MBA
USP
ESALQ

Mapping sugarcane yield in São Paulo State: applying panel data analysis and machine learning

Nélida E. Q. Silvero
João Vitor Matos Gonçalves

Agenda

- Introduction
- Material and methods
- Results
- Final considerations

INTRODUCTION

Introduction

- Brazil is one of the most important producers
- Sugarcane production occupies the third position, representing approximately 8 M ha and a total production of 654.8 Mt
- São Paulo State → Highest producer with 51% of the area and 54% of the total production
- Average yield 76 Mg ha^{-1}

Why is important to forecast sugarcane yield?

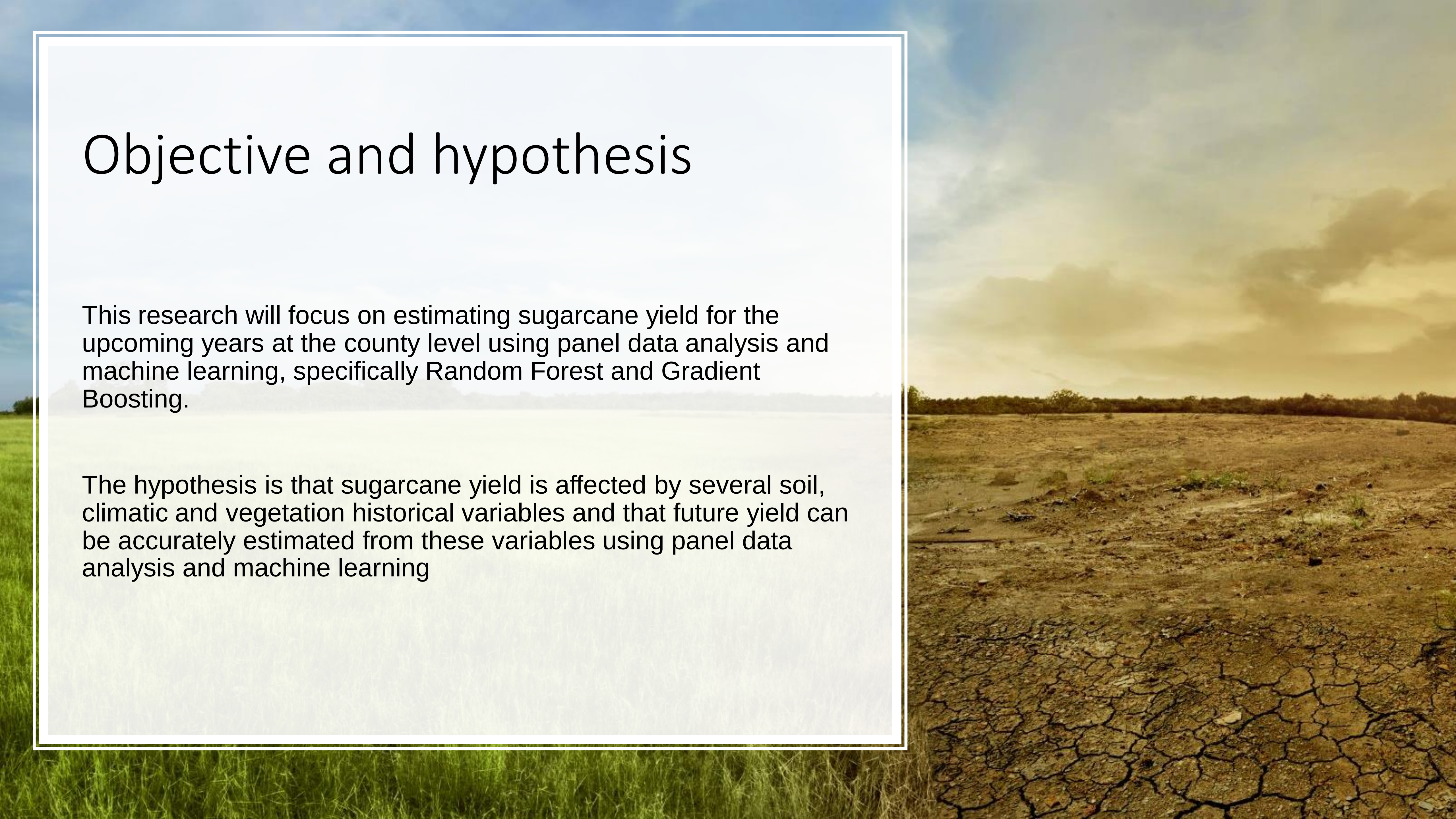
- Essential tool to support decision making processes regarding harvesting, marketing, milling, and selling strategies

- To evaluate strategies to select new cultivars that may better adapt in an area, climate risks and need for irrigation

How sugarcane yields are predicted/forecasted?

- Agrometeorological models
- Process-based models
- Machine learning

MBA USP ESALQ



Objective and hypothesis

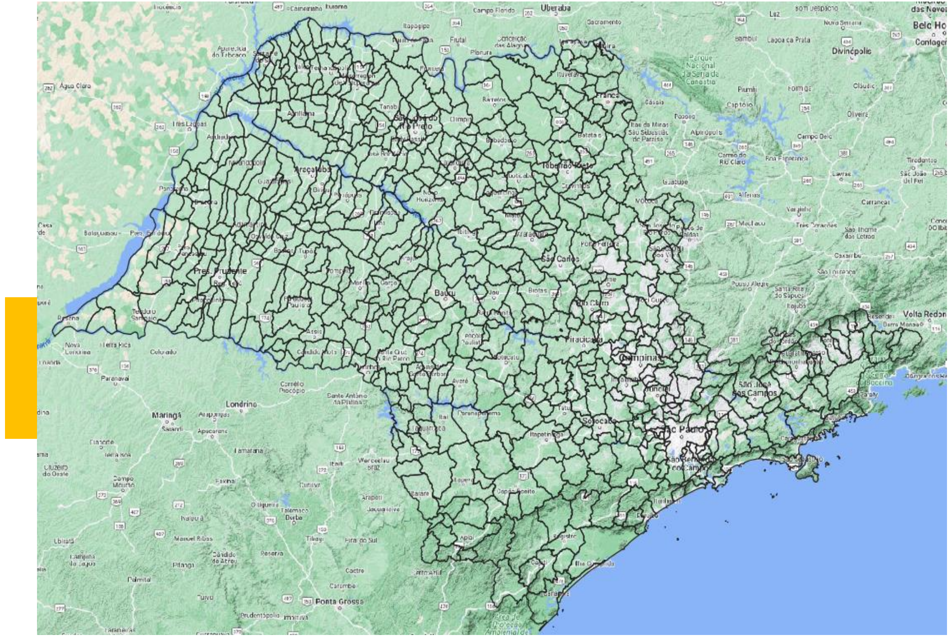
This research will focus on estimating sugarcane yield for the upcoming years at the county level using panel data analysis and machine learning, specifically Random Forest and Gradient Boosting.

The hypothesis is that sugarcane yield is affected by several soil, climatic and vegetation historical variables and that future yield can be accurately estimated from these variables using panel data analysis and machine learning

Material and methods

Study area and datasets

Municipalities of the State of São Paulo



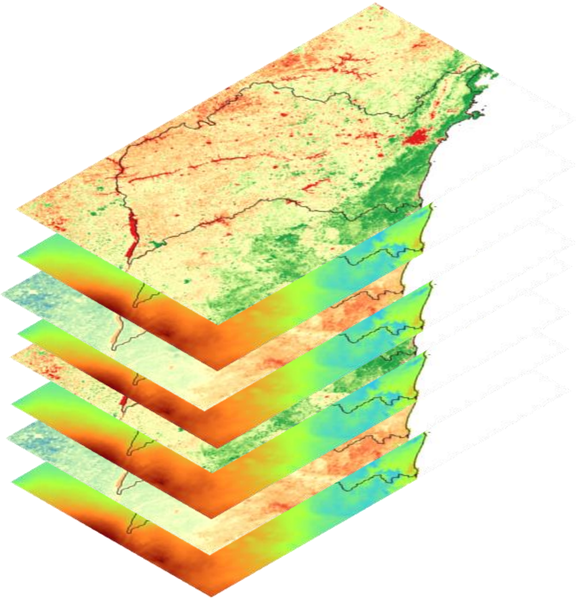
Variable Y: Sugarcane yield (kg/ha)

From Table 1612 of the SIDRA Platform
Years 2000-2020

Variables X (Years 2000-2020):

- Harvested area (AC, ha)
- Total production (QP, tn)
- Production values (VP, R$)

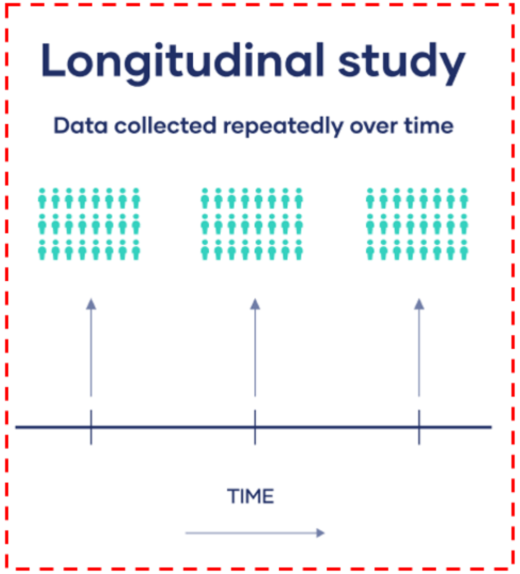
- FROM TERRACLIMATE
  - Precipitation (pr)
  - Soil moisture (soil)
  - Evapotranspiration (aet)
  - Minimum temperature(tmin)
  - Maximum temperature (tmmx)
- FROM MODIS
  - NDVI, NPP, LST



Data for the year 2021 were made available in Sep 2022 and were used as a validation set

Statistical analysis

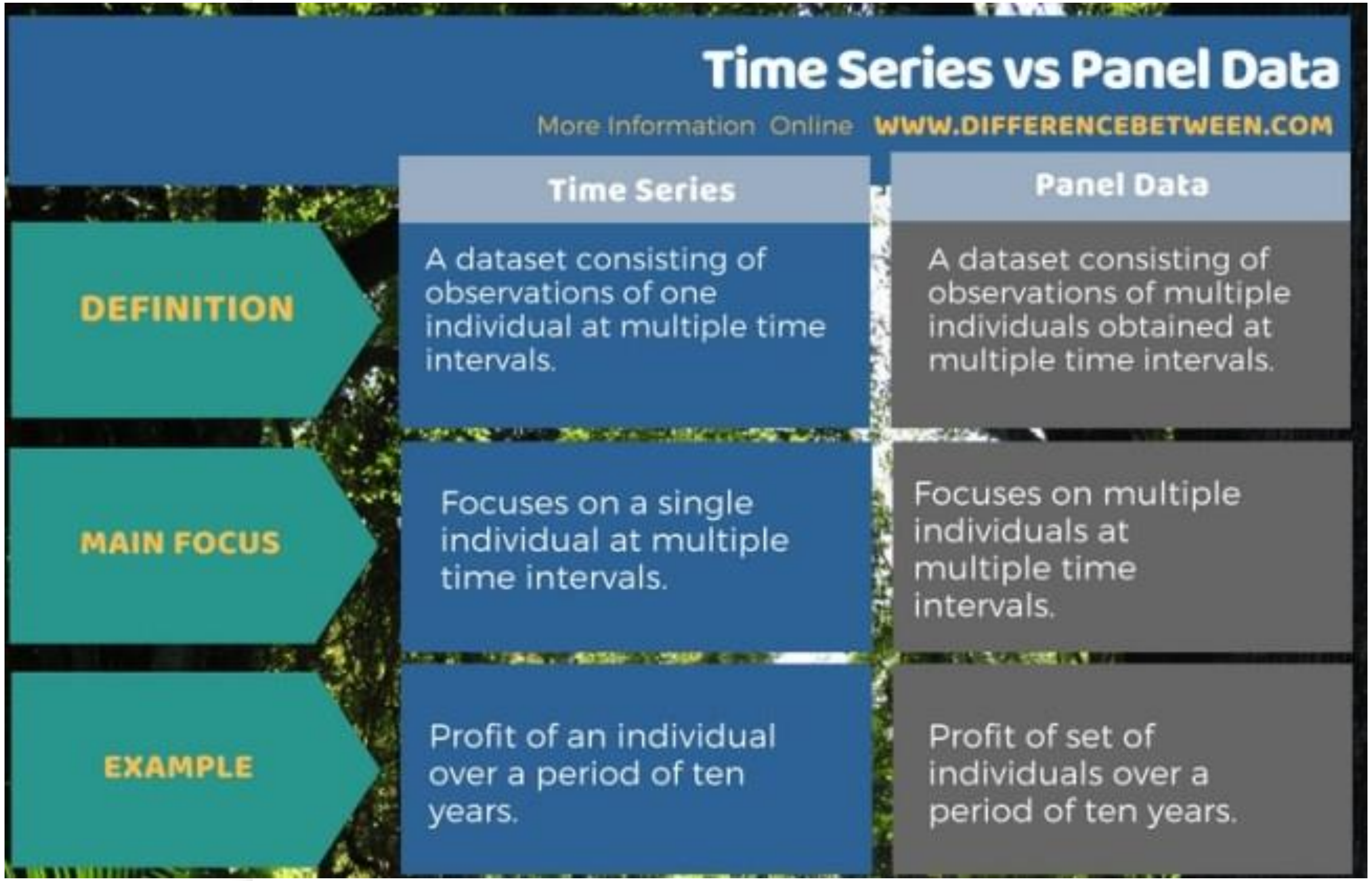
Mean and standard deviation for all years and municipalities and over time.



Panel Data Analysis

Gradient Boosting

Why Panel Data Analysis?

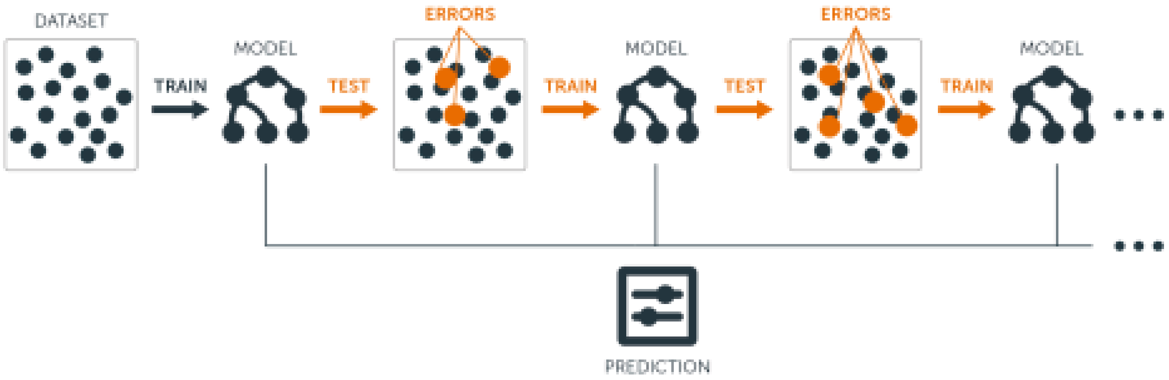


Time Series vs Panel Data

More Information Online WWW.DIFFERENCEBETWEEN.COM

	Time Series	Panel Data
DEFINITION	A dataset consisting of observations of one individual at multiple time intervals.	A dataset consisting of observations of multiple individuals obtained at multiple time intervals.
MAIN FOCUS	Focuses on a single individual at multiple time intervals.	Focuses on multiple individuals at multiple time intervals.
EXAMPLE	Profit of an individual over a period of ten years.	Profit of set of individuals over a period of ten years.

Gradient Boosting



Weak learners used sequentially to reduce errors

Usually used with decision trees

Modeling strategies

Panel Data Analysis and Gradient Boosting

Training set: Sugarcane yield and variables from years 2000 to 2014

Test set: Sugarcane yield and variables from years 2015-2020

Validation set: Sugarcane yield from year 2021

Metrics: R^2 and RMSE

Only Gradient Boosting

Hyperparameterization: Yes

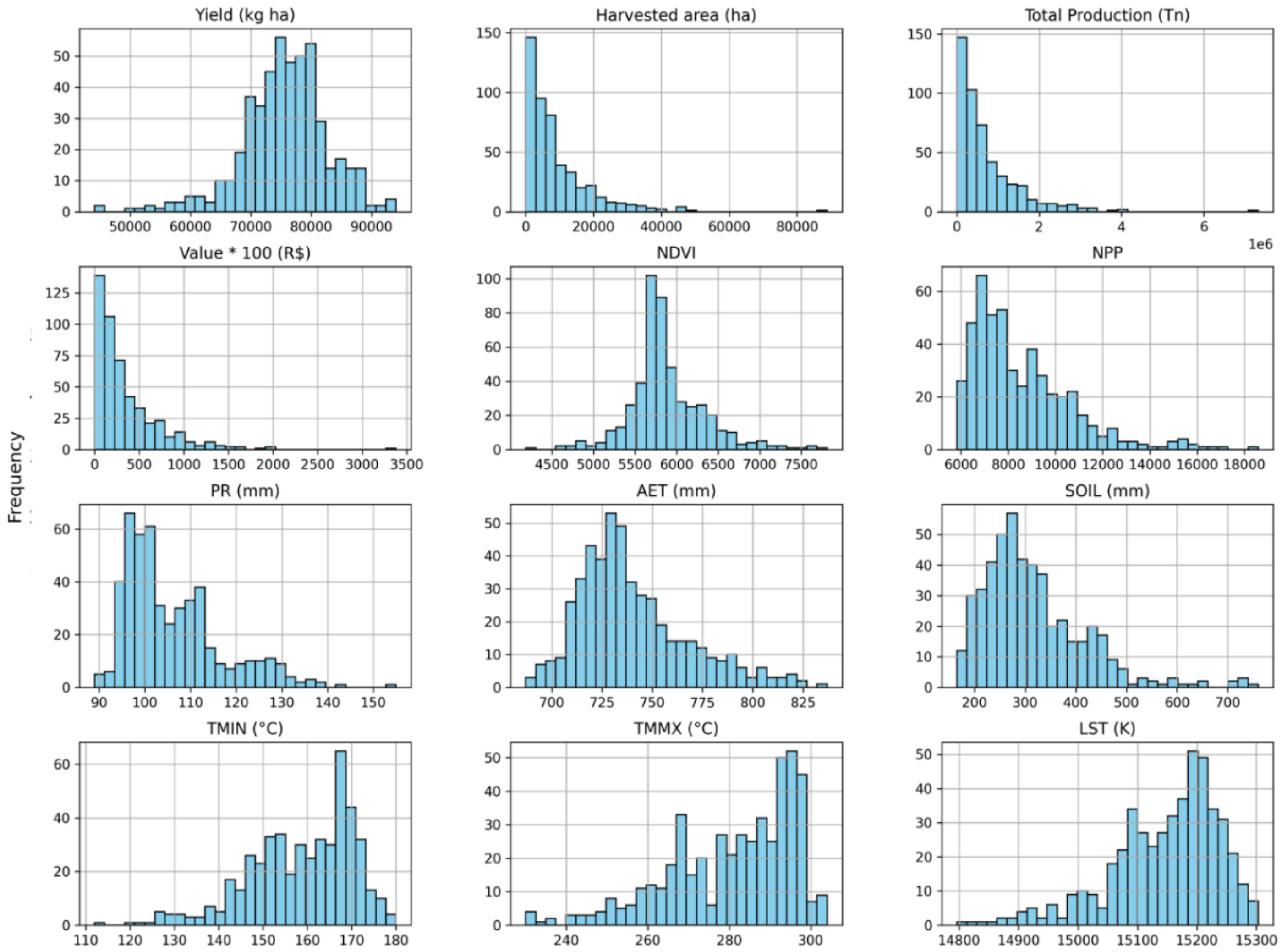
Only Panel Data Analysis

Groupings by quantiles

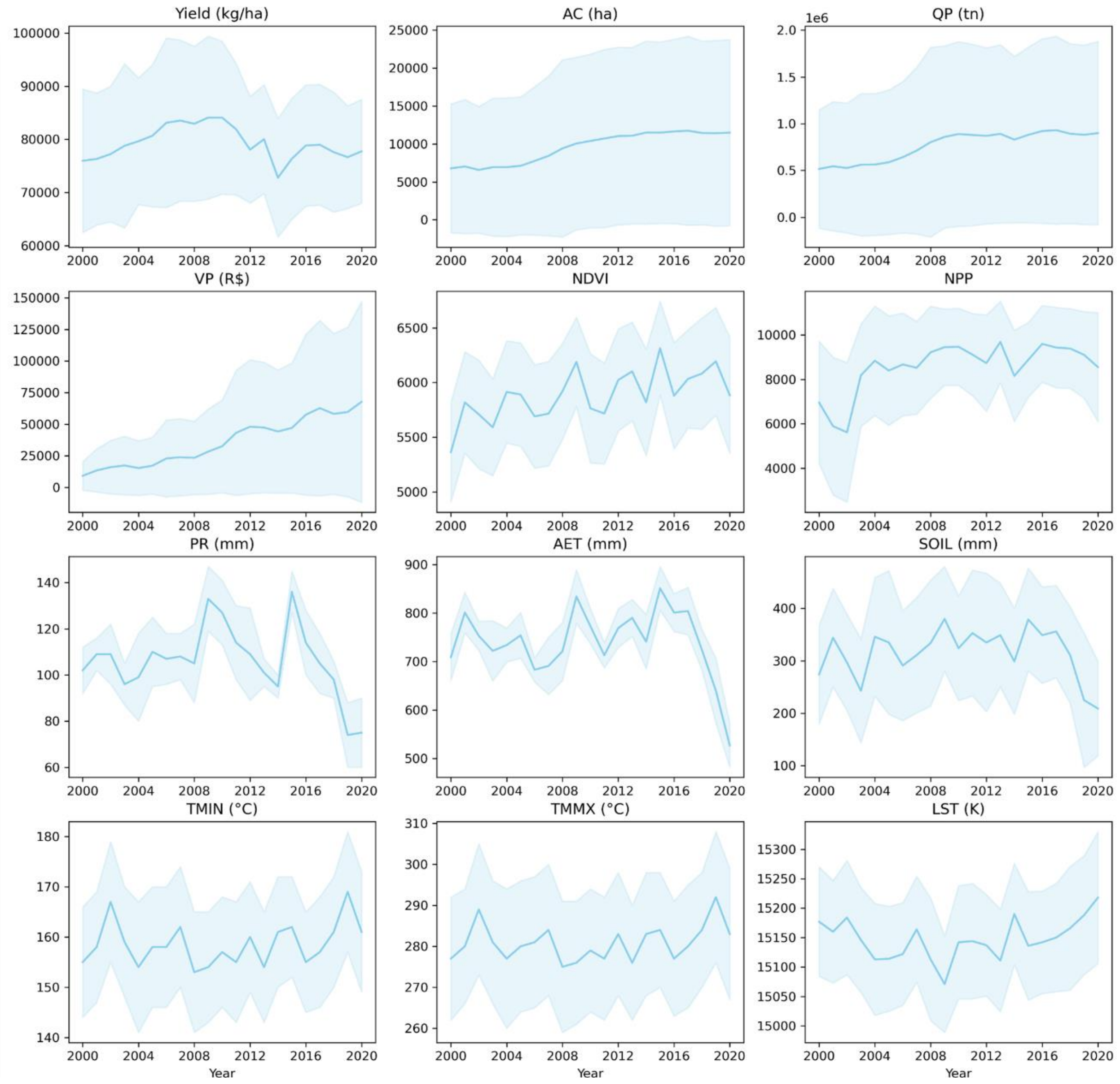
RESULTS

Descriptive statistics

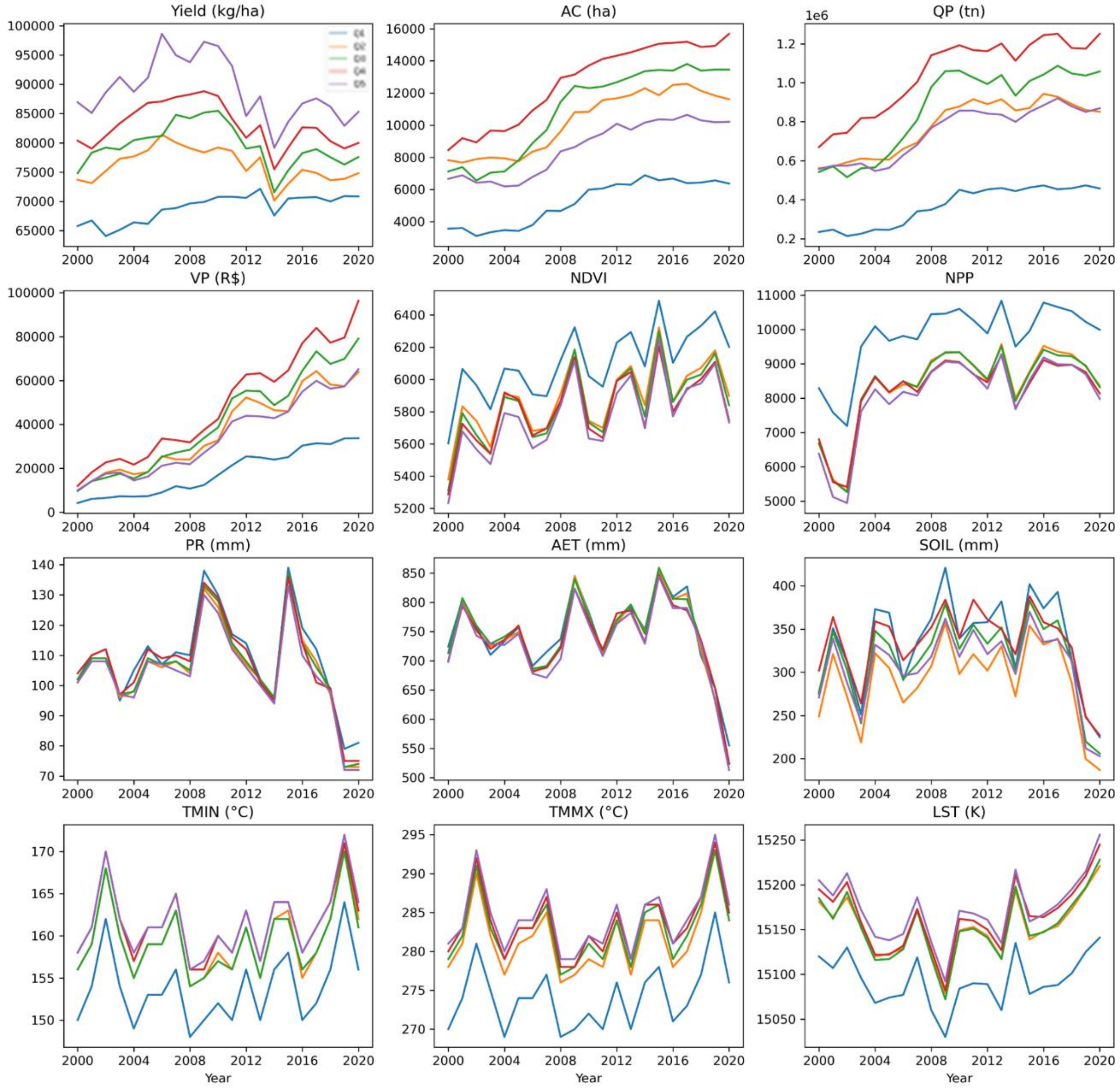
Average of all variables



All variables by year



Grouping by quantiles



Panel Data Analysis

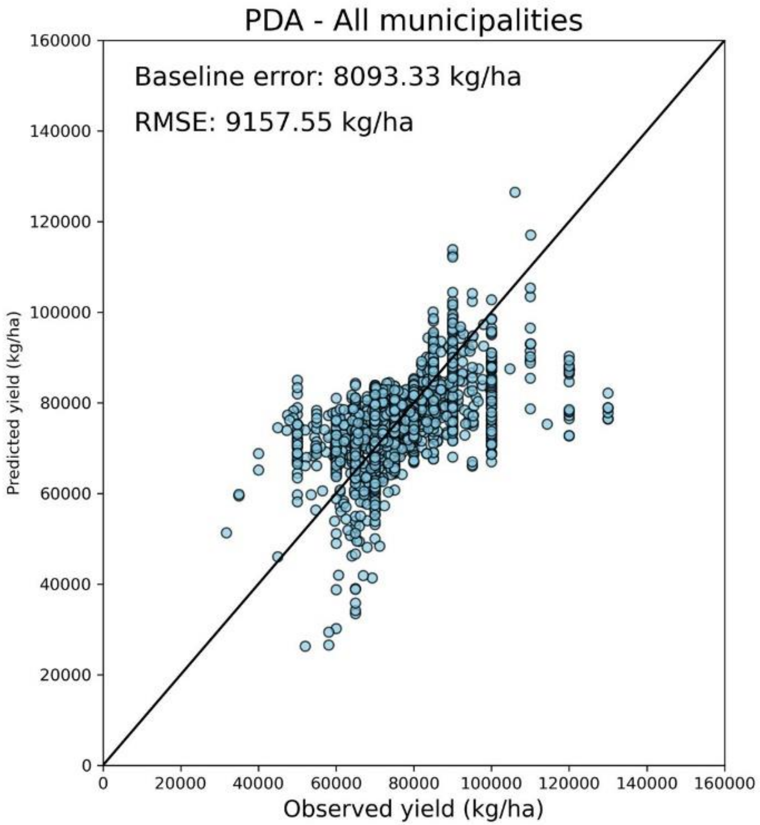
R$^2_{traininig}$: 0.4

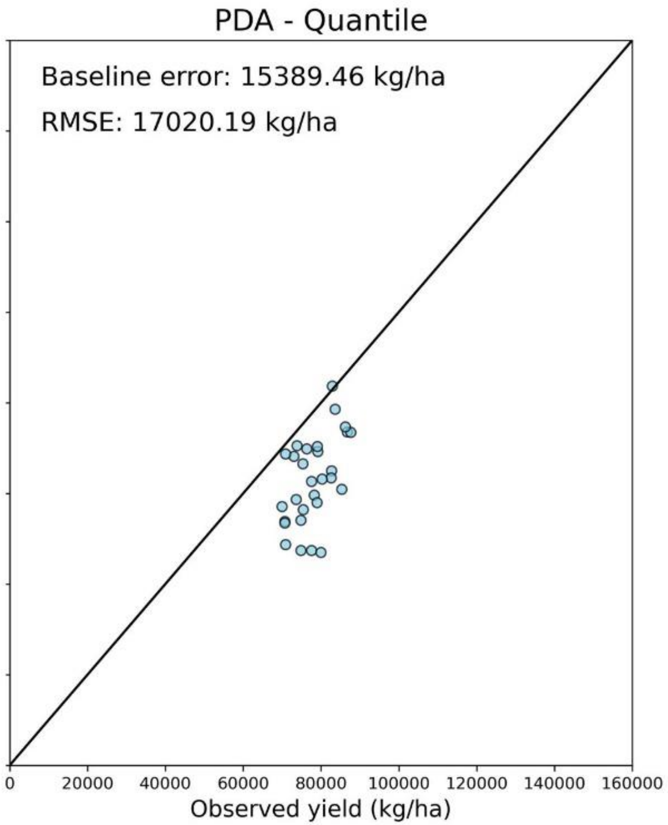
R$^2_{test}$: 0.3

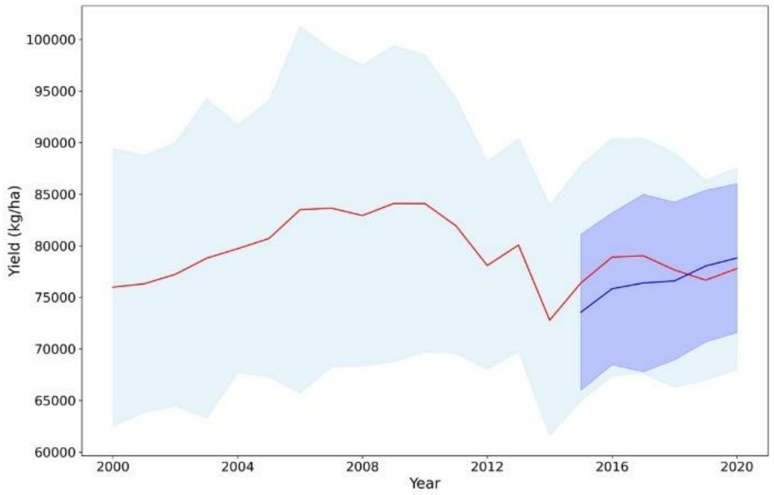
Significance: Yes (5%)

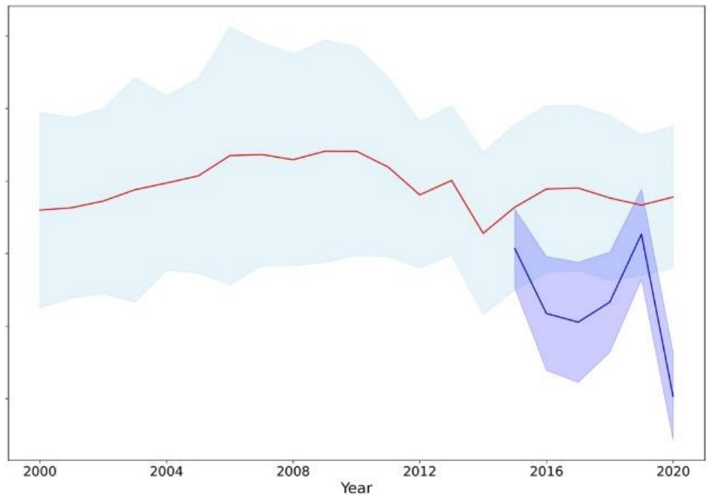
All variables were significant except soil moisture

PDA with quartiles had the worst performance









Gradient Boosting

Default

$R^2_{traininig}$: 0.70

R^2_{test}: 0.60

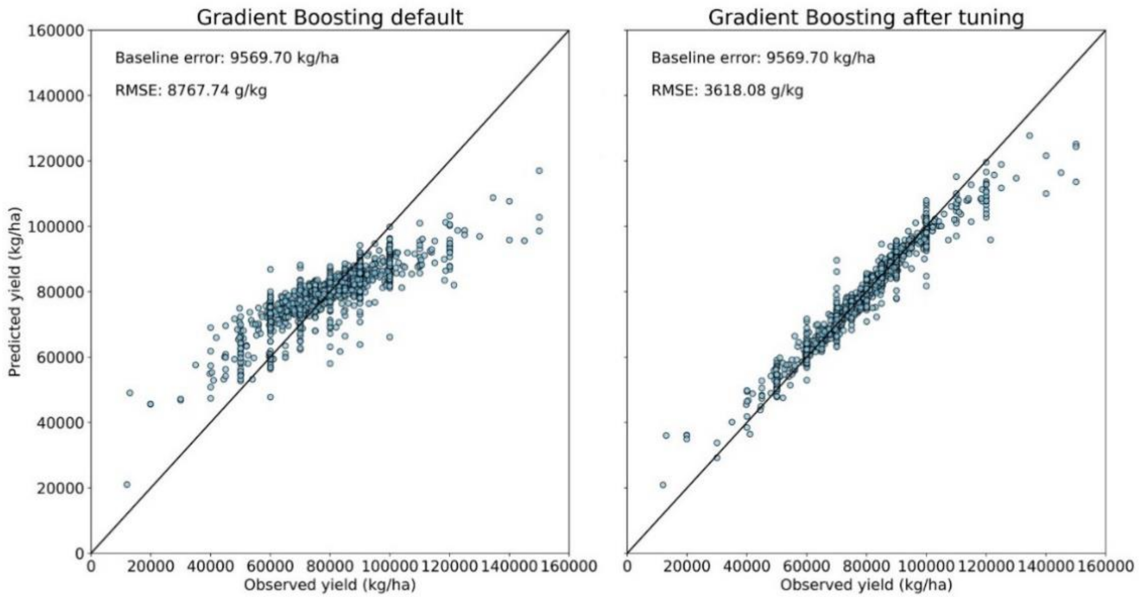
With hyperparameterization

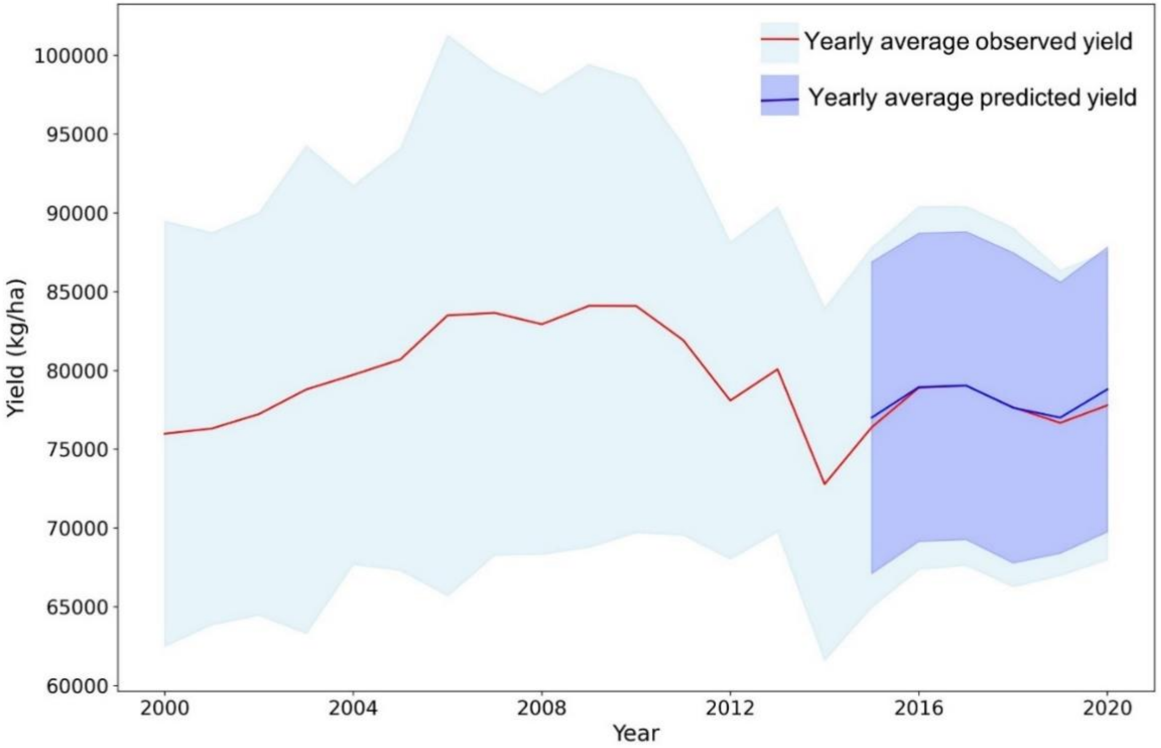
$R^2_{traininig}$: 0.92

R^2_{test}: 0.87

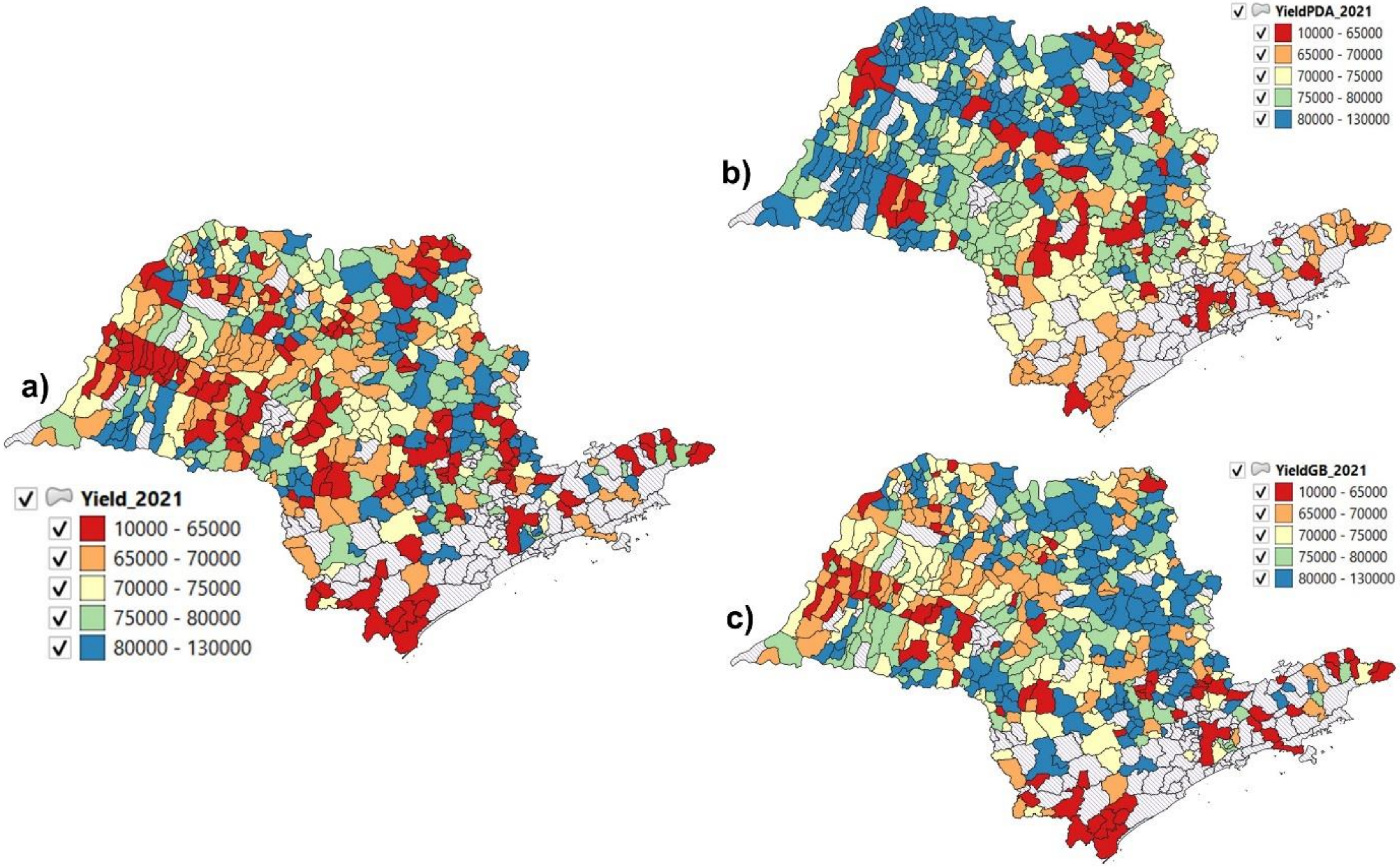
Most important variable: QP

GB with hyperparameterization was the best





Spatial distribution of sugarcane yield



FINAL CONSIDERATIONS

FINAL CONSIDERATIONS

- The importance of crop yield prediction for different stakeholders
- Gradient Boosting better than Panel Data Analysis after a set of hyperparameterization runs
- More research is needed to explore other environmental covariates and specifically management practices that were not available at the municipality level

https://neli12.github.io/

nelida.silvero@rothamsted.ac.uk

https://github.com/neli12

linkedin.com/in/nsilvero

Obrigada!