่ไม่ติดลุ่มน้ำ

(มีการเพาะปลูกน้อย – 47 จังหวัด)



MAPE

Maximum: 187.1%

3rd quartile: 7.1%

Average: 8.3%

Median: 3.3%

1st quartile: 2.2%

Minimum: 0.3%

โมเดลสามารถทำนายจังหวัดที่มีค่าผลผลิตสูง
ได้แม่นยำกว่าจังหวัดที่มีผลผลิตน้อย

53 provinces <= Average MAPE

95.3% of Avg. Total Q

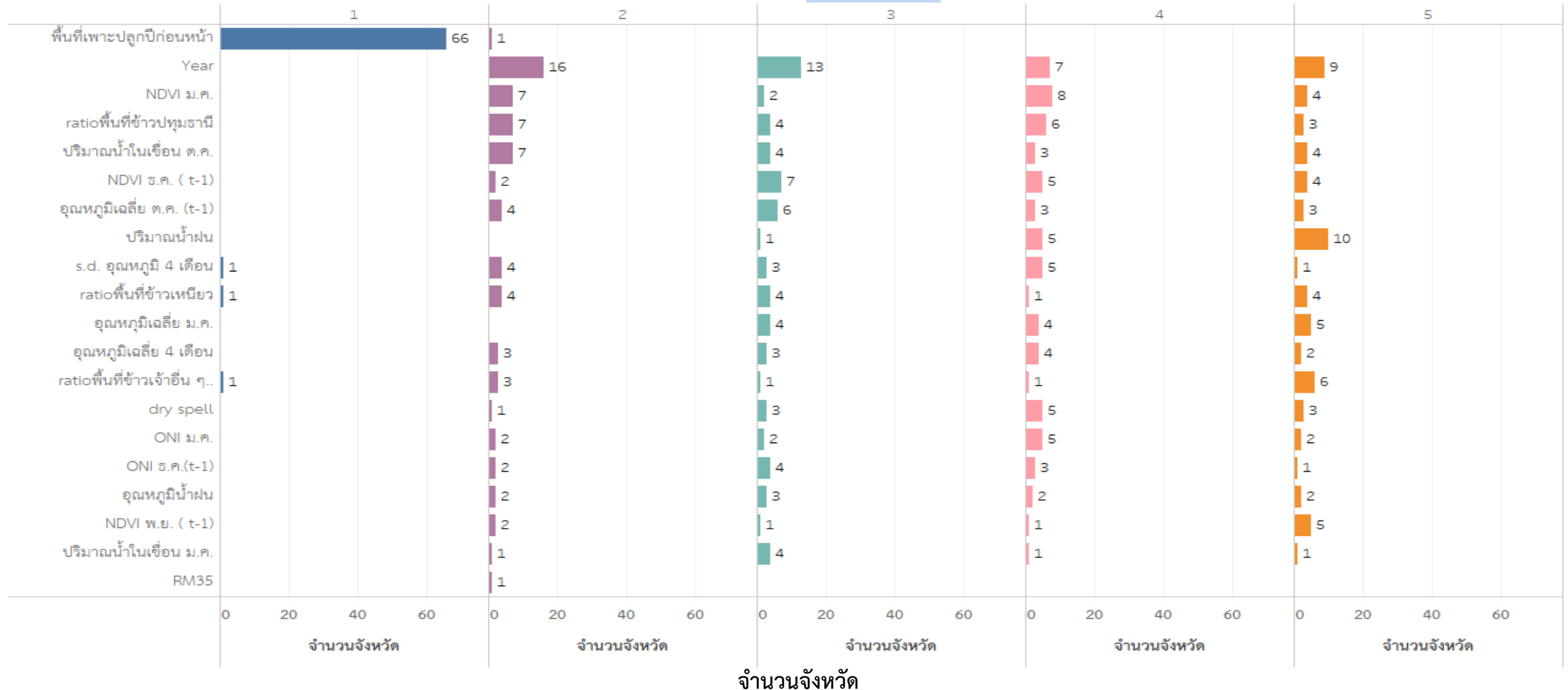
Variable Importance (Short-term Model)

Identifying factors influencing off-season rice production

20

20 variables

ลำดับ



Overview of Long-term Model Outcome

Analysis on error statistics and distribution

21

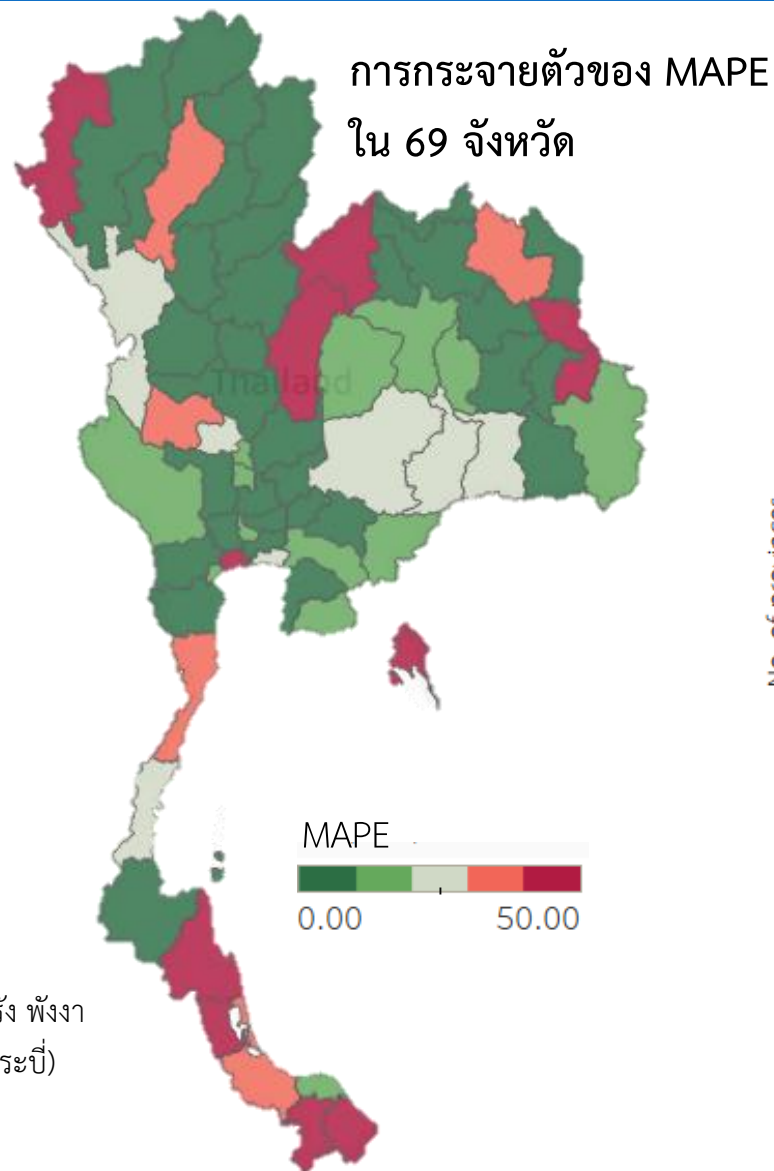
MAPE distribution

No. of provinces: 69

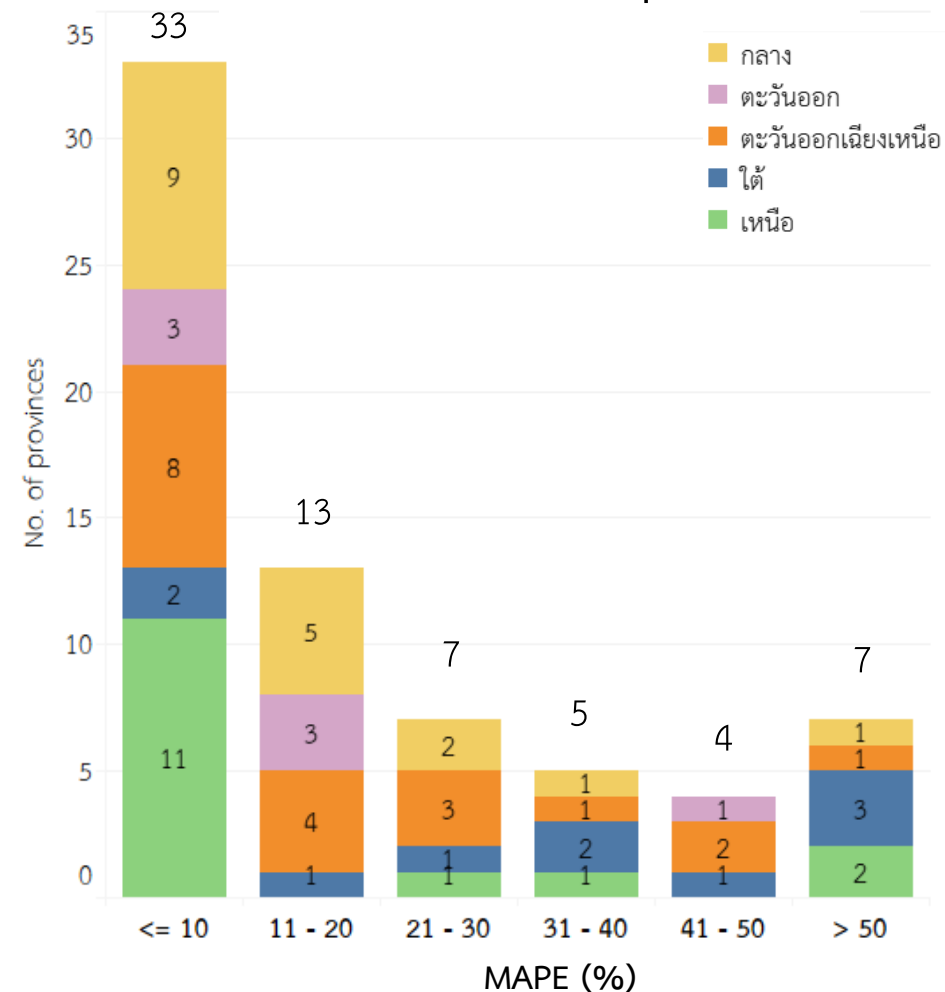
3rd quartile: 25.3%

Median: 11.2%

1st quartile: 4.5%



จำนวนจังหวัดแบ่งตามกลุ่มค่า MAPE



Long-term Model Outcome

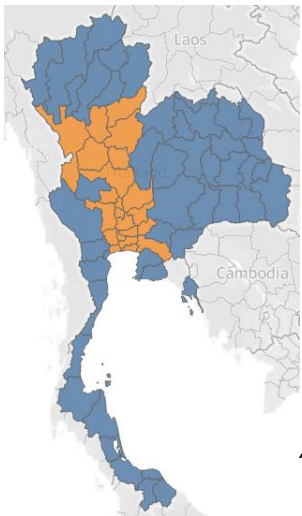
Analysis on error with respect to the level of rice production contribution

22

การกระจายตัวของปริมาณผลผลิตข้าวนาปรังเฉลี่ยปี 2018-2019 และค่า MAPE

จังหวัดที่ติดลุ่มน้ำ

(มีการเพาะปลูกมาก 22 จังหวัด)



สัดส่วน MAPE 15%

Avg. MAPE 13.31%

สัดส่วนผลผลิต 72.8%

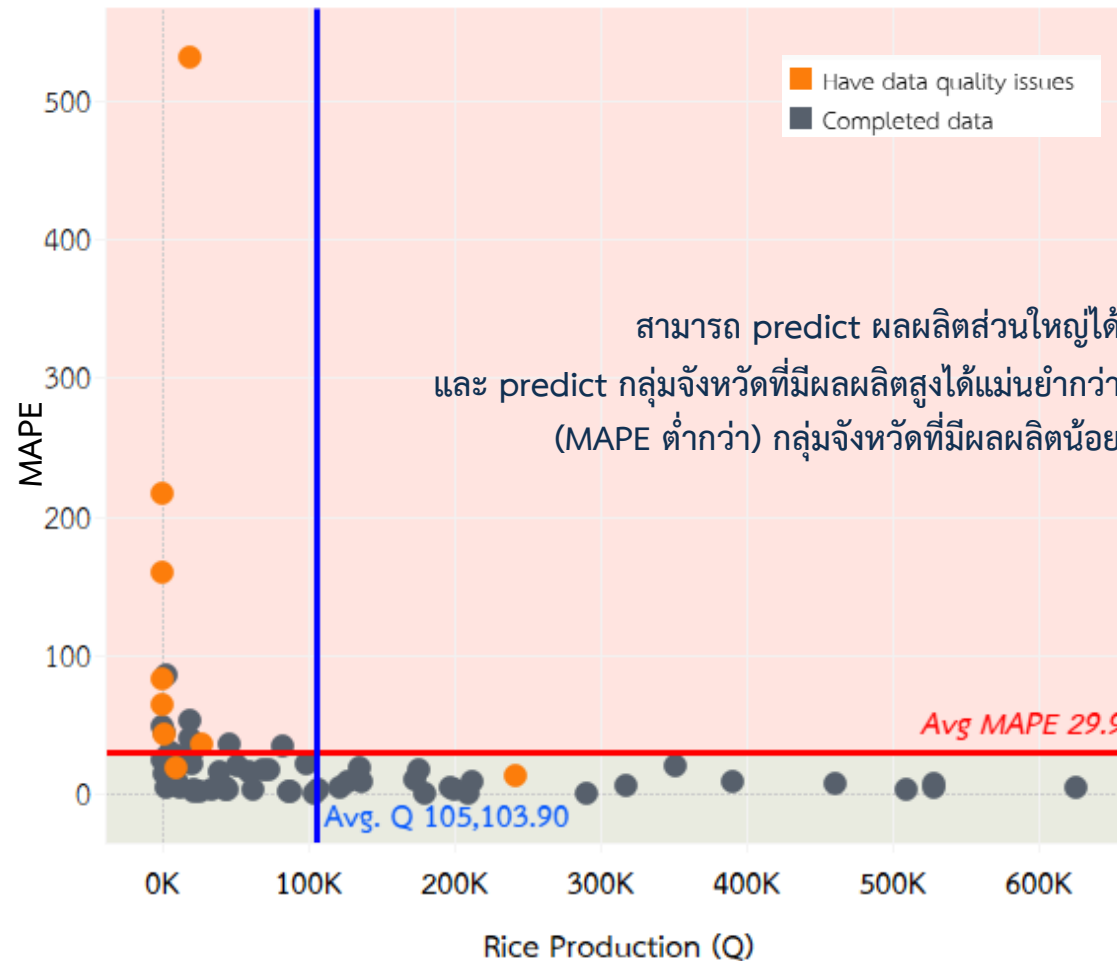
สัดส่วน MAPE 85%

Avg. MAPE 37.7%

สัดส่วนผลผลิต 27.2%

จังหวัดที่ไม่ติดลุ่มน้ำ

(มีการเพาะปลูกน้อย 47 จังหวัด)



	MAPE	Average Q (ton)
Maximum	531.3%	624,889
3 rd Quartile	25.3%	135,336
Average	29.9%	105,104
Median	11.2%	40,802
1 st Quartile	4.5%	8,557
Minimum	0.08%	49

53 provinces (76.8%)

96.7% of Avg. Total Q

จังหวัดที่มี Error สูงส่วนใหญ่มี missing values และปัญหาด้าน data quality (9 จังหวัด)

1. ฉะเชิงเทรา 2. นราธิวาส 3. ปัตตานี 4. มุกดาหาร 5. แม่ฮ่องสอน 6. ยะลา 7. สงขลา 8. อำนาจเจริญ 9. นครศรีธรรมราช

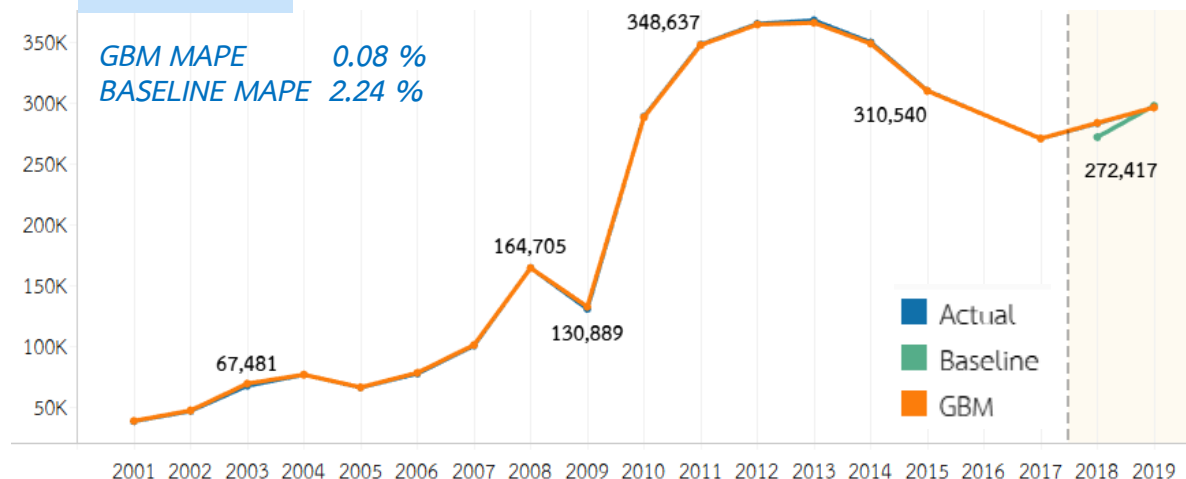
Long-term Model Predictions (province-level)

GBM's predicted rice production values in 2018 and 2019 in comparison with those of baseline model

23

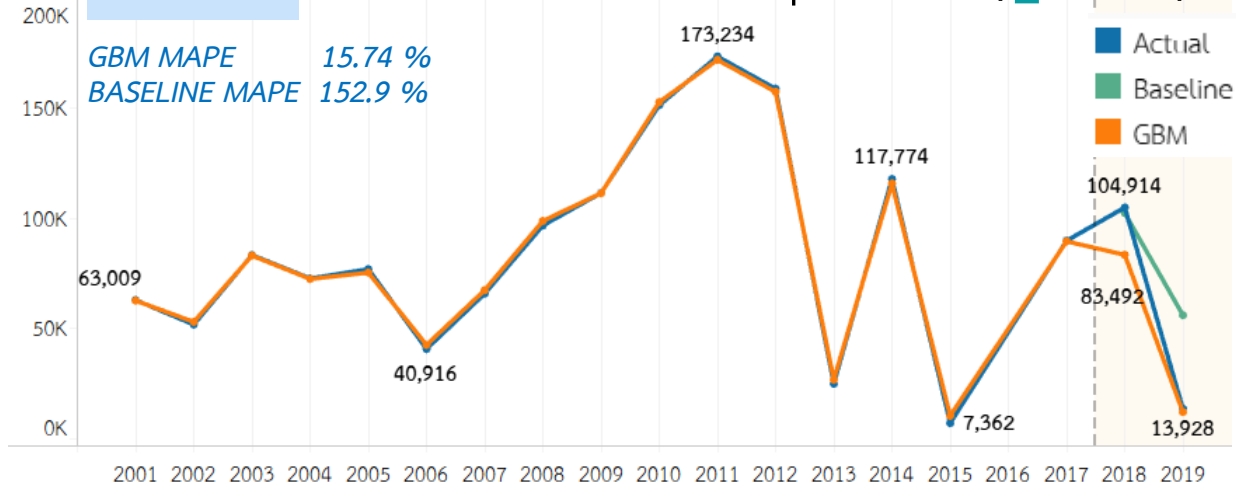
เชียงราย

Minimum MAPE



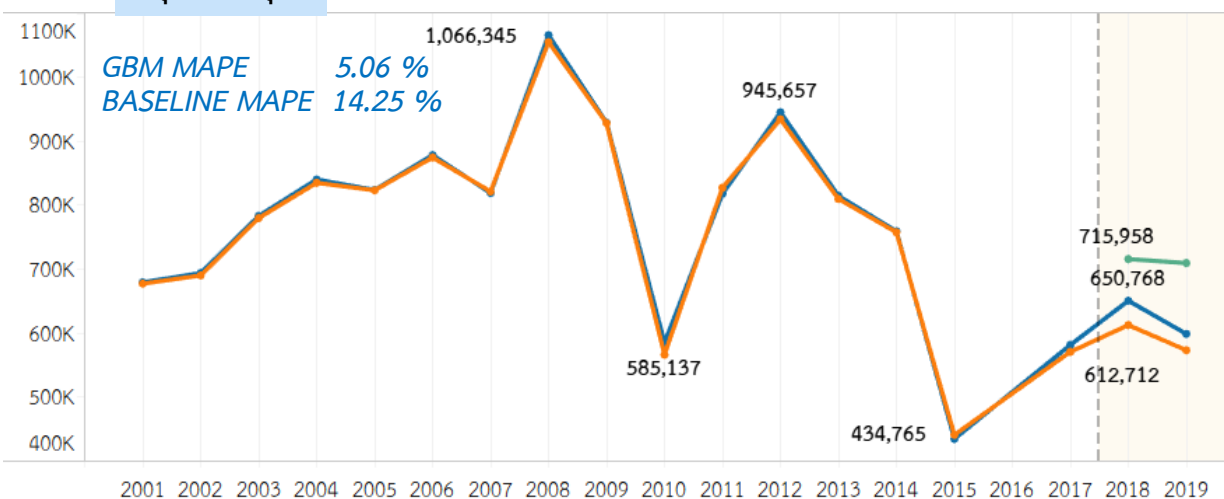
มหาสารคาม

Province with best MAPE improvement (↑137.2%)



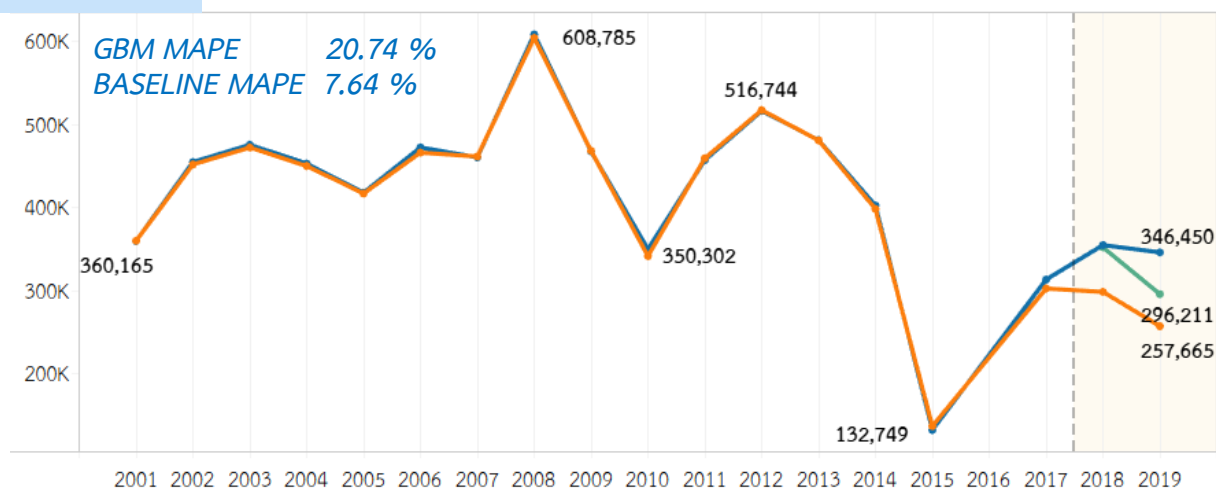
สุพรรณบุรี

Province with the highest production (8.3% of Q)



ชัยนาท

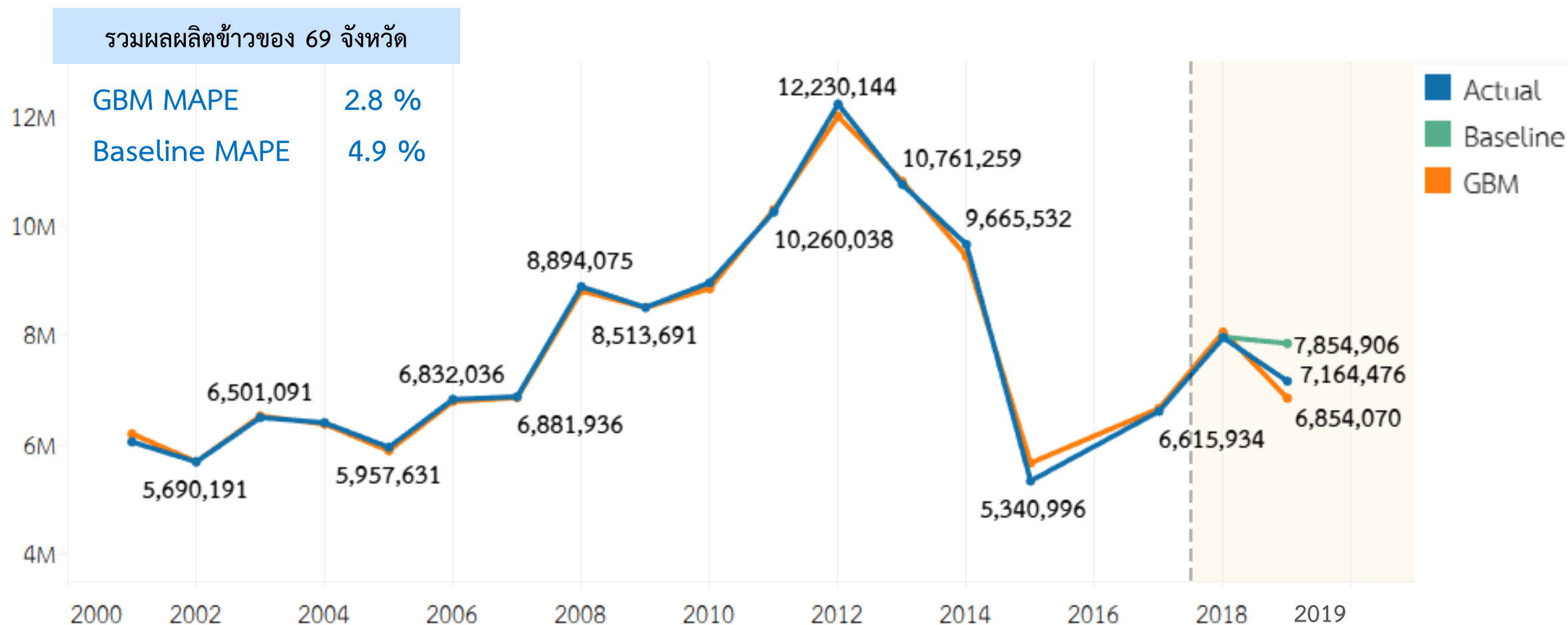
Province with moderate error and high production (4.6% of Q)



Long-term Model Predictions (aggregate-level)

GBM's predicted rice production values in 2018 and 2019 in comparison with those of baseline model

24



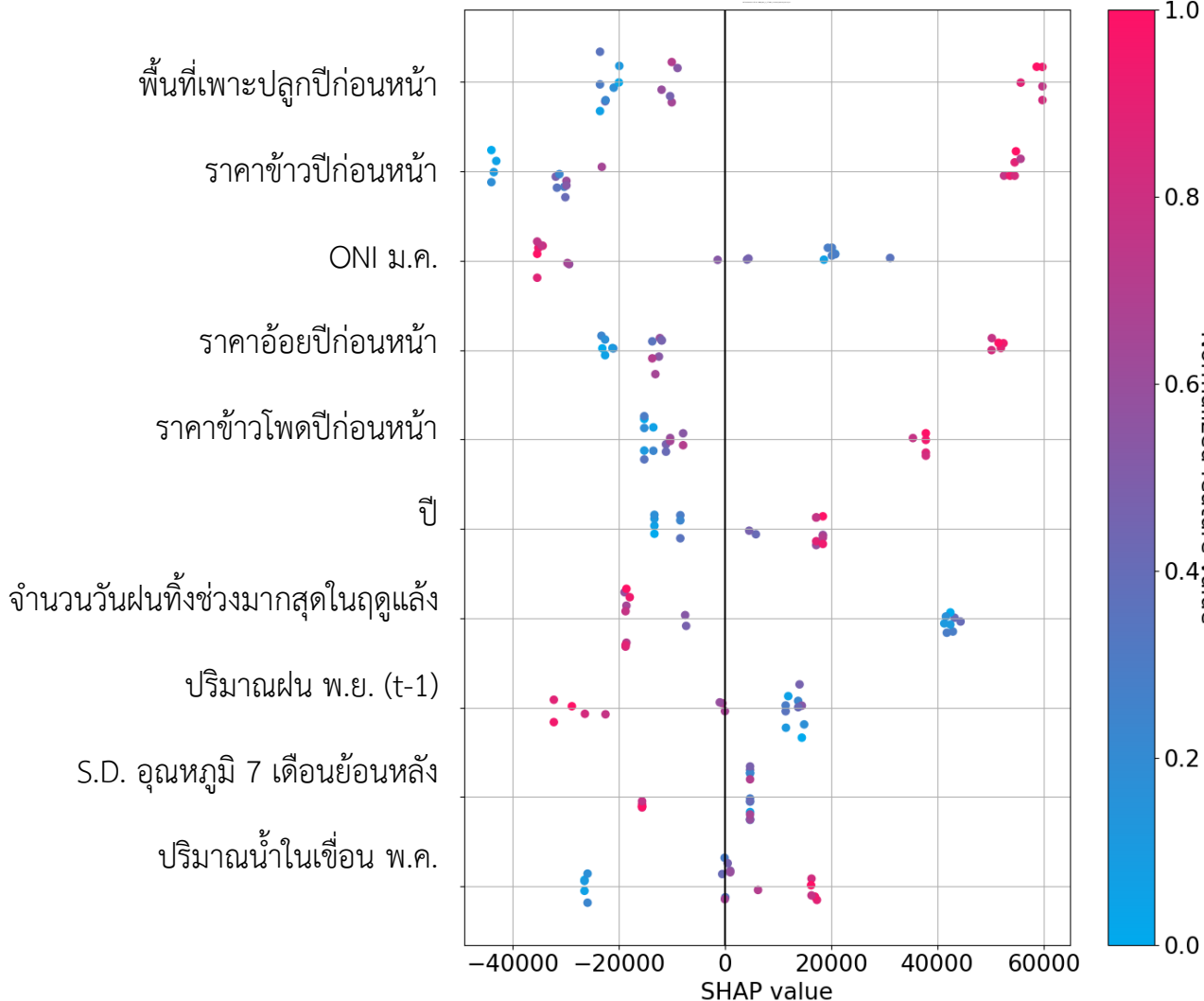
Interpreting The Impact of Variables (Long-term Model)

Using SHAP summary plot to explain variable's responsibility for a change in rice production

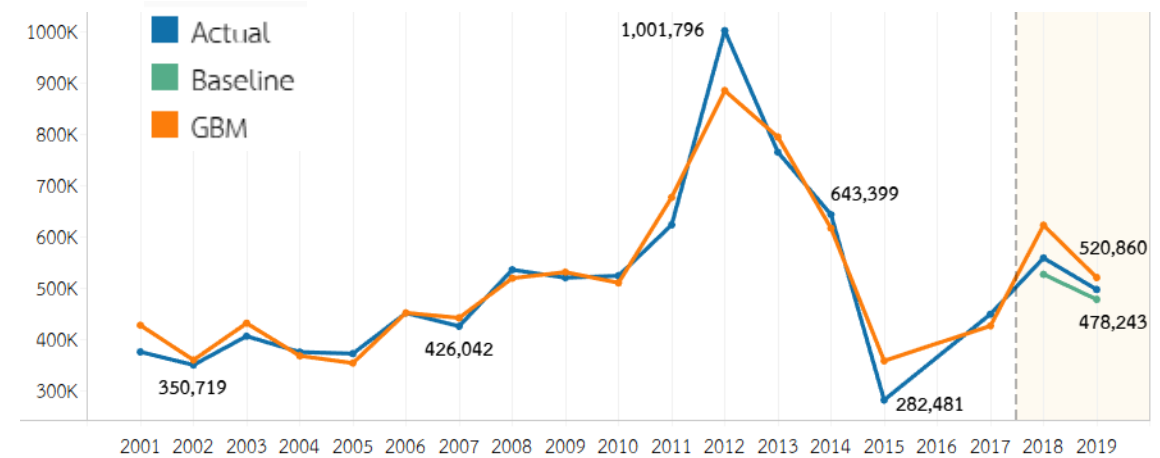
25

SHAP Summary Plot

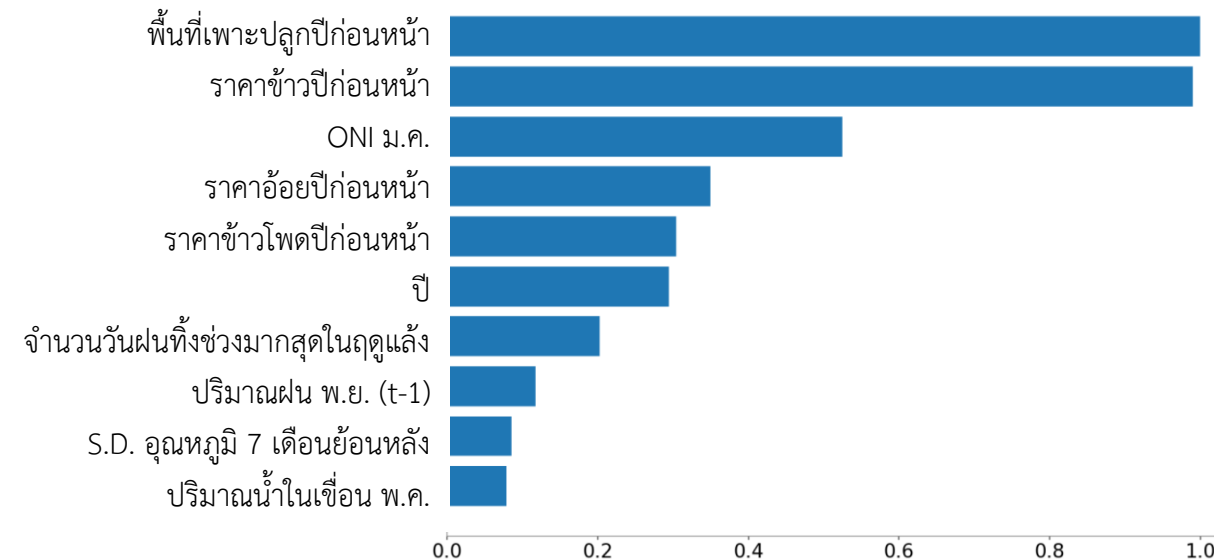
พิษณุโลก



Rice Production Trend



Variable Importance



Conclusion

26

□ Model Development and Improvement

- ▣ การใช้เทคนิค 1) Data Augmentation 2) Machine Learning ประกอบกับ 3) ข้อมูลทางเลือกเชิงกายภาพ ช่วยเพิ่มความแม่นยำของแบบจำลองอย่างมีนัยสำคัญ

□ Implications / Driving Factors

- ▣ Short-term model: พื้นที่เพาะปลูกเป็นตัวแปรที่สำคัญที่สุดในการประมาณการผลผลิต และตัวแปรเชิงกายภาพอื่น ๆ มีส่วนช่วยในการอธิบายความสัมพันธ์
- ▣ Long-term model: ราคาข้าว พืชแข่งขัน และพื้นที่เพาะปลูกในปีที่ผ่านมา เป็นตัวแปรที่มีความสำคัญในการพยากรณ์ ขณะที่ข้อมูลเชิงกายภาพมีส่วนช่วยในระดับหนึ่ง

□ Potential Applications

- ▣ พัฒนาต่อเนื่องเพื่อใช้ในงานประมาณการผลผลิตของ ธปท. และหน่วยงานที่เกี่ยวข้อง

Acknowledgement

27

- สำนักงานเศรษฐกิจการเกษตร สำนักงานพัฒนาเทคโนโลยีอวกาศและภูมิสารสนเทศ และสถาบันสารสนเทศทรัพยากรน้ำ
 - ▣ OAE – ดร.กาญจนา ขวัญเมือง คุณอัญญดา เพ็ญพร และ คุณศิริรัตน์ แก้วกำ
 - ▣ GISTDA - คุณภาณุ เนื่องจำนงค์
 - ▣ HII – ดร.สุรเจตส์ บุญญาอรุณเนตร ดร.กนกศรี ศรีนนภากร และ ดร.วินัย เขาวนวิวัฒน์
- Structural Economic Policy Department
 - ▣ ดร.จิตเกษม พรประพันธ์ คุณจิรพรรณ โอฬารธนาเศรษฐ์ คุณสมบุรณ์ หวังวณิชพันธุ์ คุณจิรัฐ เจนพึงพร และ ดร.มณฑล กปิลกาญจน์
- Data Analytics Office
 - ▣ ศ.ดร.วันประชา เขาวลิตวงศ์ คุณสุบิน เลี้ยงพันธุ์สกุล ดร.โสภณ ัญญาเวชกิจ และ คุณธนัชชนม์ จงมั่นคง
- PIER
 - ▣ ดร.โสมาตรมี จันทรรัตน์ จันทรวิไลศรี และ ดร.ลัทธพร รัตนวรารักษ์