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

Nélida Elizabet Quiñonez Silvero[1]*; João Vitor Matos Gonçalves[2]

[1] Rothamsted Research, Visiting Research Scientist, Sustainable Soil and Crops Department. Harpenden, UK
[2] Mestre em Economia. Rua da Chibata 128 – Vila Andrade; 05734-100 São Paulo, SP, Brasil
*autor correspondente: neli.silvero@usp.br

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

Abstract

Crop yield forecasting is especially important to farmers and decision makers. An idea of crop performance is possible to obtain even before the harvest take place. At regional or national levels, crop yield forecasting provides useful insights on the state of the art of commodities that are of economic importance and needs constant monitoring of their performance over time. In this study, sugarcane yield of the State of São Paulo at the municipality level was estimated using two approaches: panel data analysis and machine learning algorithms. The historical sugarcane yield from 21 years (2000-2020) was used to train and test the models and validate with the yield from the year 2021. The panel data analysis (PDA) was used because of the temporal character of the dataset and the presence of more than one individual (municipalities), characterizing longitudinal datasets that cannot be studied by time series alone. To reduce the number of individuals, the municipalities were grouped in five quantiles and the PDA carried out to try to understand if the model was improved. For the machine learning approach, Gradient Boosting (GB) and Random Forest (RF) algorithms were tested, being the first the one that presented the best results after hyperparameter tunning. The PDA both with all the municipalities and by quantiles showed the worst result, with RMSE of 9157.55 and 17020.19 kg ha^{-1}, respectively. The RF showed a very good result, but when the hyperparameter searching was applied the performance dropped considerably, indicating the presence of overfitting. The best result with an RMSE of 3619.04 ha^{-1} in the test set and 6368.71 ha^{-1} in the validation set (year 2021) was obtained with the GB algorithm, as mentioned earlier. This model mostly overestimated the sugarcane yields in the central to south parts of the State. Forecasting sugarcane yield at the municipality level in the State of São Paulo will provide useful insights on the historical and future performance of the crop that can help farmer and decision makers make more informed decisions.

Keywords: panel data analysis, sugarcane, remote sensing, time series, crop yield prediction

Mapeamento da produtividade da cana-de-açúcar no Estado de São Paulo: aplicando análise de dados em painel e aprendizado de máquina

Resumo

A previsão do rendimento das colheitas é especialmente importante para os agricultores e tomadores de decisão. É possível obter uma ideia do desempenho da cultura antes mesmo da colheita. A nível regional ou nacional, a previsão do rendimento das culturas fornece informações úteis sobre o estado da arte das commodities que são de importância econômica e precisam de monitoramento constante de seu desempenho ao longo do tempo. Neste estudo, a produtividade da cana-de-açúcar do Estado de São Paulo em nível de município foi estimada usando duas abordagens: análise de dados em painel e algoritmos de aprendizado de máquina. A produtividade histórica da cana de 21 anos (2000-2020) foi usada para treinar e testar os modelos e validar com a produtividade do ano de 2021. A análise de dados em painel (PDA) foi utilizada devido ao caráter temporal do conjunto de dados e à presença de mais de um indivíduo (municípios), caracterizando conjuntos de dados longitudinais que não podem ser estudados apenas por séries temporais. Para reduzir o número de indivíduos, os municípios foram agrupados em cinco quantis e o PDA realizado para tentar entender se o modelo foi aprimorado. Para a abordagem de aprendizado de máquina, foram testados os algoritmos Gradient Boosting (GB) e Random Forest (RF), sendo o primeiro o que apresentou os melhores resultados após o ajuste de hiperparâmetros. A PDA tanto com todos os municípios quanto por quantis apresentou o pior resultado, com RMSE de 9157,55 e 17020,19 kg ha^{-1}, respectivamente. A RF apresentou um resultado muito bom, mas quando a busca de hiperparâmetros foi aplicada o desempenho caiu consideravelmente, indicando a presença de overfitting. O melhor resultado com RMSE de 3619,04 ha^{-1} no conjunto de teste e 6368,71 ha^{-1} no conjunto de validação (ano 2021) foi obtido com o algoritmo GB, conforme mencionado anteriormente. Esse modelo superestimou principalmente a produtividade da cana-de-açúcar nas regiões centro-sul do Estado. A previsão da produtividade da cana-de-açúcar em nível municipal no Estado de São Paulo fornecerá informações úteis sobre o desempenho histórico e futuro da cultura que pode ajudar agricultores e tomadores de decisão a tomar decisões mais informadas.

Palavras-chave: dados em painel, cana-de-açúcar, sensoriamento remoto, séries temporais, produtividade, predição

1. Introduction

Sugarcane is one of the most cultivated crops in the world and Brazil is one of the most important producers (Amorim et al., 2022). Here, sugarcane production occupies the third position, just behind soybean and corn, representing approximately 8616.10 M ha and a total production of 654.8 Mtn (CONAB, 2021). The Central-South is the region where much of the produced sugarcane can be found. The São Paulo State (SP) is one of the most representative with 51% of the area and 54% of the total production, followed by Goiás, Minas Gerais and Mato Grosso do Sul (CONAB, 2021).

The average sugarcane yield in Brazil is approximately 76 Mg ha^{-1}, with maximum reaching in some cases more than 120 Mg ha^{-1}. Using a process-based model (DSSAT/CANEGRO), Monteiro and Sentelhas (2014) estimated an average sugarcane yield of 85 Mg ha^{-1} for SP, slightly higher than that reported by governmental agencies. Knowing the performance of sugarcane yield over a growing season and even before harvest has been described as an essential tool to support decision making processes regarding harvesting, marketing, milling, and selling strategies (Rahman and Robson, 2016). Besides that, estimating future yield is an important tool to evaluate strategies to select new cultivars that may better adapt in an area, climate risks and need for irrigation (Monteiro and Sentelhas, 2014). As there is an increasing interest in producing more sugarcane to meet the demands of ethanol and reduce the use of fossil fuels (Bordonal et al., 2018), having in hands the information of how much will be produced prior to the harvest will provide useful insights and future estimation of how much ethanol will be produced and processed.

There are numerous ways to estimate sugarcane yield, one of the most used being that related to agrometeorological models (Monteiro and Sentelhas, 2017) using machine learning or by using mechanistic models that aim to represent specific processes in the soil-plant-atmosphere continuum. To estimate actual sugarcane yield, climatic variables such as temperature, precipitation, solar radiation, among others are used in crop models to simulate crop growth and the possible effects of climate on sugarcane productivity. The result of these simulations has been quite accurate. However, these mechanistic models require a good knowledge and understanding of specific processes and a correct initial parameterization. Besides that, not all factors affecting sugarcane yield are considered. For example, soil was considered as one of the most important factors affecting sugarcane productivity, even more than climate (Folberth et al., 2016), but it is still poorly considered in crop models (Vereecken et al., 2016). Soil has an intrinsic potential that is mainly determined by the presence of greater or lesser clay content. This intrinsic potential may determine how much sugarcane

will be produced in a specific area, being possible to improve it by management when the potential is low (Greschuk et al., 2022).

As process-based models are still somewhat difficult to operationalize in crop yield predictions, statistical models arise as useful tools. Regression models and Machine Learning have proven their importance in this task. Studies to predict, estimate or forecast crop yield usually use historical data or information before harvest to have an idea of what would be the performance of the crop in a specific yield. In the case of using historical data, time series analysis can be an option to estimate future crop yields. However, when it comes to analize more than one geographical individual (i.e, a county or municipality), time series can be insufficient. To overcome this issue, panel data models can be useful tools to estimate or forecast future sugarcane yields. Their advantage is that unlike time series analysis, it supports the inclusion of several categories and time periods, besides other factors that can be used as predictors of sugarcane variability such as climate, soil, vegetation. Yet, panel data models are still linear regression models and although they can be better than simple time series analysis for predicting crop yields, they are unable to capture possible non-linear relationships that can be observed between environmental variables and crop yields. Machine learning (ML) models, especially those based on decision tree and ensembles were demonstrated to be more suitable for predicting crop yields. Random Forest, Gradient Boosting, Support Vector Machine are the most used models.

Since São Paulo is one of the most important producers in the country, 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. Due to data scarcity on historical yield by fields, data publicly available from governmental agencies will be used. 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.

2. Material and Methods

2.1. Study area and datasets

This research deals with estimating sugarcane yield (kg ha^{-1}) using soil, climate, and vegetation variables as predictors in a panel data analysis. The study area comprised the State of São Paulo, at the municipality level, with a total area of 248209.43 km^2.

The main dataset consists of past sugarcane yield for each municipality (642 in total) from a period of 21 years (2001-2020). This information was obtained from the SIDRA platform (https://sidra.ibge.gov.br/pesquisa/pam/tabelas) of IBGE (Instituto Brasileiro de

Geografia e Estatística). From the same platform, data such as harvested area, total production and production values were also downloaded. These datasets undergone a first pre-processing in Python, in which municipalities without sugarcane yield productivity in the whole studied period were excluded. The final dataset consisted of 486 municipalities.

Mean annual maximum and minimum temperature (°C, correction factor 0.1), accumulated precipitation (mm), soil moisture (mm, correction factor 0.1) and evapotranspiration (mm, correction factor 0.1) were obtained from TerraClimate. Vegetation parameters were represented by Normalized Difference Vegetation Index (NDVI, correction factor 0.0001) and Net Primary Productivity (NPP, kg C m^2, correction factor 0.0001), which were obtained from MODIS. Land Surface Temperature (LST, K correction factor 0.02) was also obtained from MODIS. All these environmental variables were retrieved from the Google Earth Engine platform and averaged over each municipality in QGIS (climate variables: 1km^2 resolution, vegetation and LST: 500m). All variables retrieved from the GEE platform were used in their original scale and are depicted in Appendix 1.

2.2. Statistical analysis

2.2.1. Descriptive statistics

The first step of the data analysis consisted of descriptively analyzing the dataset over municipalities and over years. Averages were calculated for each municipality for the whole period (21 years) for sugarcane yield, harvested area, total production and value. For the same variables, mean and standard deviation were calculated for each year and presented as line graphs to have an idea of the trend of these variables over time. As more than 400 municipalities were considered and was difficult to represent all of them, they were grouped and averaged in five quantiles, each representing 25% of the dataset. Spatial averages for these variables and climate, vegetation and soil variables were also presented. This spatial representation was intended to help understand the spatial variability of the variables across municipalities and over time. The statistical analysis was carried out in Python and R programming languages.

2.2.2. Panel data analysis

Panel data models describe the individual behavior both across time and across individuals, in our case, across municipalities. There are three types of panel data models: the pooled model, the fixed effects model, and the random effects model. The pooled model specifies constant coefficients and assumes that all the municipalities have the same characteristics. Since panel data include different municipalities with different characteristics, the likelihood of heterogeneity exists. Heterogeneity refers to unobserved municipality-

specific characteristics such as differences in local conditions that might affect sugarcane yield, whose differences are not considered in the regressor or independent variables.
We have an unbalanced panel.

$$Y_{it} = \beta_0 + \beta_1 \mathrm{X}_{1,it} + \beta_2 \mathrm{X}_{2,it} \ldots . + \beta_n \mathrm{X}_{n,it} + \mu_t + \omega_i + \varepsilon_{it}$$

where:

i = 1, …., N (municipalities) and t = 1,..,T (years)

Y_{it} : Yield productivity for municipality i at time t

$\mathrm{X}_{n,it}$: Environmental variable for municipality i at time t

μ_t : unobserved time-dependent error terms (factors affecting sugarcane yield productivity that vary with time but not across municipalities).

ω_i: unobserved municipality-dependent error term (location, local conditions, etc.)

ε_{it}: error term

We assume that there is unobserved heterogeneity across individuals captured ω_i.The main question is whether the individual-specific effects ω_i are correlated with the regressors. If they are correlated, we have the fixed effects model, if they are not correlated, we have the random effect models. The Hausman test was used to differentiate between fixed effects model and random effects model. The null hypothesis (p-value > 0.05) assumes that there is correlation between the individual-specific effects and the regressors, therefore, the fixed effects model should be used. In this model, dummy variables representing each individual (in our case, municipalities) are included in the model. If the null hypothesis is rejected (p-value < 0.05), the random effects model should be used. In our case, the p-value was lower than 0.05 and we used the random effects model.

The panel data analysis was carried out using data from all municipalities and the period from 2000 to 2014. The remaining years (2015-2020) were used to test the model. As data on sugarcane yield from the year 2021 is not yet available on the SIDRA platform, we predicted the sugarcane yield for this year after testing our model.

2.2.3. Machine learning

For the machine learning, all municipalities and years were treated as individual observations. Two models were tested: gradient boosting and random forest but only gradient boosting will be presented here as it had the best performance. Gradient boosting is one of the variants of ensemble methods where you create multiple weak models (decision trees) and combine them to get better performance. The gradient boosting was first carried out with the default parameters and then a grid search was applied to find the best combination of parameters. The dataset from the period 2000-2014 was used to build and test the model, and therefore partitioned into training (70%) and testing (30%), considering all

the years and municipalities as individual observations, as mentioned before (See Figure S1 in the Appendix 2 for a distribution of training and testing sets). The data from years 2015-2020 were used as the validation set. The final model was used to forecast the sugarcane yield for year 2021.

After the grid search and cross-validation procedures were applied, the following parameters were found: learning_rate=0.2, loss=huber, max_depth=20, max_features=auto, max_leaf_nodes=10, min_samples_leaf=10, min_samples_split=8, and n_estimators=700. The model was chosen based on the lowest RMSE (Root Mean Square Error). R (version 4.1) and Python (version 3.9) languages were used to analyze the data and build the models. The Google Earth Engine platform was used to obtain the environmental covariates and relevant processing procedures were performed in QGIS version 3.27 as needed.

3. Results and discussions

3.1. Descriptive statistics of yield, harvested area, total production and value

All variables were statistically described and are represented in Figures 1 and 2. Figure 1 shows histograms of sugarcane yield (Yield, kg ha^{-1}), harvested area (AC, ha), total production (QP, Tn), production value (VP, R$) and the environmental variables used to estimate the sugarcane yield: Normalized Difference Vegetation Index (NDVI), Net Primary Productivity (NPP, kg C/m^2), precipitation (PR, mm), evapotranspiration (AET, mm), soil moisture (SOIL, mm), minimum (TMIN, ºC) and maximum temperature (TMMX, ºC) and land surface temperature (LST, K) averaged over the period studied for each municipality.

Figure 2 shows the temporal variability for all variables over the same period. The average yield for all municipalities and years was 75647.63 kg ha^{-1} with a standard deviation of 7163.22 kg ha^{-1}. The temporal variability of the sugarcane yield (Figure 2, Yield), first increased until 2010, suffered a great decrease in 2014 and then showed an increase again, but more stable than the previous years. The mean harvested area for all years and municipalities was 8886.88 hectares with standard deviation of 9706.52 hectares, indicating the high variability in the area devoted to sugarcane between the municipalities, which can be confirmed by analyzing the temporal variation of the harvested area (AC) in Figure 2. It is also possible to note that the AC is constantly increasing over the years.

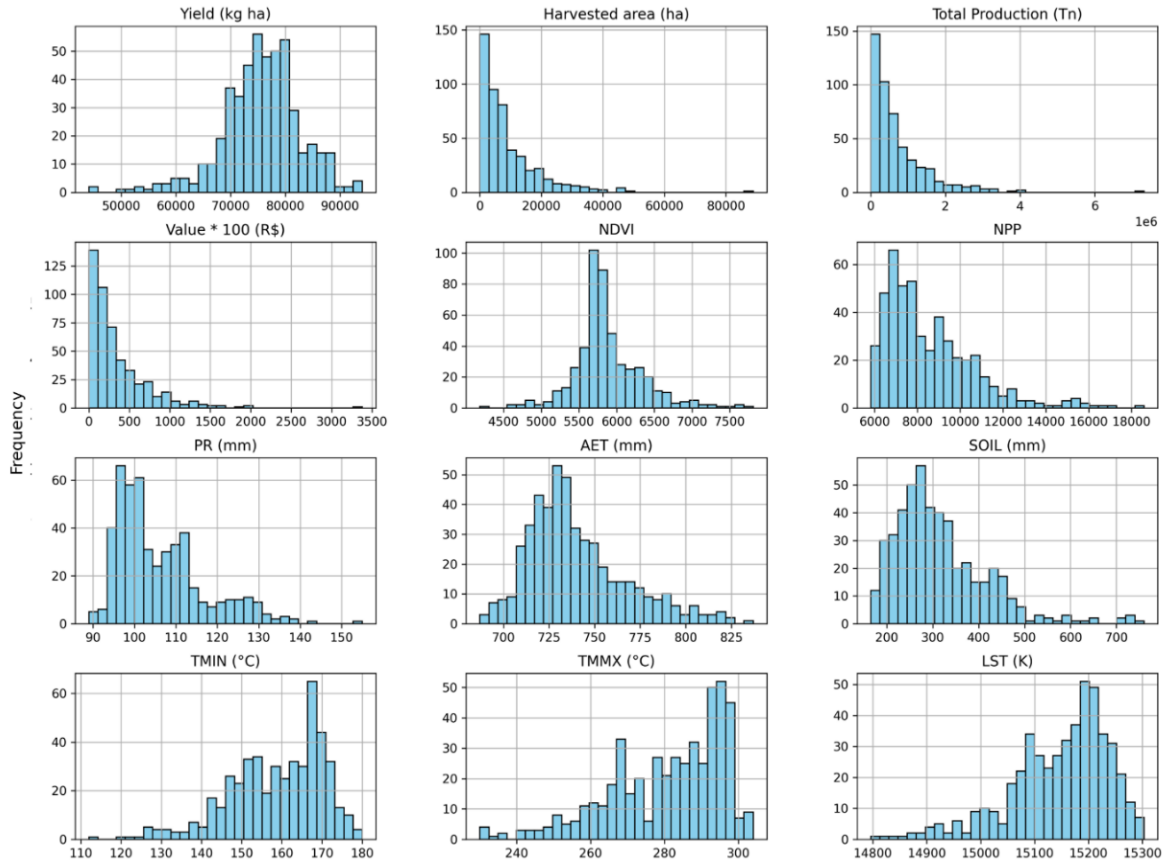


Figure 1: Distribution of the variables for all municipalities averaged over the period studied.

Although productivity has fallen in specific years, total production (QP) has shown an upward trend since the beginning of the analyzed period, as a result of the increase in the harvested area over the years (Figure 2, AC). The average total production (QP) during the period studied was 710856.70 tons with a standard deviation of 787497.70 tons, the same pattern as AC, due to the high variability between municipalities. The value of production (VP) in Figure 2 showed a considerable increase over the years, which is expected, because of economic patterns mostly related to inflation. The average value of production was 340.23 x 10 (1000 R$) with a standard deviation of 369.91 x 10 (1000 R$). Unfortunately, the SIDRA platform from where the data was collected does not inform if this value is the price of a ton of sugarcane or the average price by municipality. It is reported instead that it is a *'Derived variable calculated by the weighted average of quantity information and current average price paid to the producer, according to the harvest and commercialization periods of each product. Shipping charges, fees and taxes are not included in the price'*.

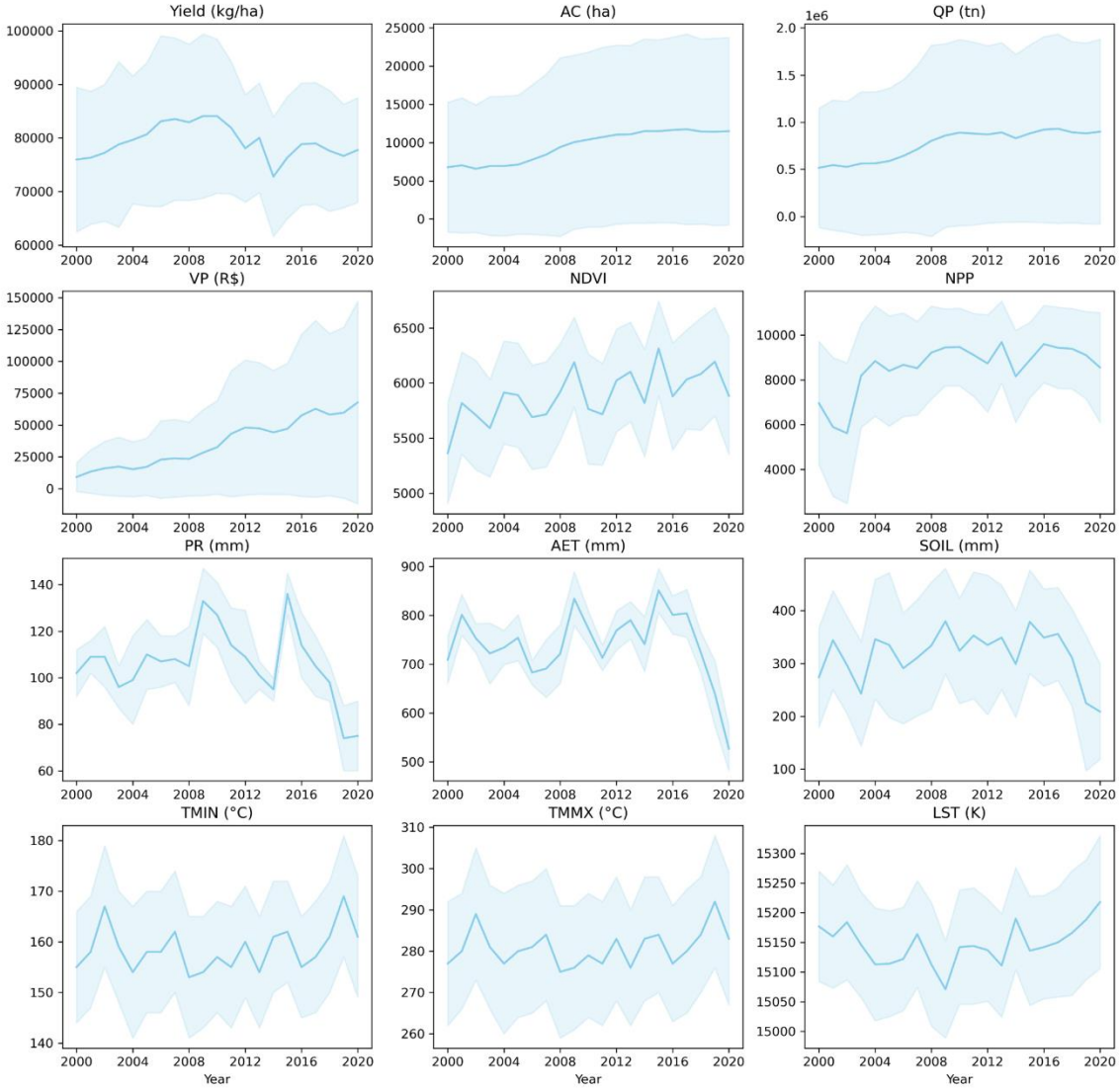


Figure 2. Average of the variables for each year. AC: harvested area, QP: total production, VP: value, NDVI: Normalized Difference Vegetation Index, NPP: Net Primary Productivity, PR: Total precipitation, AET: Evapotranspiration, SOIL: Soil moisture, TMIN: Minimum temperature, TMMX: Maximum temperature, LST: Land surface temperature. NDVI and NPP have a multiplied factor of 0.0001. AET, SOIL, TMIN AND TMMX have a multiplied factor of 0.1. LST has a factor of 0.2.

The environmental covariates used for modeling are also depicted in Figures 1 and 2. These variables were downloaded from the Google Earth Engine platform and their spatial distribution in raster format can be found in Appendix 1. The NDVI showed an average value of 0.58 (Figure 1, multiplied factor of 0.0001) with a standard deviation of 0.04, indicating that its variability over time was low, as can be observed in Figure 2. The NPP had a mean of 0.86 kg C m^2 with a standard deviation of 0.21 kg C m^2. This variable had a decrease in its value between 2000 and 2004 and then a constant increase over the next years after this decrease. Precipitation (PR), evapotranspiration (AET) and soil moisture (SOIL) had an

erratic pattern over the period studied, which a great decrease in 2020. Their averages and standard deviations were 105±10 mm, 73.9±2.7 mm, and 31.6±10.0 mm, respectively. The minimum and maximum temperature (TMIN-TMMX, °C) were also constant over the period studied, with means and standard deviation of 15.8±11 and 28±15, respectively. The land surface temperature also was constant but is showing an increasing pattern since 2016, probably due to the high exposition of soil surface. The mean and standard deviation was 302.09±1.82 K (29.75 ºC).

Figure 3. Average over each year for all the variables. AC: harvested area, QP: total production, VP: value, NDVI: Normalized Difference Vegetation Index, NPP: Net Primary Productivity, PR: Total precipitation, AET: Evapotranspiration, SOIL: Soil moisture, TMIN: Minimum temperature, TMMX: Maximum temperature, LST: Land surface temperature. NDVI and NPP have a multiplied factor of 0.0001. AET, SOIL, TMIN AND TMMX have a multiplied factor of 0.1. LST has a factor of 0.2.

With the objective of reducing the number of observations to include in the PDA, we clustered the municipalities into five groups, based on five quantiles of the sugarcane yield. The other variables were divided based on these groups (Figure 3). While the production variables (Yield, VP, QP, and AC) were properly separated in five groups, the other environmental covariables were difficult to separate, as they were not divided into their own quantiles. These quantiles were used to model sugarcane yield in the PDA but not in the ML modeling. The reasons are explained in the next section.

3.2. Model performances of PDA and ML

3.2.1. Panel data analysis with all municipalities and by quantiles

The PDA performed on all municipalities and by quantiles provided RMSE values of 9157.55 and 17020.19 kg ha^{-1}, respectively. The observed and predicted values are depicted in Figure 4. The results were not satisfactory, especially for the model built using the quantiles. With all municipalities, the model was better but still not enough to explain the variability of the sugarcane yield.

The values predicted for the period 2015-2020 for both models is also presented in Figure 4. The estimated values for all municipalities were more similar to the observed values by the end of the period. Those from the PDA with quantiles, conversely, were quite far from the observed values, not even similar to them at any time. This can be explained by the fact that, although the groups were well divided by yield, the explanatory variables did not follow this pattern, being impossible to obtain a satisfactory model if the environmental variables are not able to explain the variability of the sugarcane yield.

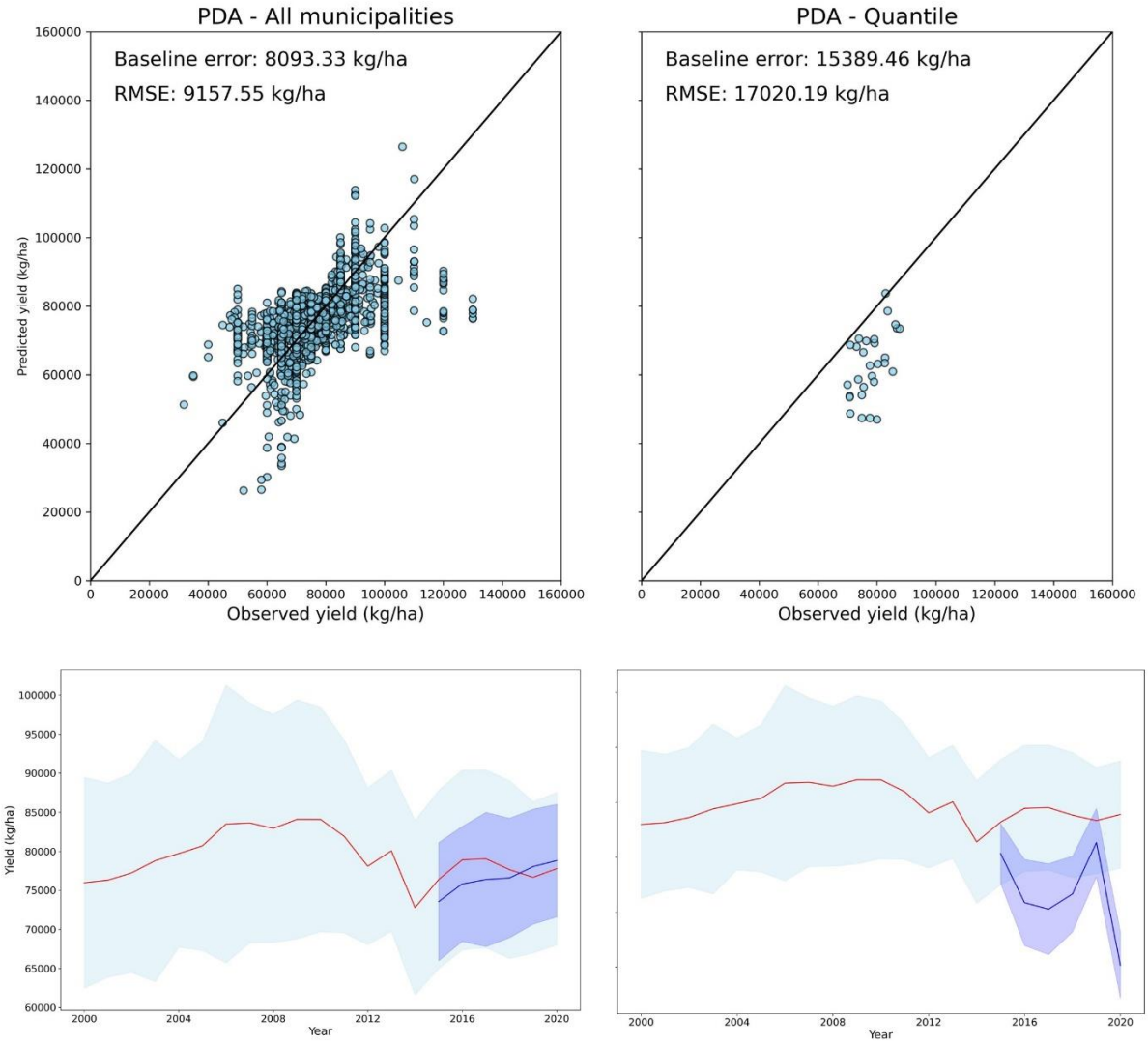


Figure 4. Prediction of sugarcane yield for the test set covering the years 2015-2020 (blue line with blue shadow) compared to observed values (red line). Observed vs predicted sugarcane yield for test set (period 2015-2020)

Among the variables used in the PDA for all municipalities, the accumulated precipitation had the highest negative impact (-27.773) followed by harvested area (-4.342) and soil moisture (-3.068). NDVI also had a negative impact, but it was the lowest (-0.9348). The variables that had the highest positive impact were minimum (54.03) and maximum (45.78) temperature, indicating that an increase in these variables would favor the increase in sugarcane yield, explained by the fact that C4 plants are more efficient under high temperature (Sage and Kubien, 2007). These results confirm the findings of Guo et al. (2021), who reported a little increase in sugarcane yield after the increase in temperature when evaluating sugarcane resilience in a subtropical region in China. Regarding the precipitation, contrary that was found here, the authors reported an improvement in

sugarcane yield after an increase in precipitation, but this was hold until a certain limit, above which the sugarcane yield tends to decrease. The LST also had a positive impact on sugarcane yield (14.84), but this result is not following common patterns, as an increase in LST is indicating a high emission from surfaces, which may indicate bare soil, without residues or a drier surface. QP, VP and AET had a negligible influence in the model (coefficients less than 0.05).

Although the PDA performed on all municipalities had a better performance than that observed for the quantiles, the results are still poor. This can be due to the non-linear nature of the relationship between the sugarcane yield and the environmental covariates that were used to build the model. There are not so many studies using PDA to forecast sugarcane yield but there are plenty of examples that used PDA to forecast the yield of other crops. Most of the works based on PDA used it to measure the effect of one or two variables in crop yield variability. Yun and Graming (2022), for example, used county-level corn yield and time-invariant location characteristics in panel regressions to estimate the effects of weather on yields. Their results were significant, and they found that the fixed model effect was one of the best. Similar works were developed in India to evaluate the projected sugarcane yield under different climate change scenarios, estimating how much the sugarcane yield would change with a 1-unit change in climatic factors (Jyoti and Singh, 2020). Different from our work, they estimated the change in sugarcane yield in different future periods (2040, 2060, 2080, and 2100). The authors used state-wise information from the 1970-2017 period. Their results suggested a decrease in sugarcane yield, mostly impacted by temperature and rainfall.

In a similar work, Silva et al. (2019) studied the effects of climate change on sugarcane production in a specific region in Brazil during the period comprising the years from 1990 to 2015. Different from this work, they did not aim at estimating or forecasting sugarcane yield, but instead how change in climatic factors would affect sugarcane yield. They found a positive correlation of rainfall with sugarcane production whereas a negative correlation with temperature was observed and concluded that under intensified climate change, sugarcane yield can be lowered and contribute to great losses in sugarcane productivity.

3.2.2. Machine learning to forecast sugarcane yield

As the PDA showed poor results, two ML models were tested to find out if the performances can be improved. Random Forest (RF) and Gradient Boosting (GB), with their default hyperparameter values were run with the same set of variables used in the PDA. As the RF also showed a poor performance and specially a high overfitting, it was excluded, and

the results are no longer presented here. The analysis continued with the GB, first with the default parameters and then with tuning (testing and selecting a set of hyperparameters). The results are presented in Figure 5.

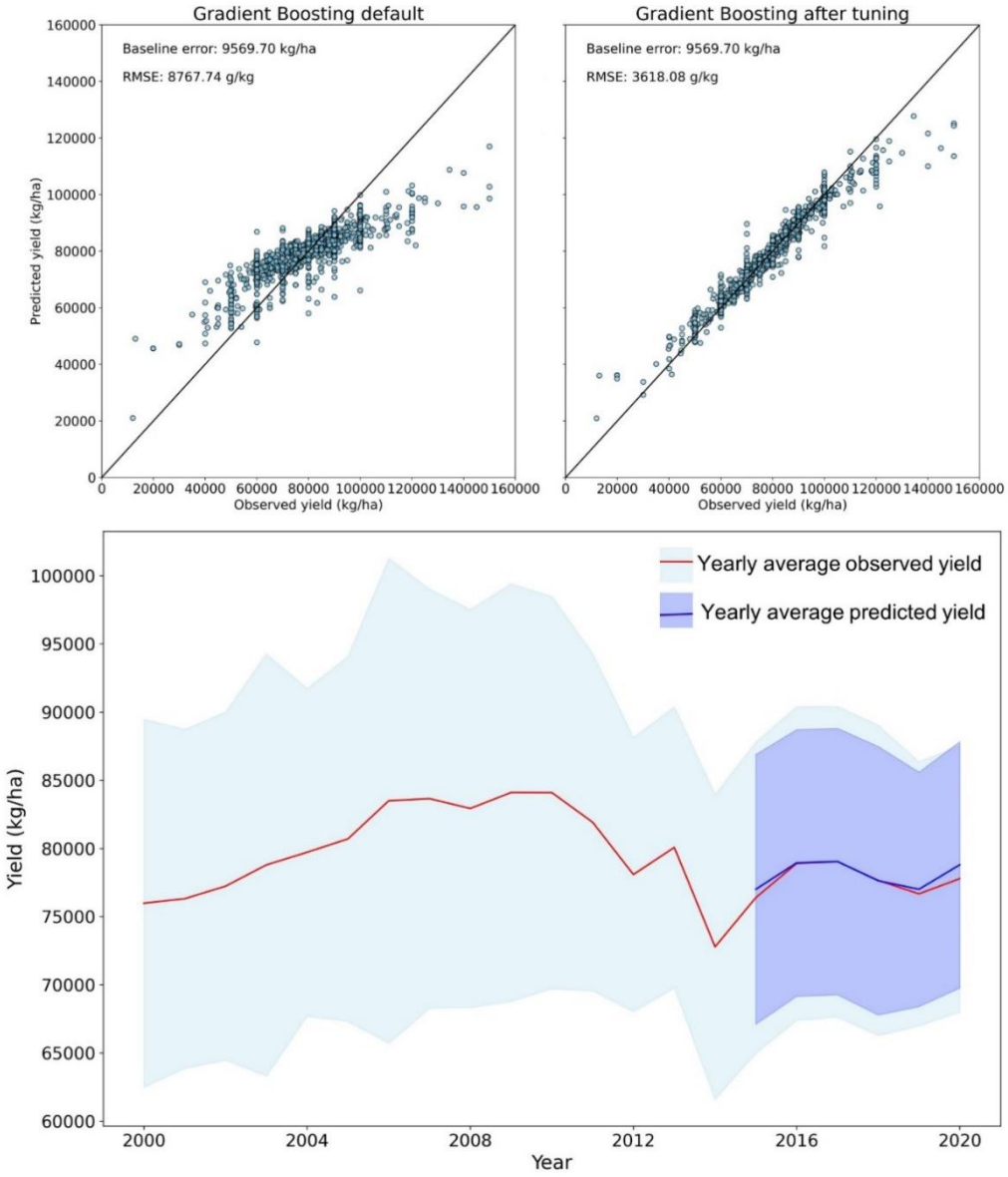


Figure 5. Observed and predicted sugarcane yield (kg ha^{-1}) after applying gradient boosting without hyper parameterization (Gradient Boosting default, top left) and with tuning (Gradient Boosting after tuning, top right). Observed (2000-2020) vs predicted sugarcane yield (2015-2020) after applying the Gradient Boosting model (bottom).

The RMSE of the GB default (with no hyperparameter tuning) was 8767.74 kg ha^{-1}, which was slightly better than the baseline error (9569.70 kg ha^{-1}, Figure 5). After the hyper parameterization, the model improved considerably, and we found an RMSE of 3618.08 kg ha^{-1} and R^2 of 0.92. This was the best result obtained after a series of tuning. The estimated

average sugarcane yield for the period 2015-2020 (test set) is also presented in Figure 5 and it can be observed that the predicted average values are almost equal to the observed values, demonstrating the excellent performance obtained with this model.

The importance of the variables was measured by the SHAP algorithm, which is presented in Appendix 2, Figure S2. Although the known importance of the environmental covariates used here to explain the sugarcane variability, the production variables (i.e., QP, AC, and VP) followed by NPP and soil moisture were the most important, particularly the QP. When the QP increases, there is a tendency to increase the sugarcane yield. This can be probably related to better local soil conditions or specific management practices. Unfortunately, due to data scarcity and the level analyzed in this study, it was not possible to include variables related to soil management or sustainable practices. Usually, in ML models, temperature, rainfall, soil types and management practices are more commonly used (Klompenburg et al., 2020) and they can explain crop yield variability and forecast future trends. In the case presented here, production variables such as harvested area and total production were included due to the low predictability of the environmental covariates used. This can be explained by the fact that this is a large-scale analysis and the variables used were averaged over municipalities, which probably had a negative influence in explaining the relationship with the variables and the sugarcane yield.

There are increasing interest in using ML models to estimate or forecast crop yield due to their ability to learn the non-linear relationship between the yield and the environmental covariates. Gradient Boosting, particularly, is a decision-tree-based algorithm which has proved very powerful in several studies. Chea et al. (2022) tested different models and environmental covariates to estimate sugarcane Brix content. They tested different combinations of variables and concluded that Gradient Boosting was the best model to estimate the Brix content. Among the combination of variables used, vegetation indices, estimated sugarcane height, and rain event data were the most suitable. Huber et al. (2022) used the Extreme Gradient Boosting (XGBoost) algorithm and compared its performances in predicting crop yield with those from Deep Learning (DL). Although they found that DL outperformed XGBoost, they also acknowledge the fact that DL's interpretability is still low when compared to decision trees and highlighted the importance of using SHAP to understand the influence of the variables used. The variables used in their work were MODIS reflectance bands, MODIS temperature and DAYMET precipitation and vapor pressure, being the reflectance bands the most important predictors in their model.

Poudyal et al. (2022) aimed at predicting sugarcane yield and select genotype by using hyperspectral imagery and the gradient boosting algorithm. In their case, they used images from specific months (April, July, and September) to determine the best timing of

yield prediction, which is somewhat different from the approach presented here, where historical yields were used to estimate the yield of specific years and forecast the yield of a year (2021). They found that the sugarcane yield was accurately predicted by using images from July. This type of research can help select specific-date satellite images to estimate sugarcane crop yields at the regional level. In this research, the images were averaged for each year of the period studied (21 years), which can negatively influence the relationship between the variables and sugarcane productivity. Future works can focus on selecting the best single-date images to explain the sugarcane yield instead of averaging over a year.

3.3. Applying PDA and GB to the validation set

After the evaluation of the PDA and GB models, they were applied to the data from the year 2021. At first, the data from this year was devoted to forecasting purposes, because the observed values of sugarcane yield were not available yet in the SIDRA platform when this research started. At the time the data analysis was almost complete, the data were made available and therefore it is possible to use this data as an external validation. The results of the observed and predicted values are presented in Figure 6. The RMSE for both models were higher than that observed in the testing set and that of PDA (11663.29 kg ha^{-1}) was even higher than the baseline error (9569.70 kg ha^{-1}). Therefore, we can confirm the suitability of the Gradient Boosting with hyperparameterization for predicting and forecasting the sugarcane yield over the State of São Paulo.

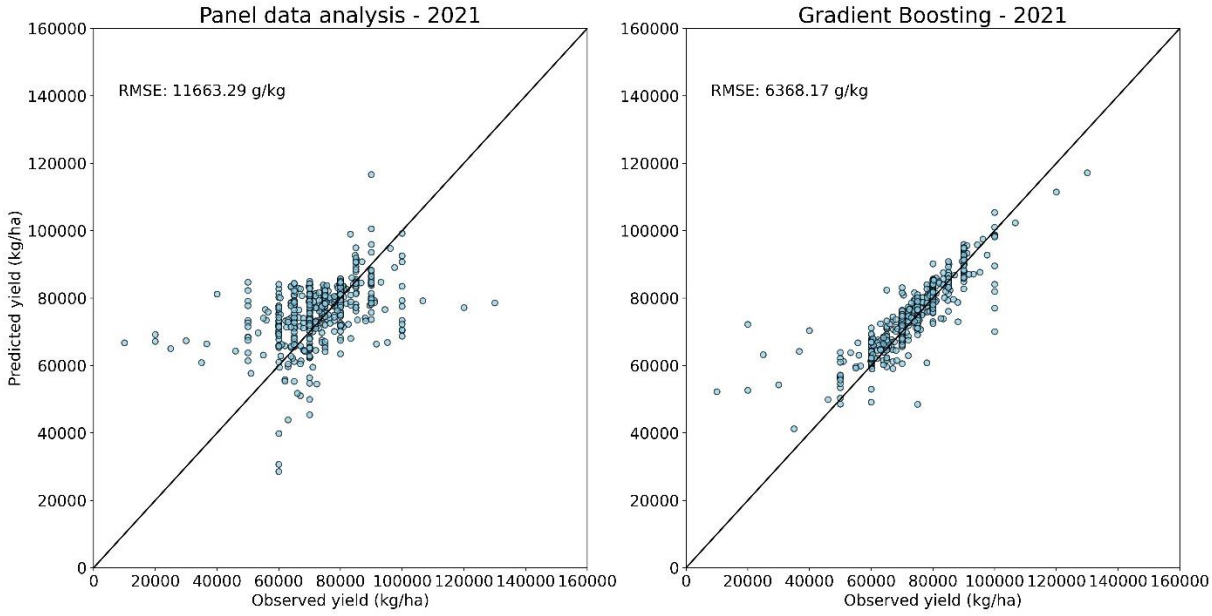


Figure 6. Observed vs predicted sugarcane yield for year 2021 (validation set) after applying the Panel Data Analysis and Gradient Boosting model.

To spatially visualize the results of each model (PDA and GB), the predicted value for each municipality in year 2021 is presented in Figure 7. The observed values for the year 2021 are also presented to allow comparison. The areas were each model predicted lower or higher yields and their differences is presented in Figure S3 of Appendix 2. As the GB model was the best in performance, it also performed well on this external dataset of year 2021. As can be seen from Figure 7 and Figure S3, most of the predicted yield values with GB were higher than the observed yield, but this difference was not superior to 10000 kg. This higher predicted yield was concentrated on the central part of the state.

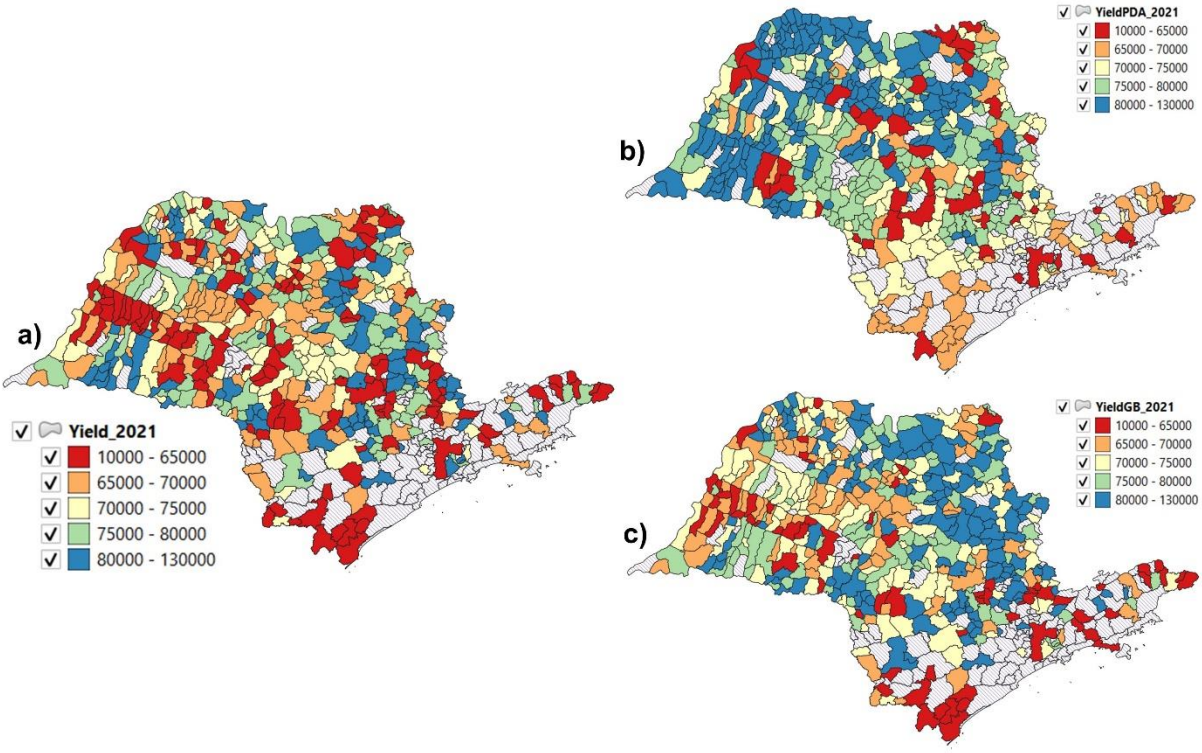


Figure 7. Observed sugarcane yield for year 2021 (a) and predicted sugarcane yield for year 2021 by panel data analysis (b) and gradient boosting (c). Sugarcane yield is expressed in kg ha^{-1}

The predicted values for year 2021 from the PDA were even worst than those results observed with the test set (2014-2020, Figure 4), with a high RMSE and overestimation of the yield, especially in the north part of the state (see Figure S3 in Appendix 2). In this region, the values were overestimated in almost 25000 kg, which is very high. With this model only a few municipalities showed predicted yield below the observed yield.

4. Final considerations

Crop yield prediction or forecasting is a useful tool for different stakeholders from farmers to politicians when it comes to understand how crops will perform in a specific season. In this research, PDA and ML algorithms were tested to find out the best one to estimate sugarcane yield using historical data. The PDA did not show a satisfactory performance for estimating sugarcane yield at the municipality level neither when divided as quantiles. On the ML side, Random Forest performed well with the default hyperparameters, but the overfitting was extremely high. Gradient Boosting was the best algorithm, with RMSE lower than 4000 kg ha^{-1}, improved by a set of hyper parameterization runs. In the validation set, which was the year of 2021, again the GB algorithm was better than the PDA.

The PDA was not able to explain historical and future sugarcane yield variability. This is probably due to the linear nature of the algorithm, which the GB algorithm was able to explore by building a set of decision trees and weak learners sequentially. Estimating or forecasting sugarcane yield can benefit stakeholders and for building public policies related to the availability of sugarcane for a specific season. Although this research was focused on using historical sugarcane yield to test the models, future research can focus on developing models using environmental variables such as satellite images of specific dates and months before harvesting start. This can be more useful locally or at specific farms, but at the municipality level stakeholders would have an idea of how much a state or a region will produce. Finally, more research is needed to explore other environmental covariates and specifically management practices that were not available at the municipality level and when this research was conducted.

Code and data availability

All datasets are freely available in the SIDRA and Google Earth Engine platforms. Python, R and GEE code with a description of the main steps can be found in the links: https://github.com/neli12/time-series-productivity-sp, https://www.youtube.com/watch?v=Od1_6oM1NXU&list=PLhpCzBIsNmgpH4vahrHl7ybSivwzhBRFN.

References

Amorim, M.T.A., Silvero, N.E.Q., Bellinaso, H., Rico, A.M.R., Greschuk, L., Rabelo, L., Demattê, J.A.M., 2022. Impact of soil types on sugarcane development monitored over time by remote sensing. Precision Agriculture 1–21.

Chea, C., Saaengprachatanarug, K., Posom, J., Saikaew, K., Wongphati, M., Taira, E. 2022. Remote Sensing Applications: Society and Environment 26: 100718

CONAB. 2021. Sugarcane Crop Tracking 2020/2021 (in Portuguese: Acompanhamento de Safra brasileira de cana 2020/2021). National Supply Company of Brazil, 4(4), 1–62.

De Oliveira Bordonal, R., Luís, J., Carvalho, N., Lal, R., Barretto De Figueiredo, E., Gonçalves De Oliveira, B., Scala, N. La, 2018. Sustainability of sugarcane production in Brazil. A review. Agronomy for Sustainable Development 38: 13–36.

Folberth, C., Skalský, R., Moltchanova, E., Balkovič, J., Azevedo, L.B., Obersteiner, M., van der Velde, M., 2016. Uncertainty in soil data can outweigh climate impact signals in global crop yield simulations. Nature Communications 71(7): 1–13.

Greschuk, L., Demattê, J.A.M., Silvero, N.E.Q. Soil potential productivity for the entire Brazil. In press.

Guo, H., Huang, Z., Tan, M., Ruan, H., Awe, G. O., Are, S. A., Abegunrin, T. P., Hussain, Z., Kuang, Z., Liu, D. 2021. Smart Agricultural Technology 1:100014

Huber, F., Yunsshchenko, A., Stratmann, B., Steinhage, V. 2022. Extreme Gradient Boosting for yield estimation compared with Deep Learning approaches. Computers and Electronics in Agriculture 202: 107346.

Jyoti, B., Singh, A.K. 2020. Projected sugarcane yield in different climate change scenarios in Indian State: A state-wise panel data exploration. International Journal of Food and Agricultural Economics 8(4): 343-365

Klompenburg, T., Kassahum, A., Catal, C. 2020. Crop yield prediction using machine learning: A systematic literature review. Computers and Electronics in Agriculture 177: 105709

Monteiro, L.A., Sentelhas, P.C., 2014. Potential and actual sugarcane yields in Southern Brazil as a function of climate conditions and crop management. Sugar Tech 16(3): 264–276.

Monteiro, L.A., Sentelhas, P.C., 2017. Sugarcane yield gap: can it be determined at national level with a simple agrometeorological model? Crop Pasture Science 68: 272–284.

Poudyal, C., Costa, L. F., Sandhu, H., Ampatzidis, Y., Odero, D. C., Arbelo, O. C., Cherry, R. H. 2022. Sugarcane tield prediction and genotype selection using unmanned aerial vehicle-based hyperspectral imaging and machine learning. Agronomy Journal 114:2320-2333.

Rahman, M.M., Robson, A.J., Rahman, M.M., Robson, A.J., 2016. A novel approach for sugarcane yield prediction using Landsat time series imagery: A case study on Bundaberg Region. Advances in Remote Sensing 5, 93–102.

Sage, Rowan F., Kubien, David D. 2007. The temperature response of C3 and C4 photosynthesis. Plant, Cell and Environment 30: 1086-1106.

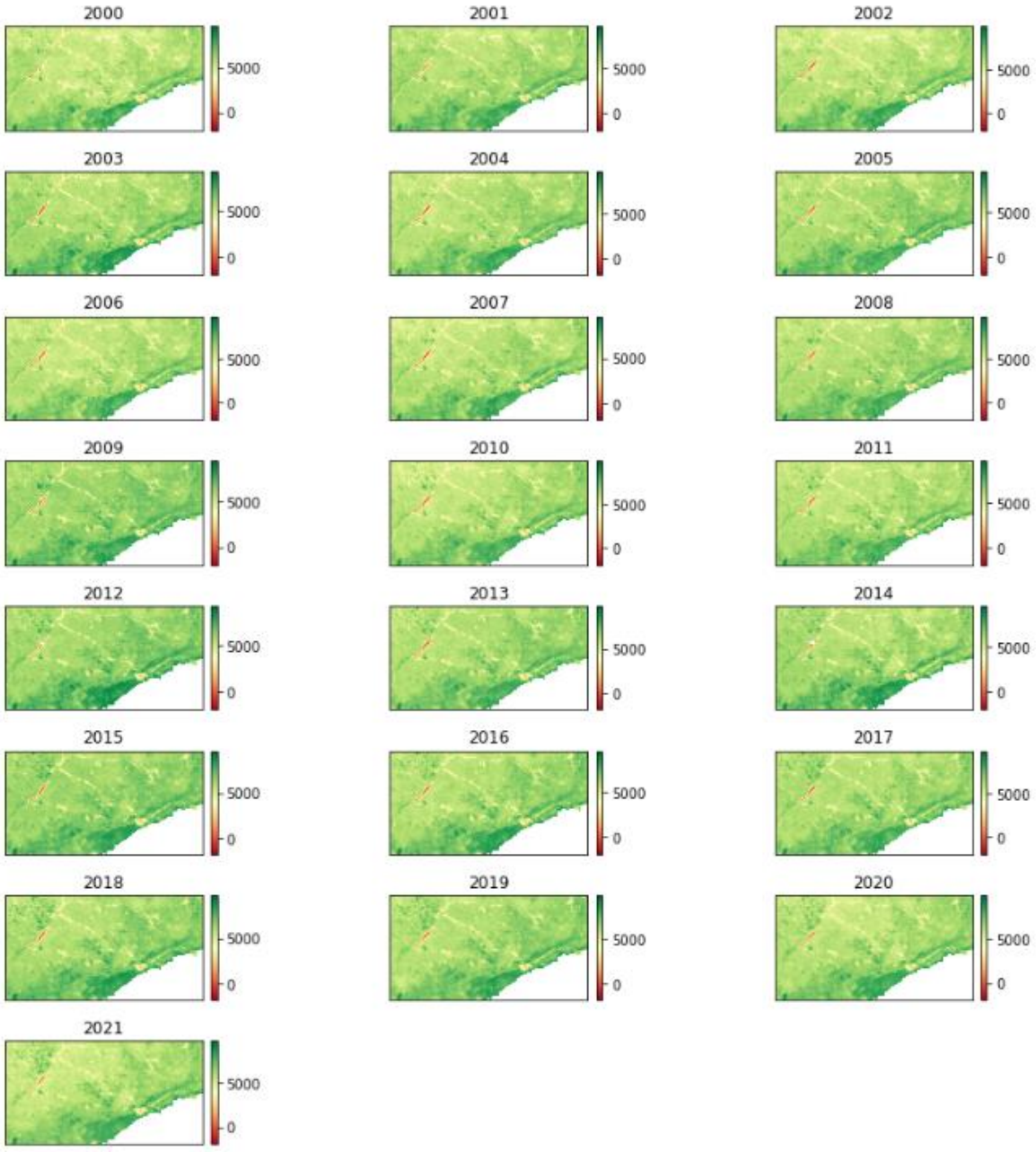
Silva, W. K. M, de Freitas, G. P., Coelho Junior, L. M., Pinto, P. A. L. A., Abrahão, R. 2019. Effects of climate change on sugarcane production in the State of Paraíba (Brazil): a panel data approach (1990-2015). Climatic Change 154:195-209

Vereecken, H., Schnepf, A., Hopmans, J.W., et al. 2016. Modeling soil processes: Review, key challenges, and new perspectives. Vadose Zone Journal 15(5):1-57.

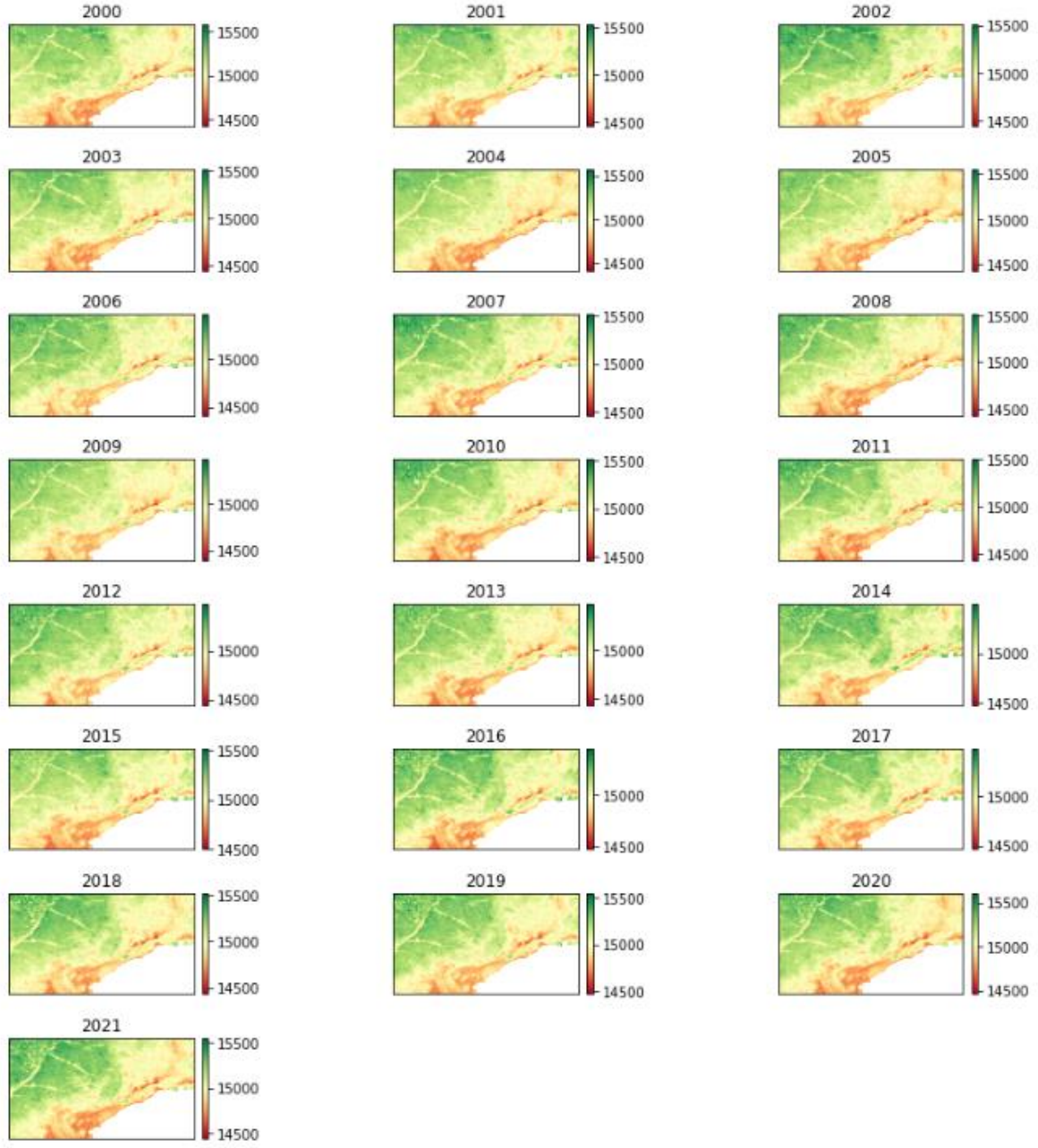
Yun, S. Graming, B. 2022. Spatial panel models of crop yield response to weather: Econometric specification strategies and prediction performance. Journal of Agricultural and Applied Economics 54(1):53-71

Appendix 1 – Spatial distribution of the environmental covariates

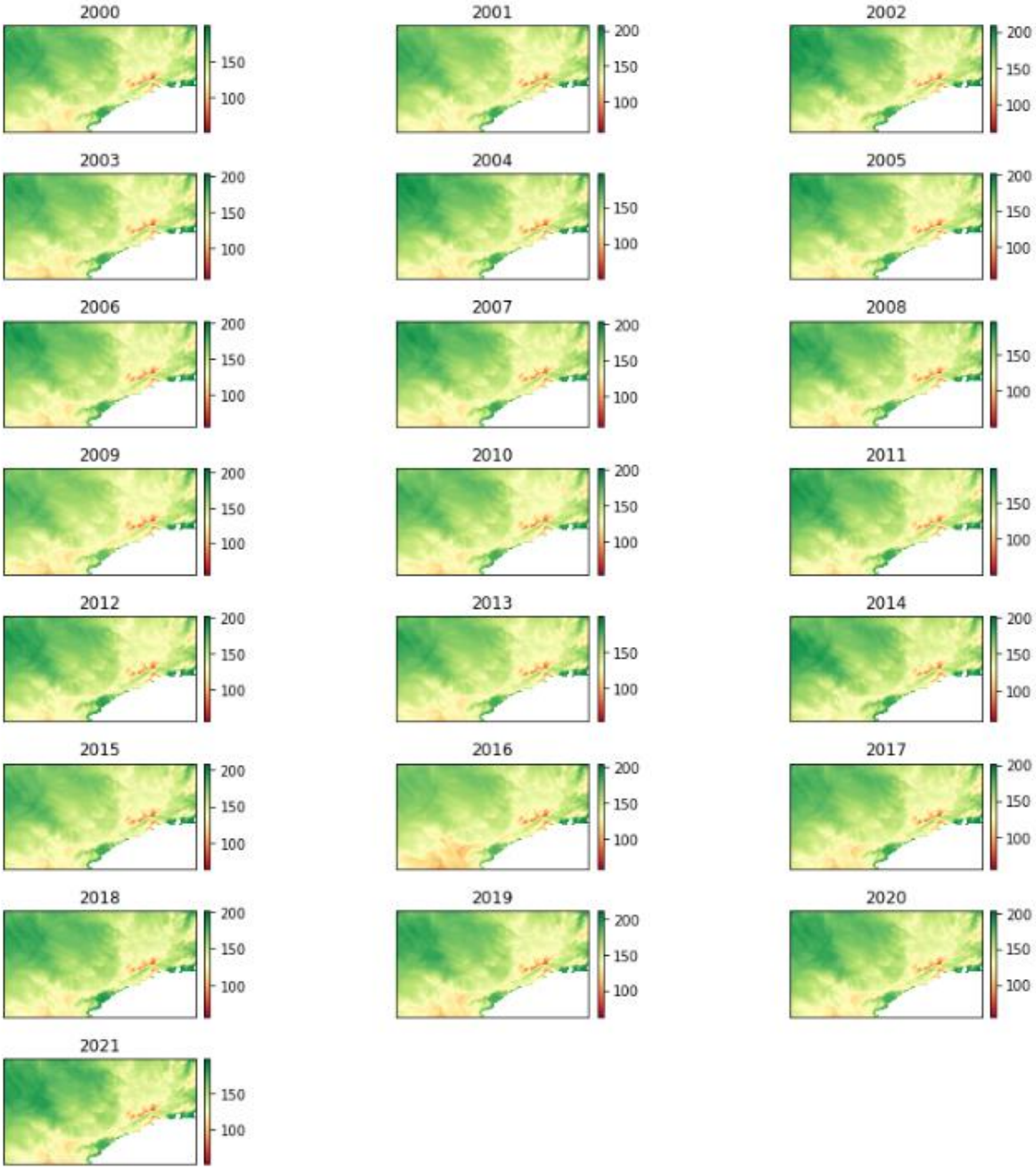
NDVI (Normalized Difference Vegetation Index) over the years (correction factor 0.0001)



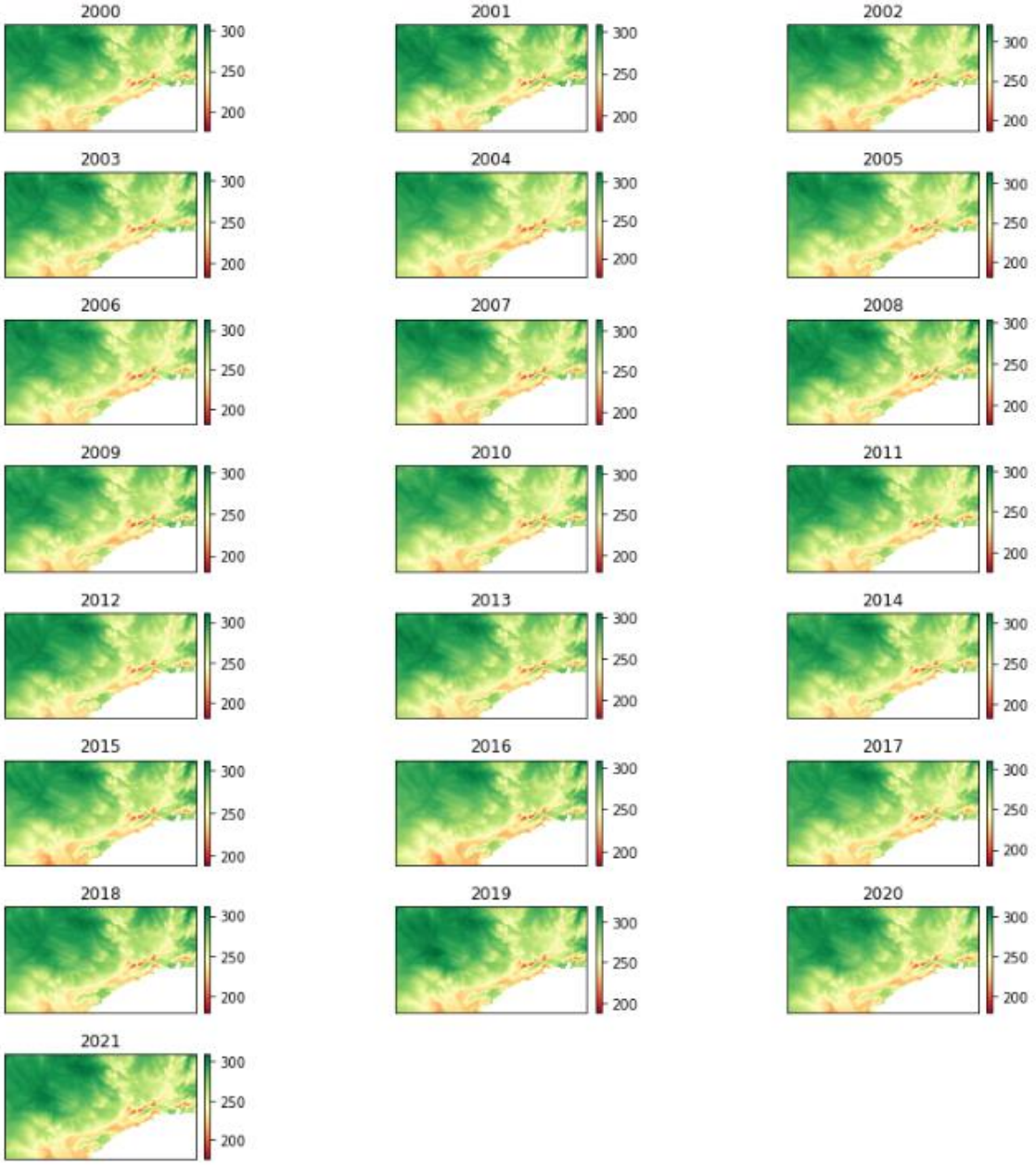
LST (Land Surface Temperature, K) over the years (correction factor 0.02)



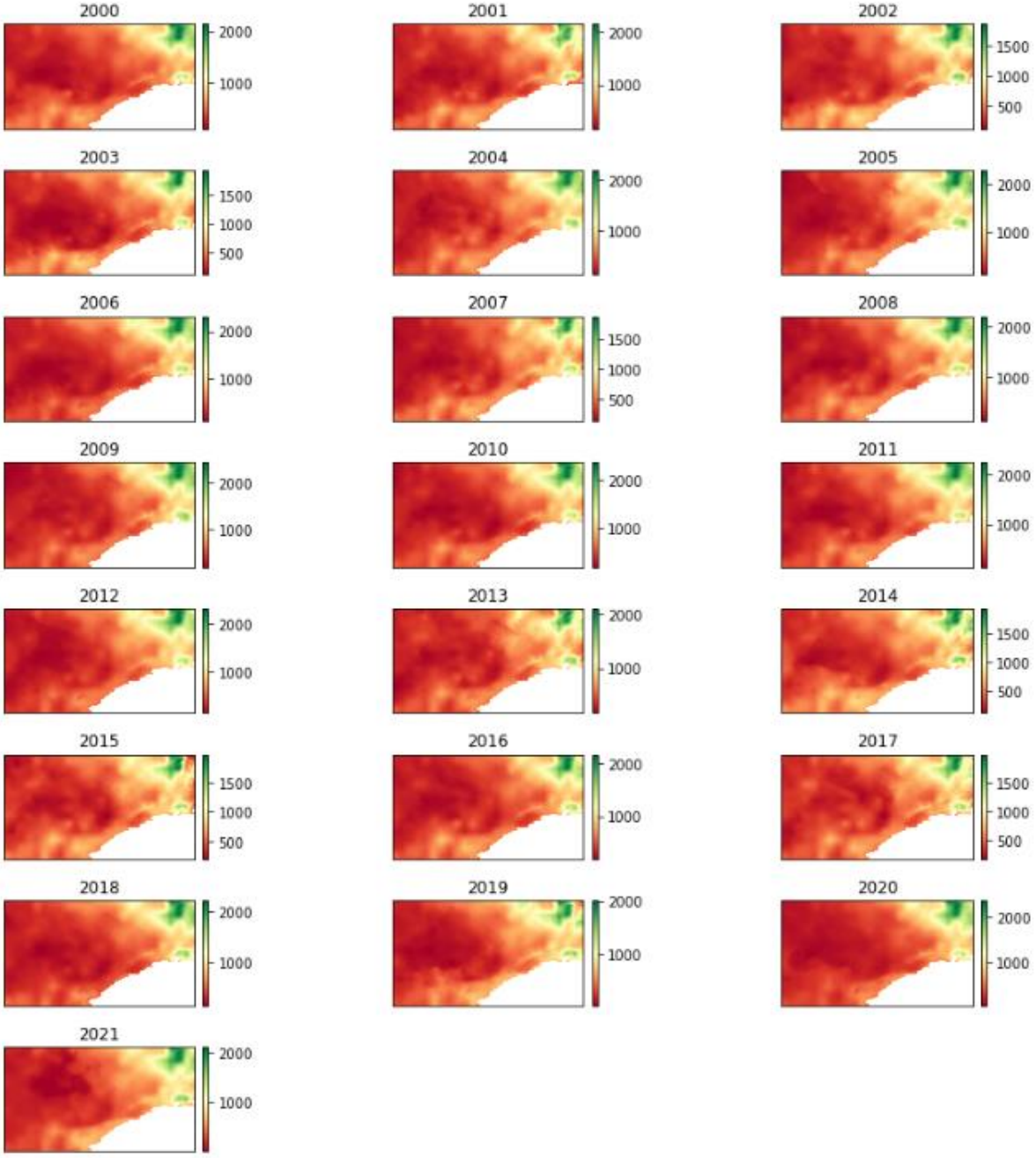
TMIN (Minimum temperature, ºC) over the years (correction factor 0.1)



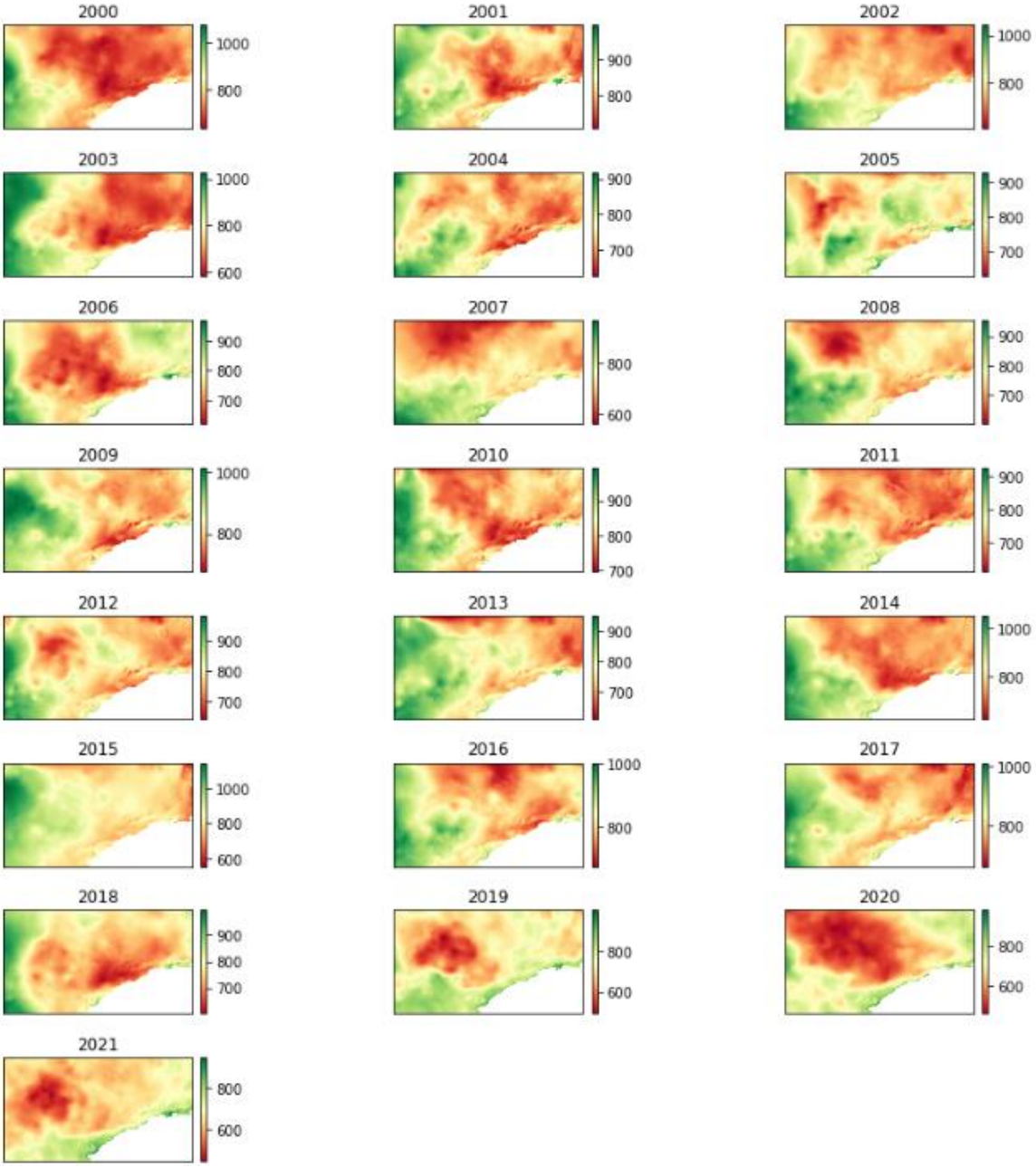
TMMX (Maximum temperature, ºC) over the years (correction factor 0.1)



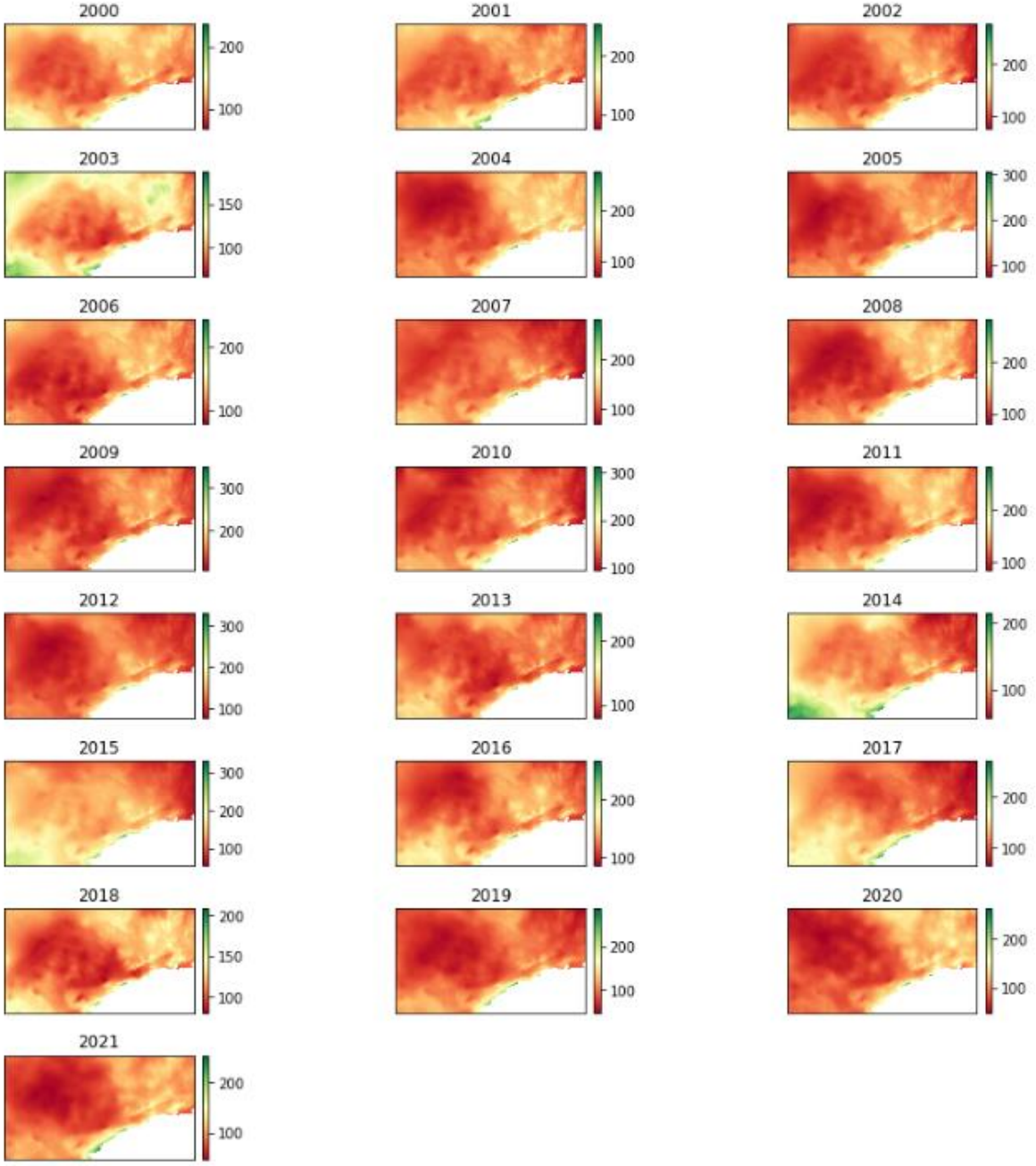
SOIL (Soil moisture, mm) over the years (correction factor 0.1)



AET (evapotranspiration, mm) over the years (correction factor 0.1)



PR (precipitation accumulation, mm) over the years



Appendix 2

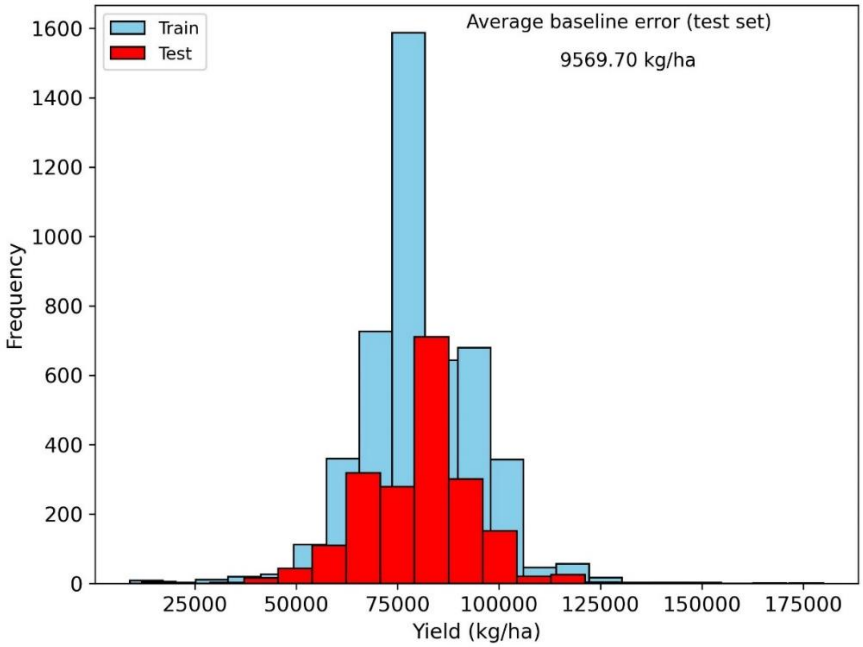


Figure S1. Histograms of training and validation sets used in the machine learning framework for the period 2000-2014. The average baseline error of the test set (average yield minus observation values) is also depicted.

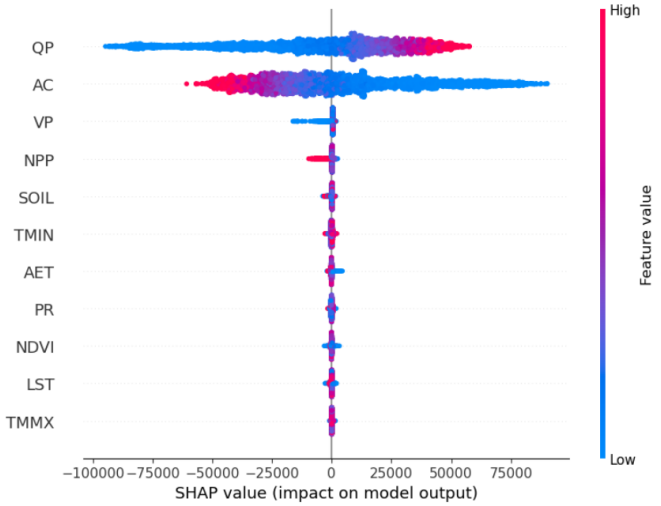


Figure S2. Contributions of the environmental covariates to the GB model. QP: Total production, AC: harvested area, VP: Production value, NPP: Net Primary Productivity, SOIL: soil moisture, TMIN: Minimum temperature, AET: evapotranspiration, PR: precipitation, NDVI: Normalized Difference Vegetation Index, LST: Land surface temperature, TMMX: Maximum temperature

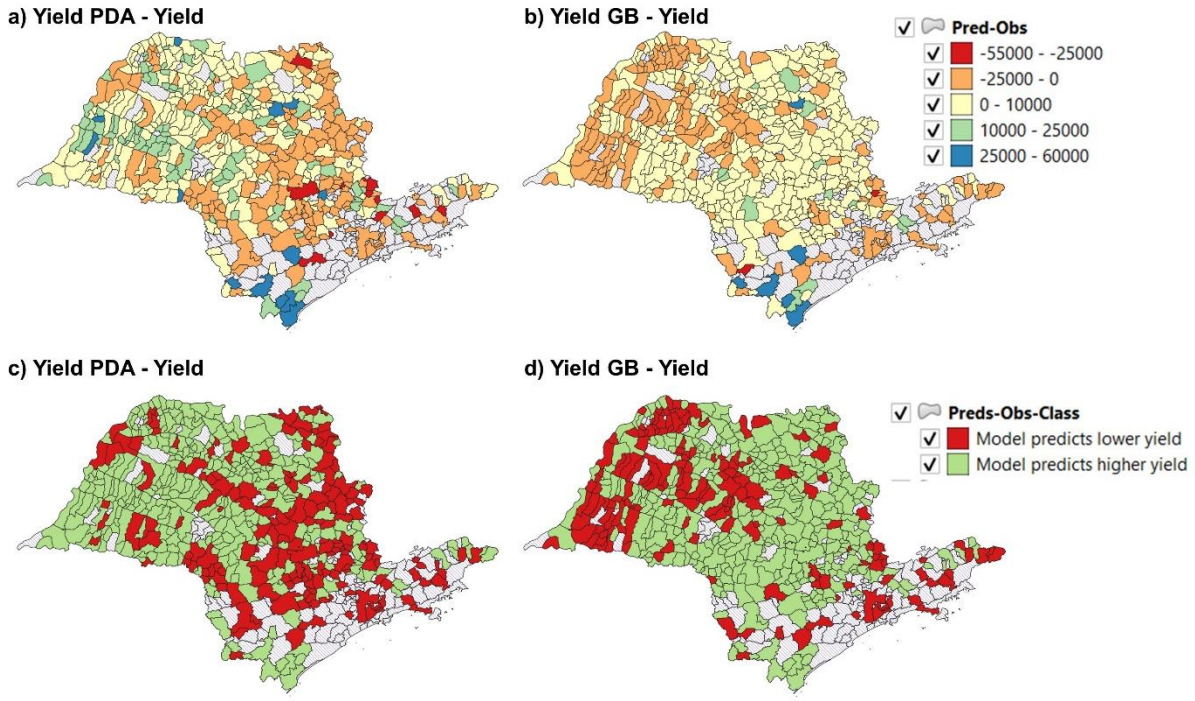


Figure S3. Differences between predicted yield and observed yield by panel data analysis (a) and gradient boosting (b). Bottom figures (c, d) represent the area where the models predicted lower or higher yield than that observed.