Imbalanced Datasets and Crop Yield Prediction

Application of Preprocessing Techniques for Regression Tasks in Agriculture

July 31st, 2024

ASABE Annual International Meeting 2024 – Anaheim, California

Mariaelisa Polsinelli[1,2], Morteza Mesbah[2], Zhiming Qi[1], Matt Ramsay[3]

1. McGill University, Department of Bioresource Engineering, Montreal, Quebec, Canada H9X 3V9

2. Agriculture and Agri-food Canada, Charlottetown Research and Development Centre, Charlottetown, Prince Edward Island, Canada C1A 4N6

3. Prince Edward Island Farmers' Association and Oyster Cove Farms, Charlottetown, Prince Edward Island, Canada C1E 0E5

mariaelisa.polsinelli@mail.mcgill.ca

McGill

Agriculture and Agri-Food Canada

Agriculture et Agroalimentaire Canada

Introduction

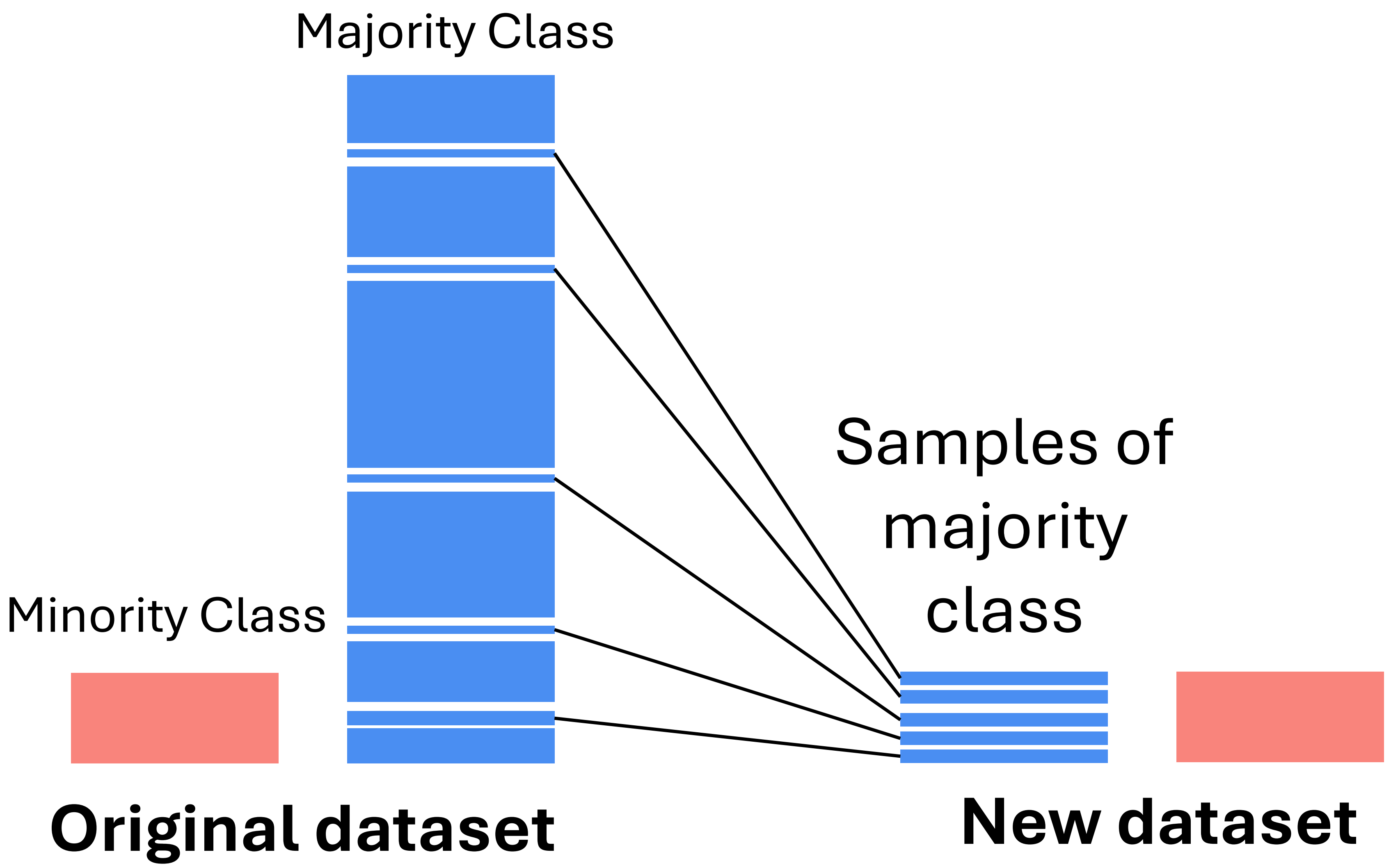
- Increase in drought and extreme weather years

- Limited data

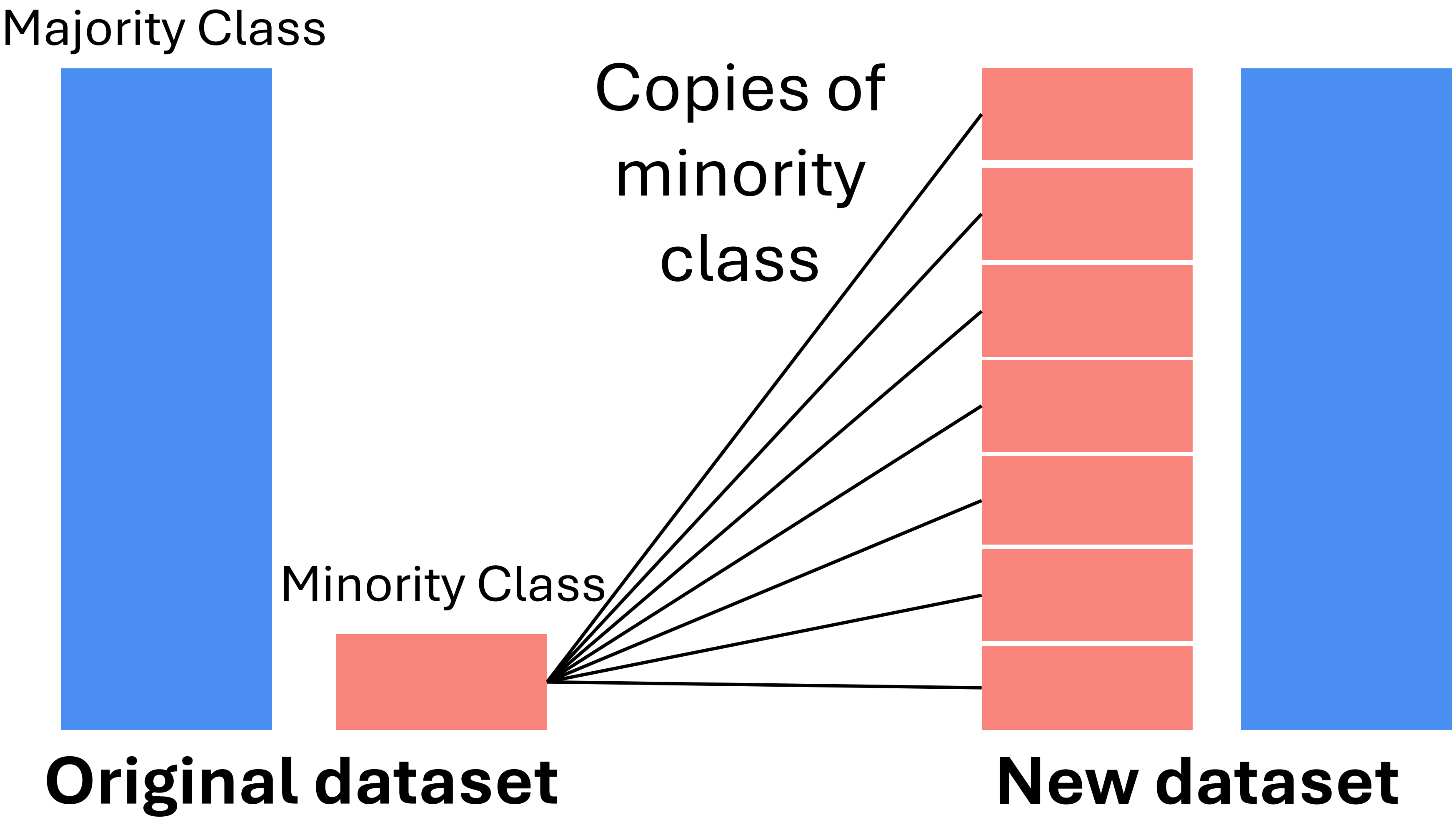
- Unbalanced datasets

- Poorer Predictions

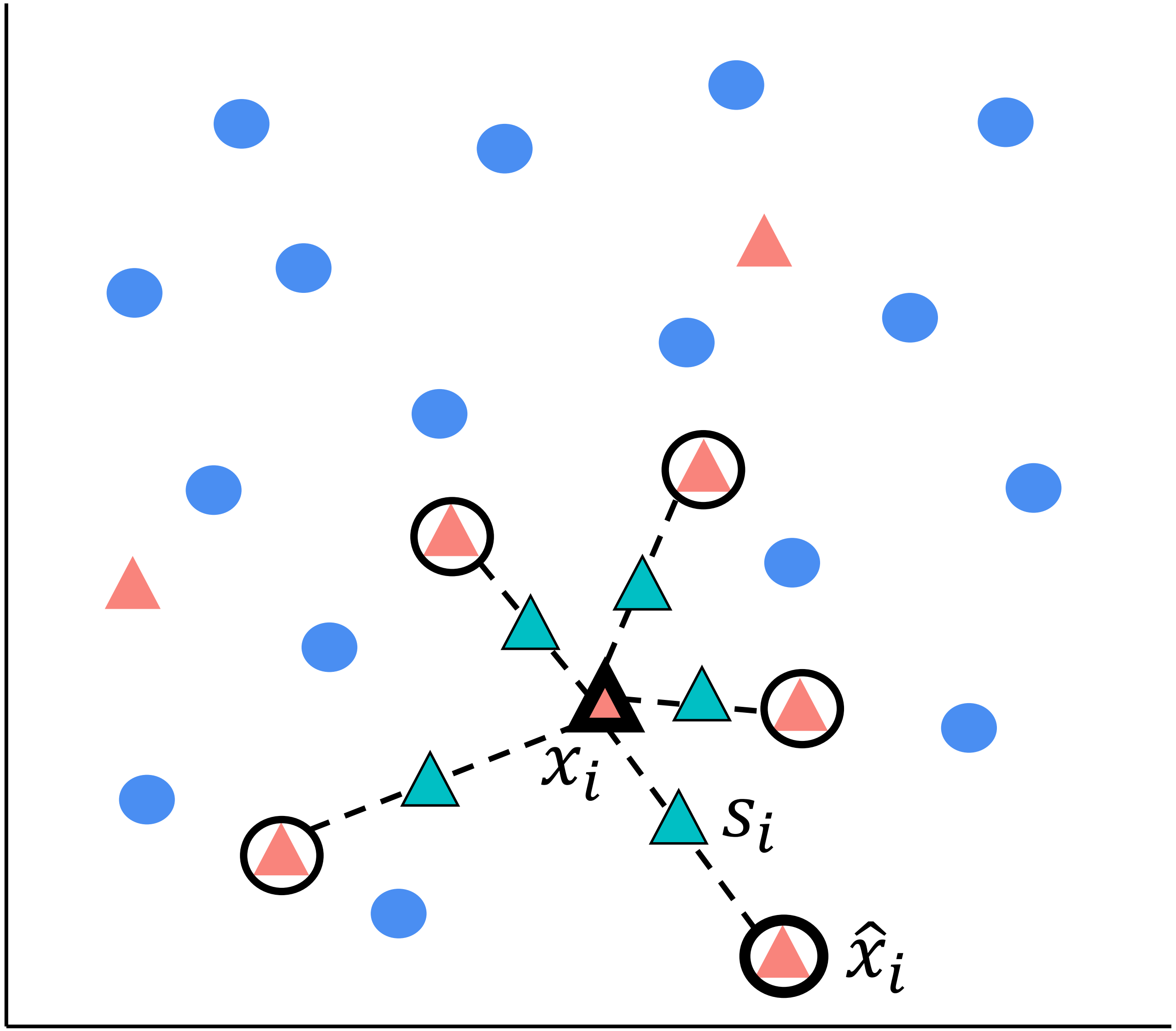
Undersampling (US)



Oversampling (OS)



SMOTER (synthetic minority oversampling technique regression)



Majority class

Minority class

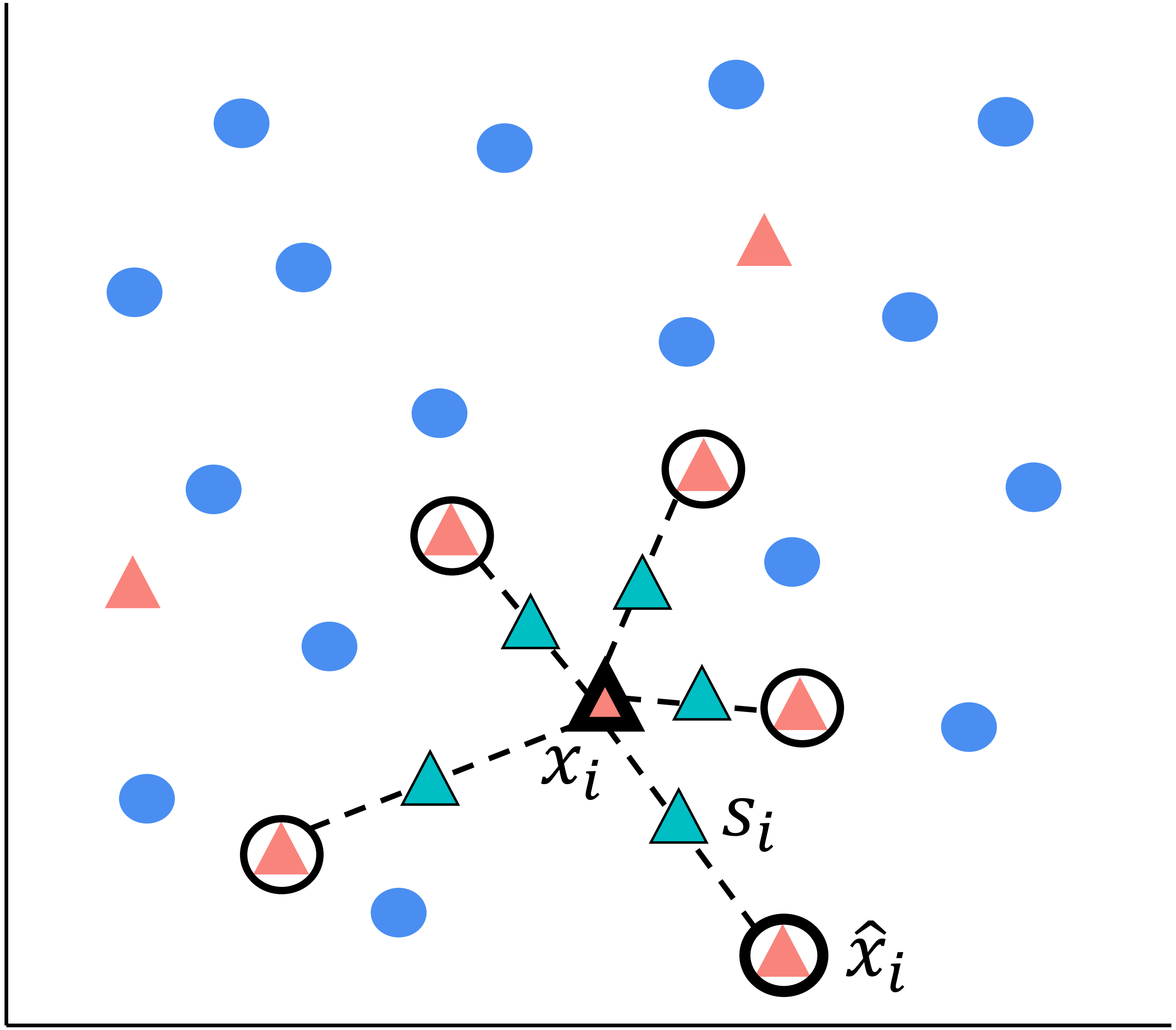
Randomly selected minority class datapoint

Randomly selected k-nearest neighbours

Synthetic minority instance

(Chawla et. al, 2002)

SMOTER (synthetic minority oversampling technique regression)



Continuous

$$s_i = x_i + RANDOM(0,1)(x_i - \hat{x}_i)$$

Categorical

Randomly select between x_i and $\hat{x}_i$

Target

$$s_{i_y} = \frac{d_2 * \hat{y}_i + d_1 * y_i}{d_1 + d_2}$$

$$d_1 = \mathrm{DISTANCE}\left(s_{i_y}, \hat{x}_i\right)$$

$$d_2 = \mathrm{DISTANCE}\left(s_{i_y}, x_i\right)$$

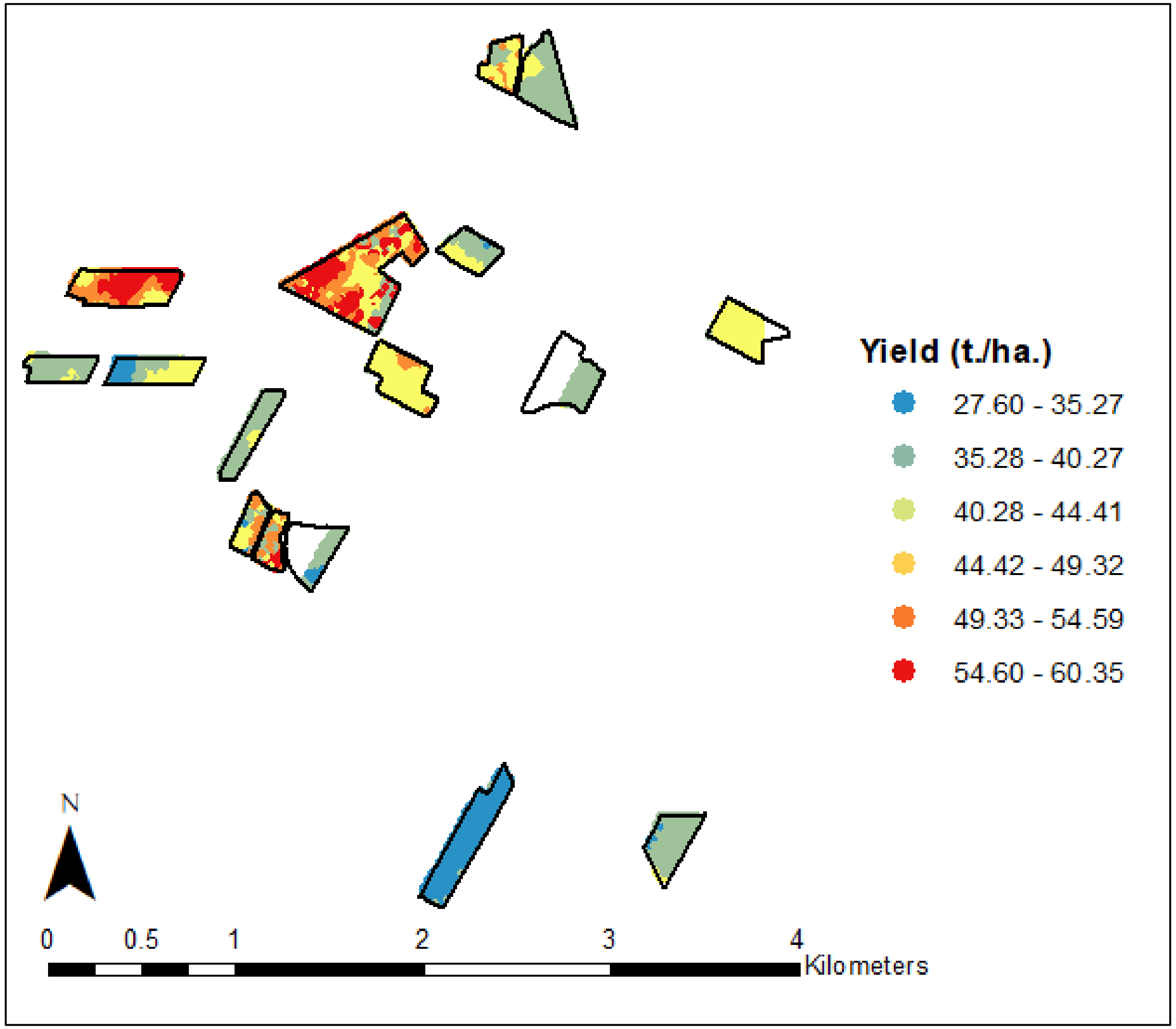
(Chawla et al., 2002, Torgo et al., 2013, 2015)

Objectives

- Evaluate the effects of applying the three preprocessing techniques
- Assess the quality of the synthetic data generated by SMOTER

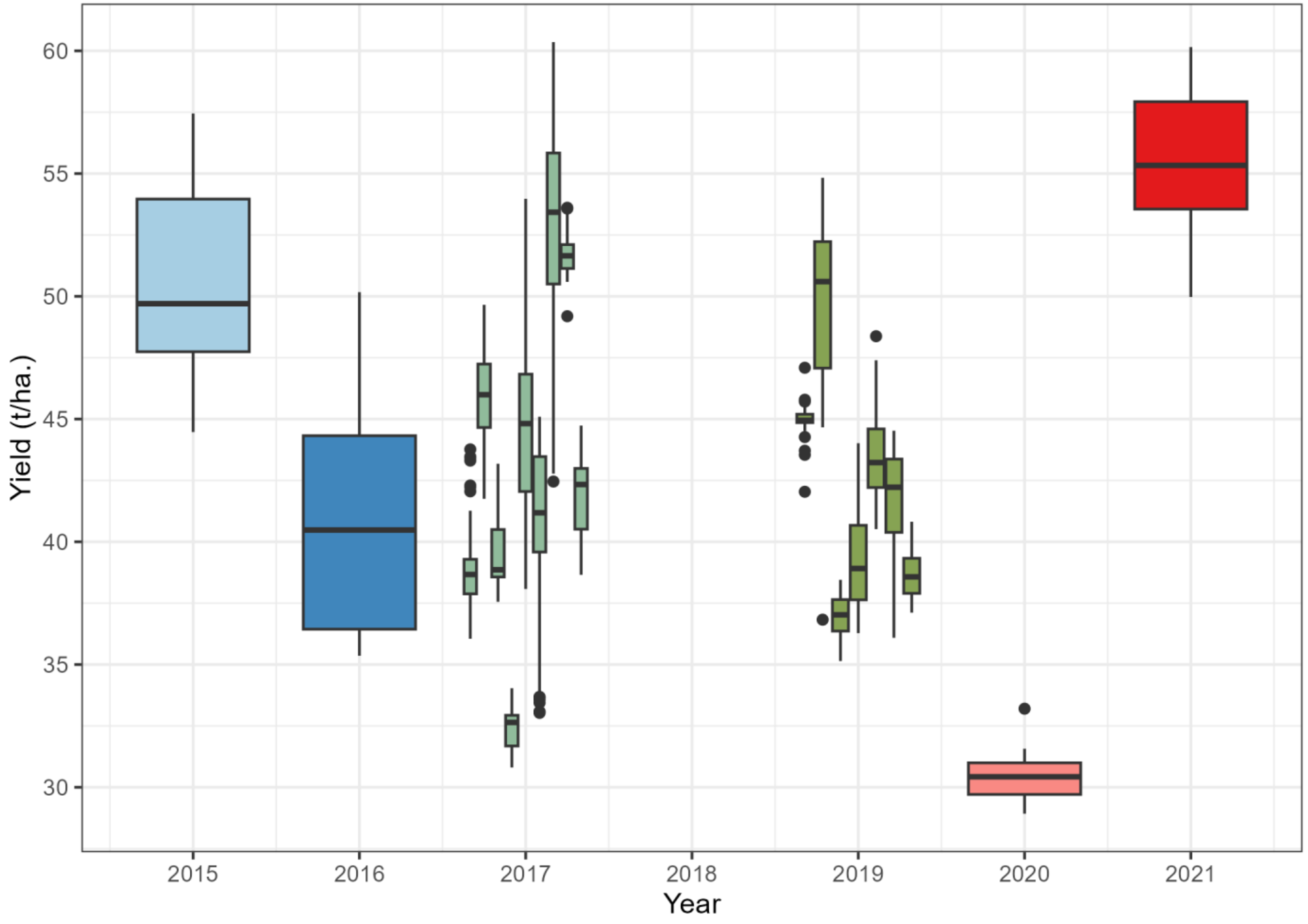
Data

Imagery ©2024 NASA, TerraMetrics, Map data ©2024 Google 100 km

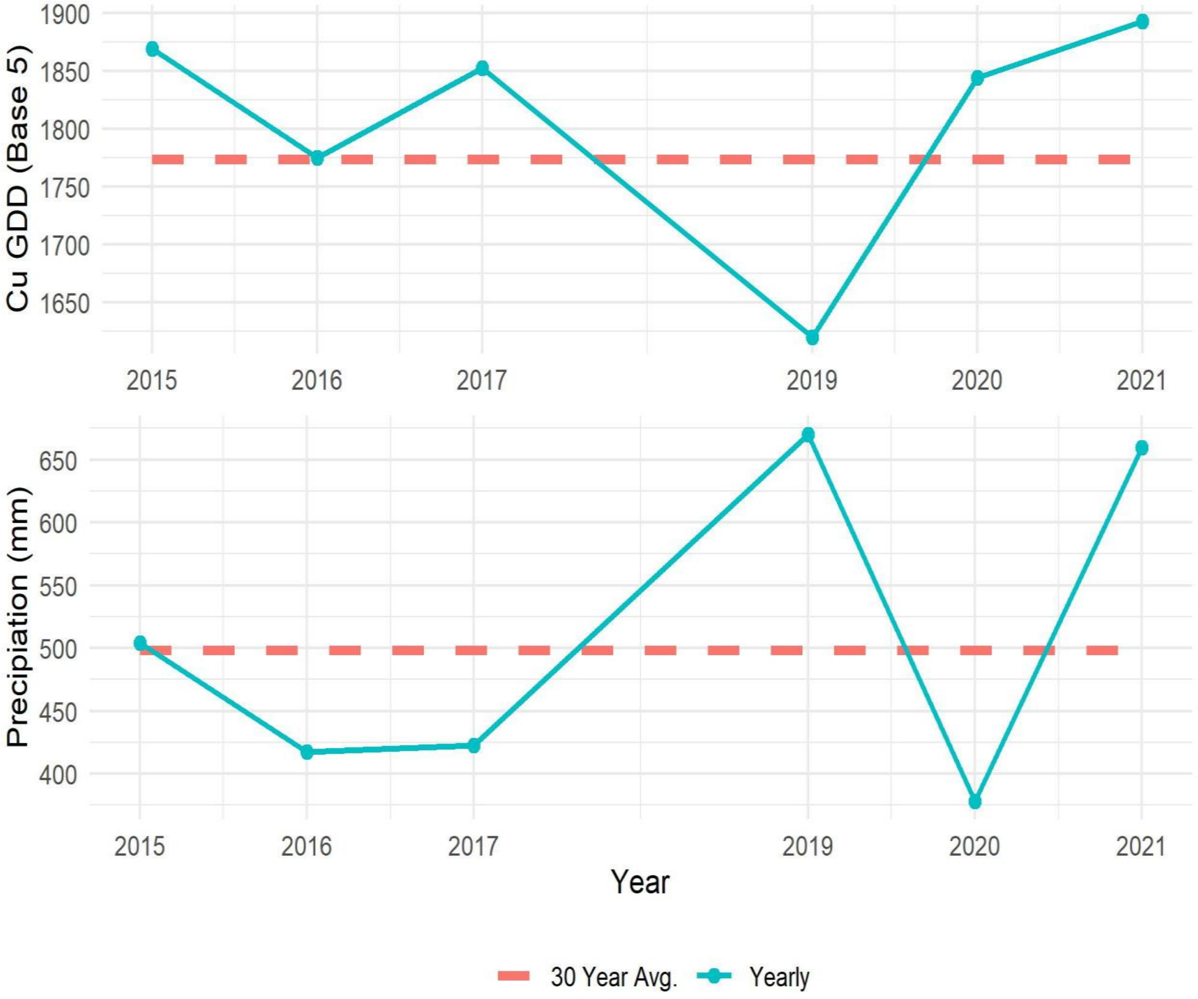


- Russet Burbank Potato
- 20 Field-Years

Data

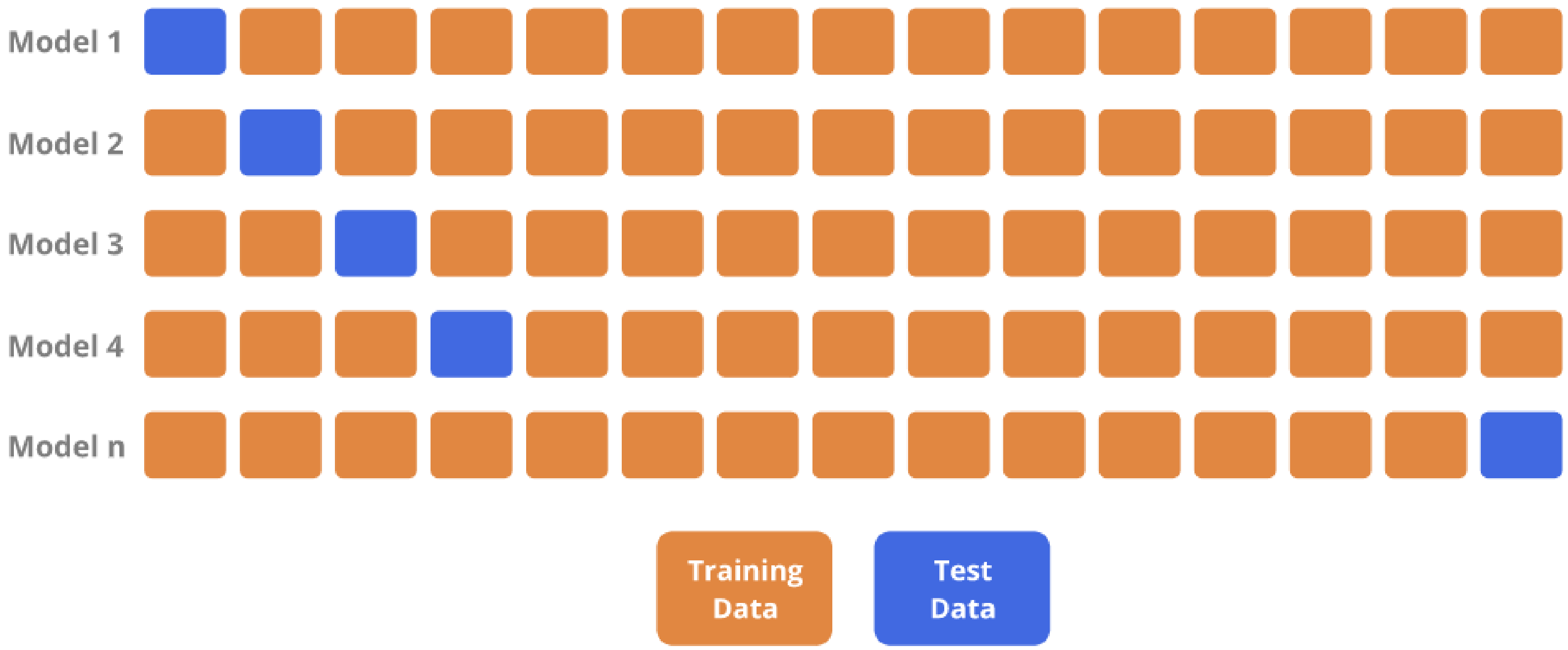


Yield (t/ha.)
60
55
50
45
40
35
30
2015
2016
2017
2018
2019
2020
2021
Year



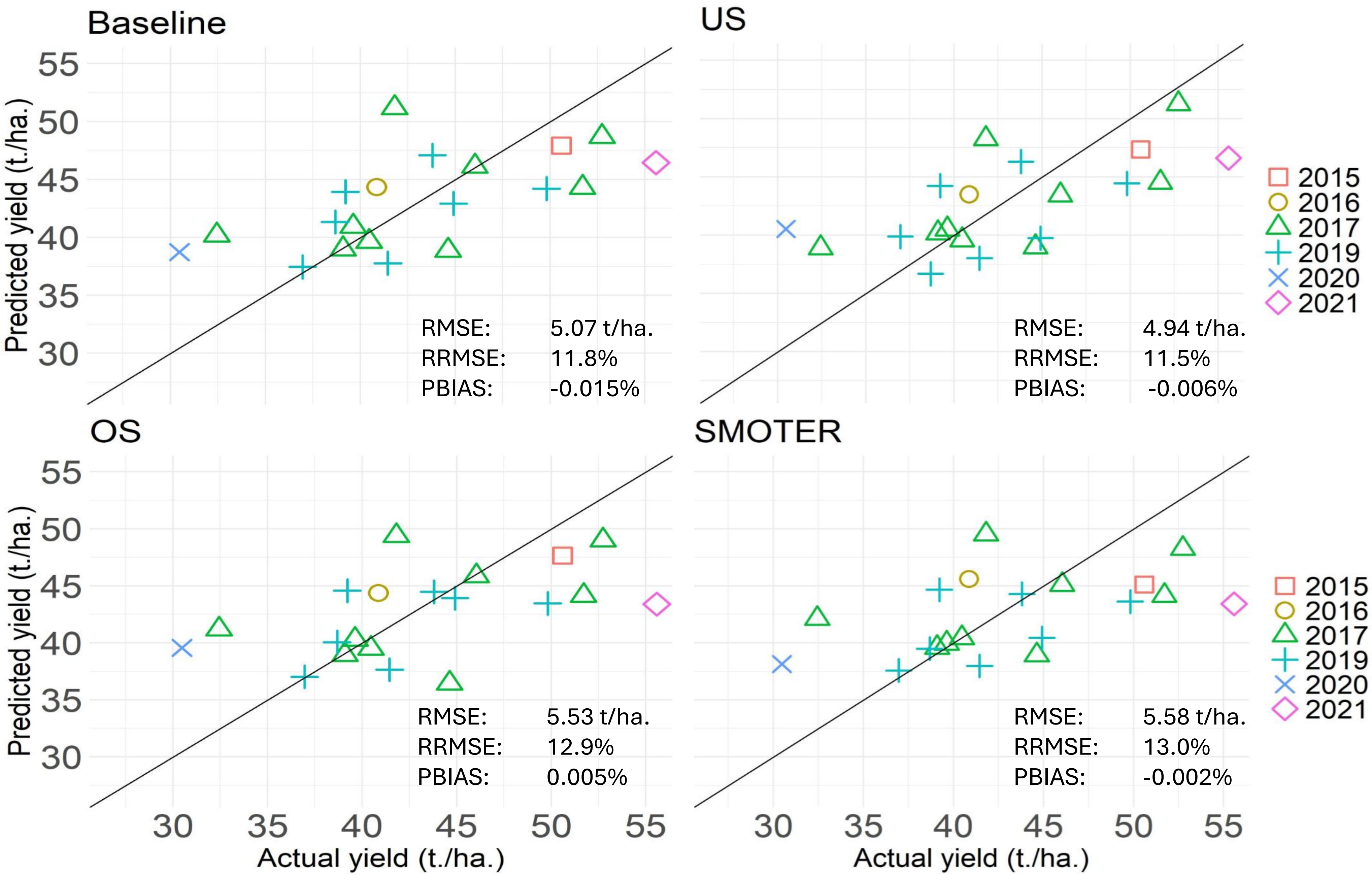
Cu GDD (Base 5)
1900
1850
1800
1750
1700
1650
2015
2016
2017
2019
2020
2021
Precipitation (mm)
650
600
550
500
450
400
2015
2016
2017
2019
2020
2021
Year
30 Year Avg.
Yearly

Random Forest Models



- 2-fold cross validation
- Grid search
  - *Mtry* – number of features sampled at each node split
  - *Ntree* – number of trees built
  - *Nodesize* – number of terminal nodes (tree size)

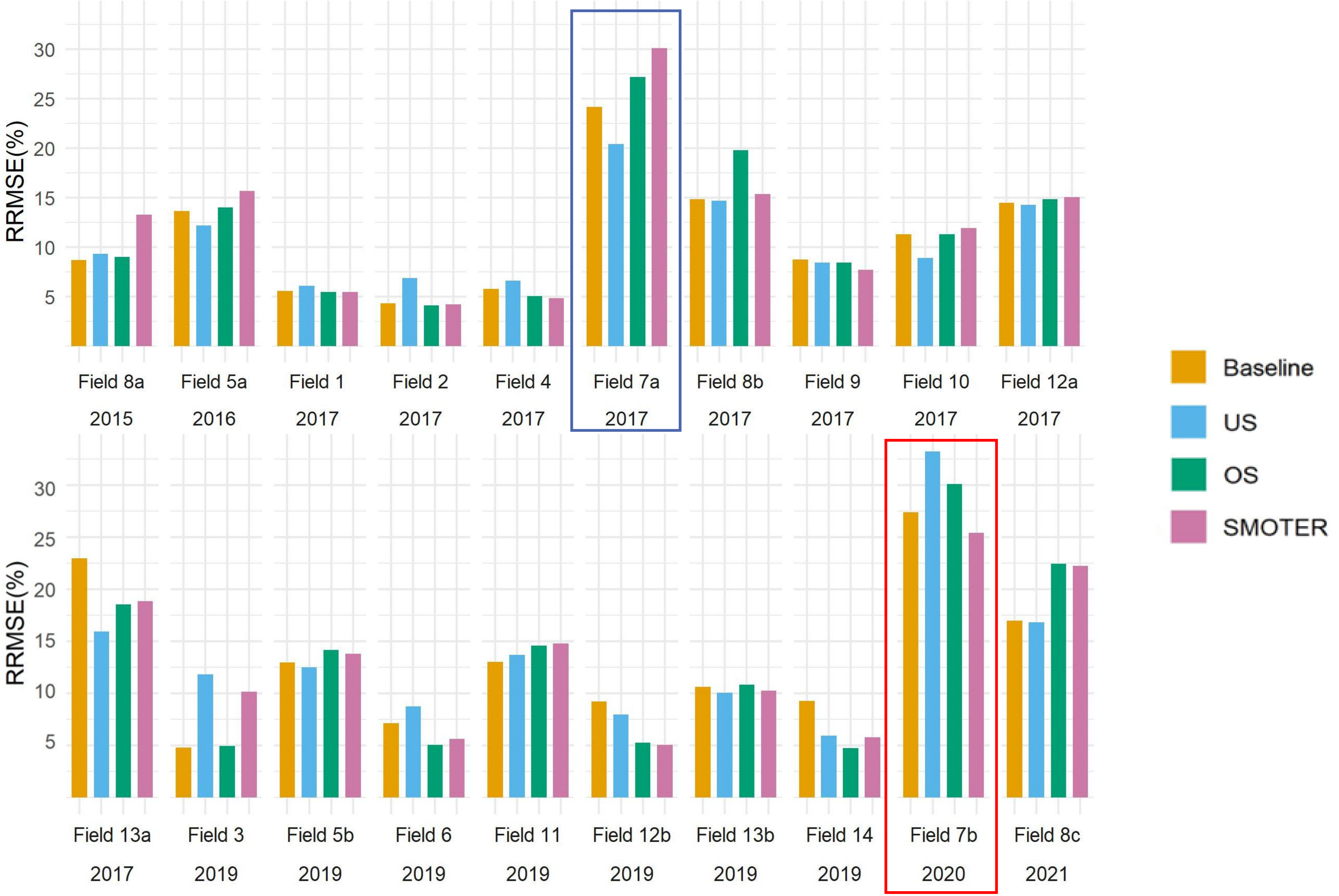
(Gareth et al., 2013)


Baseline
US
OS
SMOTER
Predicted yield (t./ha.)
Actual yield (t./ha.)
55
50
45
40
35
30
RMSE: 5.07 t/ha.
RRMSE: 11.8%
PBIAS: -0.015%
RMSE: 4.94 t/ha.
RRMSE: 11.5%
PBIAS: -0.006%
RMSE: 5.53 t/ha.
RRMSE: 12.9%
PBIAS: 0.005%
RMSE: 5.58 t/ha.
RRMSE: 13.0%
PBIAS: -0.002%
2015
2016
2017
2019
2020
2021

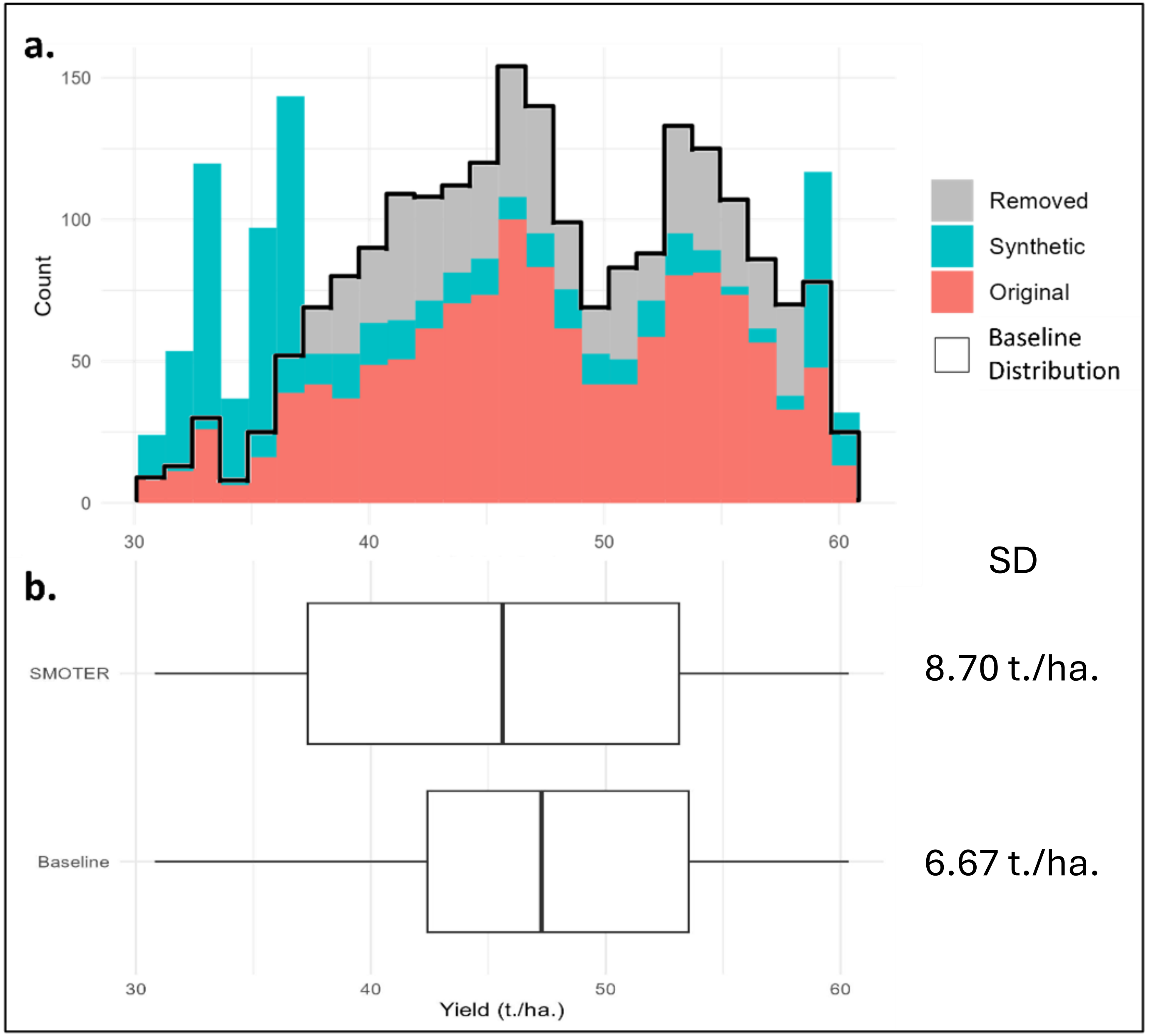
Wins/Losses

		Best	Win	Loss
Average Whole-Field Predicted Yield	baseline	4	-	-
	OS	2	7	13
	SMOTER	5	9	11
	US	9	12	8
RRMSE	baseline	3	-	-
	OS	3	9	11
	SMOTER	5	10	10
	US	9	12	8
PBIAS	baseline	2	-	-
	OS	4	9	11
	SMOTER	5	8	12
	US	9	12	8

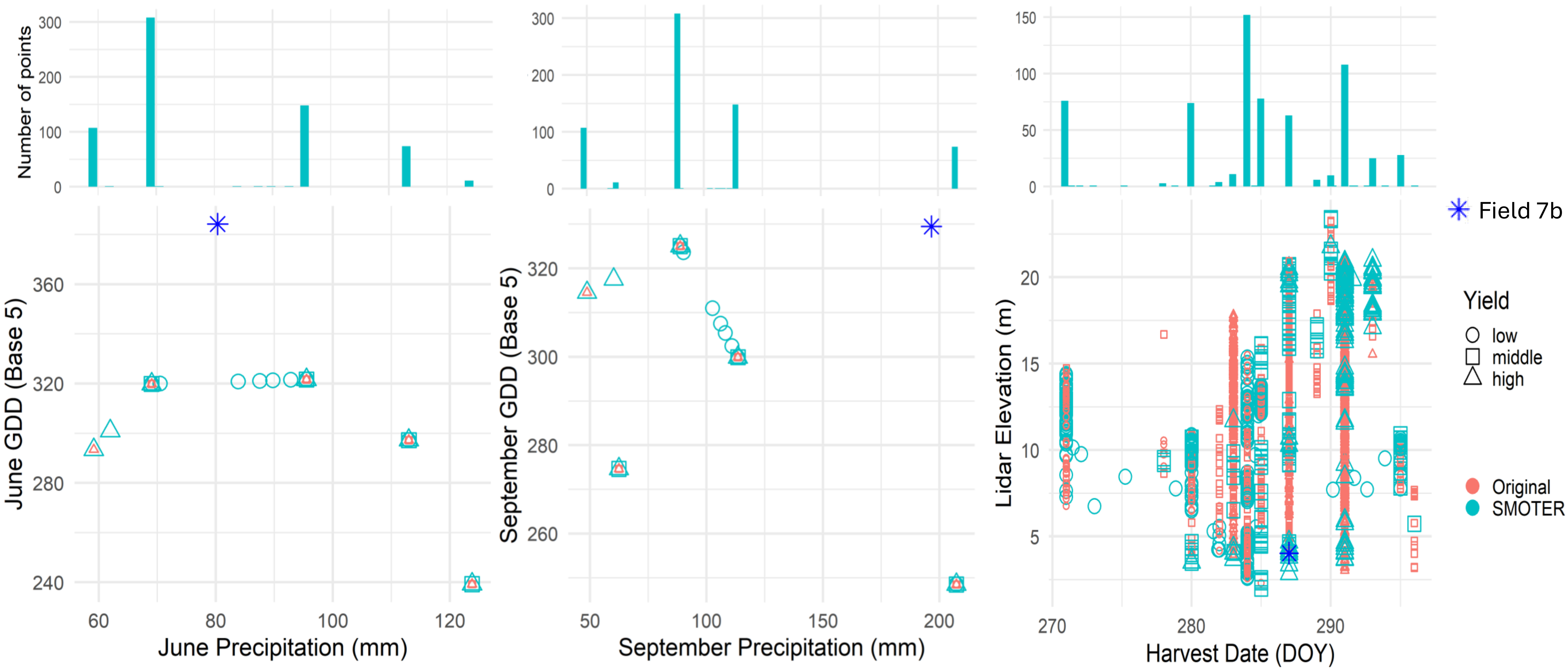
RRMSE
per field-year



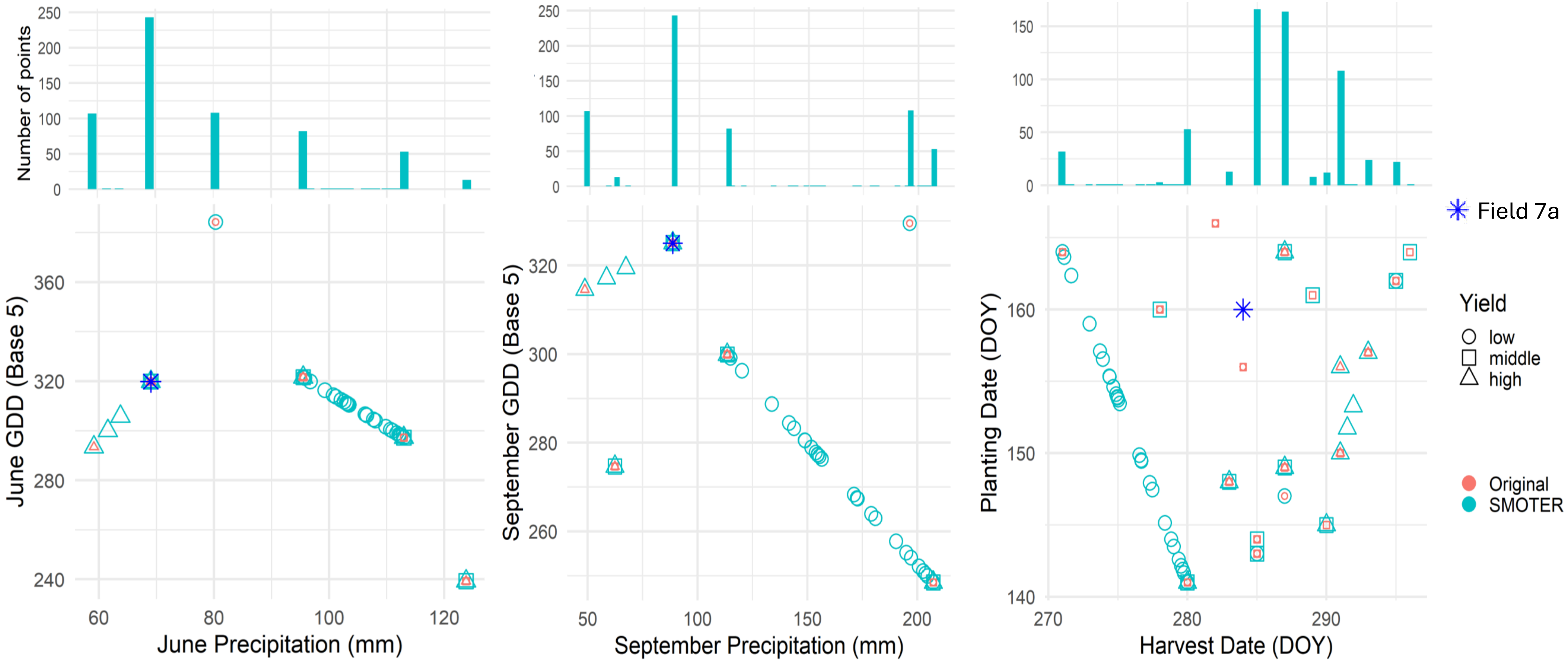
Field 7b – New Training Set Distribution with SMOTER



Field 7b (Drought Year) – SMOTER Feature Space



Field 7a (2017) – SMOTER Feature Space



Conclusion

- US performed best overall
- SMOTER had best performance for drought year
- SMOTER increases the variability within feature space
- Potential to improve predictions for cases of limited data due to drought or extreme years

- Fine-tuning of SMOTER parameters
- Future experiments with more drought or extreme years
- Expand to more models

References

Branco, P., Ribeiro, R. P., & Torgo, L. (2016). UBL: An R package for Utility-based Learning (arXiv:1604.08079). arXiv. https://doi.org/10.48550/arXiv.1604.08079

Breiman, L. (2001). Random Forests. Machine Learning, 45, 5–32. https://doi.org/10.1023/A:1010933404324

Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: Synthetic Minority Over-sampling Technique. Journal of Artificial Intelligence Research, 16, 321–357. https://doi.org/10.1613/jair.953

Chemchem, A., Alin, F., & Krajecki, M. (2019). Combining SMOTE Sampling and Machine Learning for Forecasting Wheat Yields in France. 2019 IEEE Second International Conference on Artificial Intelligence and Knowledge Engineering (AIKE), 9–14. https://doi.org/10.1109/AIKE.2019.00010

Ebrahimy, H., Wang, Y., & Zhang, Z. (2023). Utilization of synthetic minority oversampling technique for improving potato yield prediction using remote sensing data and machine learning algorithms with small sample size of yield data. ISPRS Journal of Photogrammetry and Remote Sensing, 201, 12–25. https://doi.org/10.1016/j.isprsjprs.2023.05.015

Jamieson, P. D., Porter, J. R., & Wilson, D. R. (1991). A test of the computer simulation model ARCWHEAT 1 on wheat crops grown in New Zealand. Field Crops Research, 27, 337–350. https://doi.org/10.1016/0378-4290(91)90040-3

Kuhn, M. (2008). Building Predictive Models in R Using the caret Package. Journal of Statistical Software, 28(5). https://doi.org/10.18637/jss.v028.i05

Le, K. N., Jeong, J., Reyes, M. R., Jha, M. K., Gassman, P. W., Doro, L., Hok, L., & Boulakia, S. (2018). Evaluation of the performance of the EPIC model for yield and biomass simulation under conservation systems in Cambodia. Agricultural Systems, 166, 90–100. https://doi.org/10.1016/j.agsy.2018.08.003

Liaw, A., & Wiener, M. (2002). Classification and Regression by randomForest. 2.

Mohammed, R., Rawashdeh, J., & Abdullah, M. (2020). Machine Learning with Oversampling and Undersampling Techniques: Overview Study and Experimental Results. 2020 11th International Conference on Information and Communication Systems (ICICS), 243–248. https://doi.org/10.1109/ICICS49469.2020.239556

Torgo, L., Branco, P., Ribeiro, R. P., & Pfahringer, B. (2015). Resampling strategies for regression. Expert Systems, 32(3), 465–476. https://doi.org/10.1111/exsy.12081

Torgo, L., Ribeiro, R. P., Pfahringer, B., & Branco, P. (2013). SMOTE for Regression. In L. Correia, L. P. Reis, & J. Cascalho (Eds.), Progress in Artificial Intelligence (pp. 378–389). Springer. https://doi.org/10.1007/978-3-642-40669-0_33

Wang, J., Wang, J., Shi, Y., Zhou, H., & Liao, L. (2019). A recursive update model for estimating high-resolution LAI based on the NARX neural network and MODIS times series. Remote Sensing (Basel, Switzerland), 11(3), 219. https://doi.org/10.3390/rs11030219

Thank you

- US performed best overall
- SMOTER had best performance for Field 7b drought year
- SMOTER increases the variability within feature space
- Potential to improve predictions for cases of limited data due to drought or extreme years

- Fine-tuning of SMOTER preprocessing
- Future experiments with more drought or extreme years
- Expand to more models

mariaelisa.polsinelli@mail.mcgill.ca
mariaelisa.polsinelli@agr.gc.ca