Detrending Fieldprint Platform Yield-based Metrics Using NASS Data

Agustin Alesso, Ph.D.[1], Nicolas Martin, Ph.D.[1]

[1]Department of Crop Sciences, University of Illinois at Urbana-Champaign

Status of Document:	Version 1 – Effective 04 28 2021
Contact Details:	ecoronel@fieldtomarket.org

TABLE OF CONTENTS

INTRODUCTION

About Field to Market

Field to Market: The Alliance for Sustainable Agriculture brings together a diverse group of grower organizations: agribusinesses; food, beverage, apparel, restaurant, and retail companies; conservation groups; universities; and public sector partners to define, measure, and advance the sustainability of food, fiber, and fuel production in the United States.

Field to Market seeks to drive continuous improvement in the sustainability of commodity crop production by uniting the agricultural supply chain and key stakeholders around a common measurement framework; nearly 150 diverse stakeholders across the food and agriculture value chain have joined Field to Market to scale sustainable agriculture. Field to Market engages in broad communication and collaboration with stakeholders to ensure a coordinated, outcomes-based approach to sustainable agriculture grounded in science.

Field to Market has developed the Fieldprint® Platform, a pioneering assessment framework that empowers brands, retailers, suppliers, and farmers at every stage in their sustainability journey to measure the environmental impacts of commodity crop production and identify opportunities for continuous improvement.

About the Fieldprint Platform

Farmers can access this free and confidential tool through the online Fieldprint Platform or associated farm-management software that integrates the Platform's metrics and algorithms. Brands, retailers, and suppliers can access aggregated data from farmers who opt-in to participate in projects registered in the Continuous Improvement Accelerator. Over 50% of acres measured in the Fieldprint Platform are analyzed via an API run by one of Field to Market's seven qualified data management partners.

Crop management information entered into the Fieldprint Platform is analyzed and processed into a Fieldprint Analysis, which presents the sustainability performance of a farmer's unique operation. This free, online tool helps growers voluntarily analyze how their management choices impact natural resources and operational efficiency.

Objectives

One challenge with analyzing Fieldprint Platform data is detecting improvements in metrics scores that are highly sensitive to yearly variations in weather, yield, management, and other factors that may fall outside of farmers' control. This may preclude being able to make meaningful comparisons or detect trends due to management in multi-year projects.

The objective of the work described in this report was to develop a procedure to remove the effect of inter-annual yield variation or trend from yield-scaled sustainability metrics, such as the Energy Use or Greenhouse Gas Emissions Metrics calculated by the Fieldprint Platform. By removing the overall annual effect on yields, it is hypothesized that more accurate comparisons can be conducted between metric scores, and relevant trends can be detected.

DATA

Fieldprint Platform Data

The demonstration data provided by Field to Market to the University of Illinois at Urbana-Champaign constituted a sample of farm-level productivity and management practices processed by the Fieldprint Platform. The following is a sample of a few of the first few columns of 5 records taken randomly from the original data file.

Table 1. Glance of Fieldprint Platform data output

Grower ID	Field Name	Year	Location	Crop	Yield	Yield Units
298	22a432e3	2008	Floyd County, TX	Cotton	1678	lb / acre
3422	35898fec	2013	Sedgwick County, KS	Wheat (winter)	34	bushel / acre
3918	f3a0f296	2017	Desha County, AR	Rice	185	bushel / acre
297	9275ec4f	2013	Hale County, TX	Cotton	1297	lb / acre
298	9e3a213d	2008	Floyd County, TX	Cotton	1678	lb / acre

The data file contains 839 records and 171 columns. Only relevant columns for analysis were retained.

To evaluate the effect of the detrending approaches on metrics comparisons, those fields with less than two records for the same crop were omitted from the analysis. In total, 169 fields had multiple-year data for the same crop. The following graph shows the distribution of fields per crops:

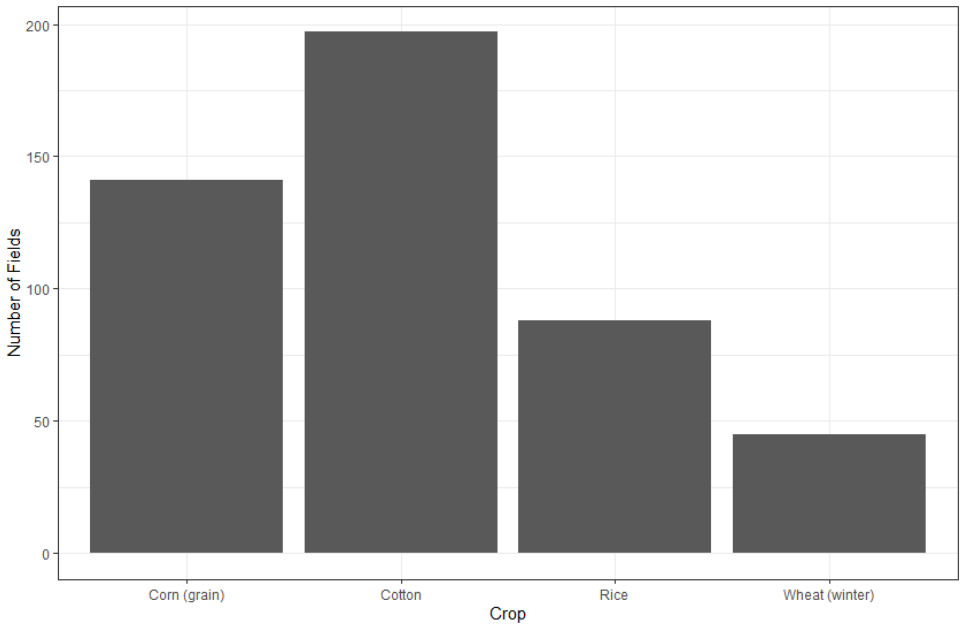

Figure 1. Distribution of crops within the subset of fields with more than two records

After initial exploration, yield units were homogenized. Then, yield-scaled metrics like Land Use, Irrigation Water Use, Energy Use, and GHG Emissions metrics were recalculated. Yield units were converted to pounds per acre for all crops.

The following graph shows yield data of the several crops by state across years. There were several extreme yield values entered into the Fieldprint Platform, likely the result of erroneous data entry. Using boxplot to identify outliers, these data points were set aside pending review by the users who entered the data. In addition, those records with a yield below 100 lb/acre were removed. Once the data inputs are revised, and the Fieldprint Analysis calculated again, the data points can be brought back into the analysis.

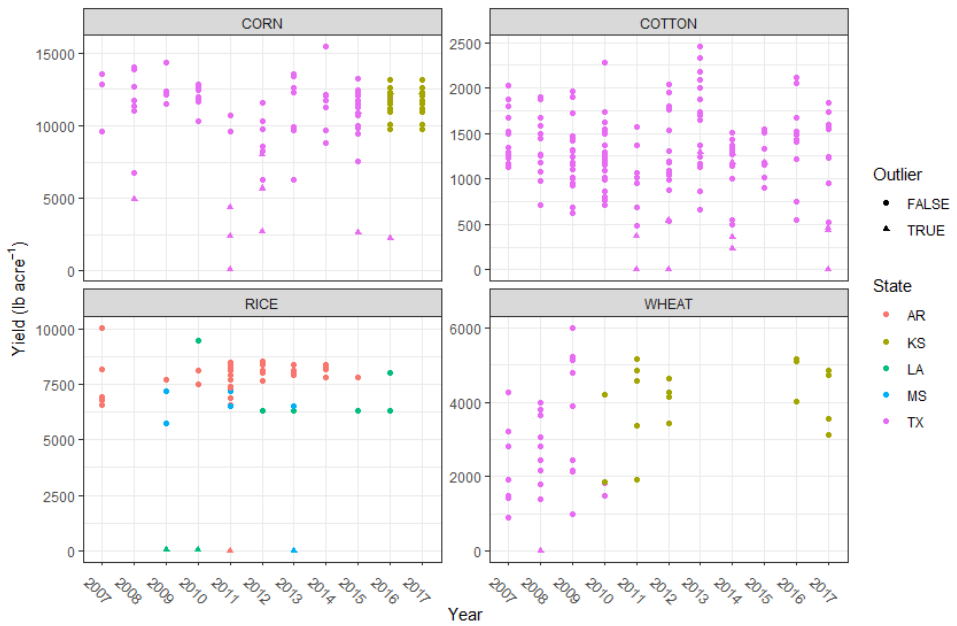

Figure 2. Homogenized yield data by crop and state across years

In total, 29 data points were classified as outliers due to extreme yields; however, it did not change the number of fields with more than two observations.

NASS Data

To assess yield trends, USDA NASS county-level yield values were downloaded via API queries. The USDA NASS data can also be manually downloaded via their online Quick Stats platform. For each crop, state, county, and year included in the Fieldprint Platform data file, variables about yield values and area planted were retained. A sample of the data extracted from USDA NASS is shown below:

Table 2. Glance of USDA NASS records

Crop	Production Practice	State	County	Year	Area Planted	Yield
CORN	NON-IRRIGATED	KS	DONIPHAN	2014	88400	11704.0
CORN	ALL PRODUCTION PRACTICES	MS	LEFLORE	2016	68300	10847.2
WHEAT	ALL PRODUCTION PRACTICES	KS	SMITH	2011	131000	2538.0
CORN	ALL PRODUCTION PRACTICES	TX	ELLIS	2012	19800	5482.4
WHEAT	NON-IRRIGATED	KS	DICKINSON	2007	159000	1080.0

In those cases where data from more than one production practice were available, only the category `ALL PRODUCTION PRACTICES` was retained, which is usually a blended value of all rainfed, irrigated, and non-irrigated production for a given county.

APPROACH 1: LINEAR TREND

The first approach was to model yield trends for each county based on county-level yield estimates obtained from NASS. This approach assumes that the trend observed on NASS data at the county level would capture the global effect of weather on yield for that particular year. This global effect would allow the adjustment of yields observed at field-level as follows:

$$Y'_{ij,t} = Y_{ij,t} - (\hat{Y}_{i,t} - \overline{Y}_{i,.})$$

where: $Y'_{i,t}$ is the adjusted yield for field j from county i in year t; $Y_{ij,t}$ the observed yield for a field j from the county i in the year t; $\hat{Y}_{i,t}$ is the predicted yield for county i in year t and $\overline{Y}_{i,.}$ is the mean yield for county i across all the years in the series.

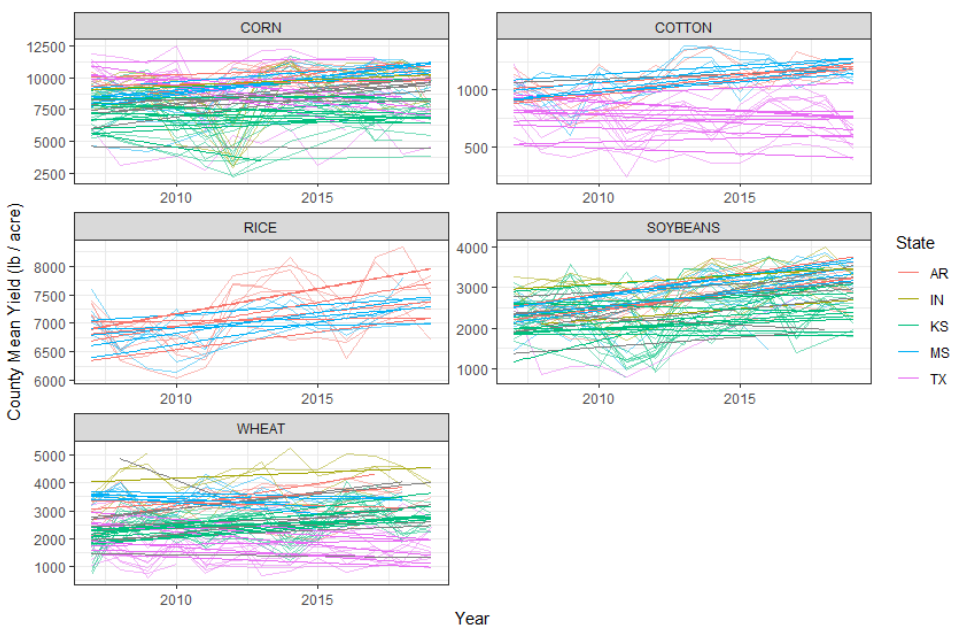

Figure 3. Evolution of yields across years and linear trend adjusted at the county level

Trends vary a great deal among counties, even within the same state. As it can be seen in the following histogram, the distribution of the R^2 statistic, which account for variability explained by the linear trends, showed a wide range of variation.

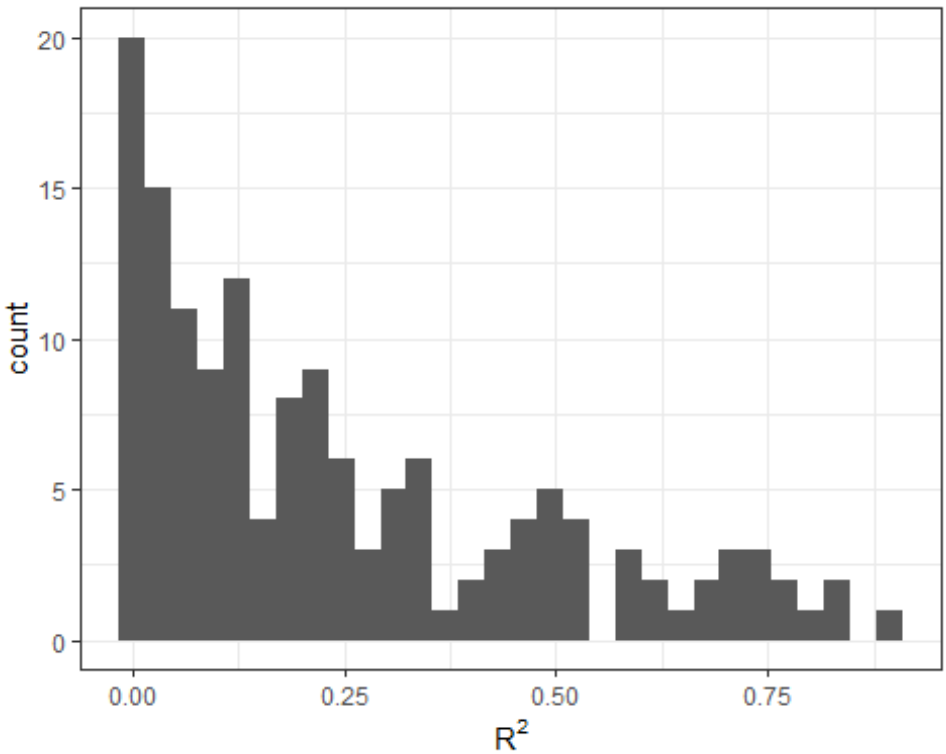

Figure 4. Distribution of R^2 of the linear trend models by county

The amount of variation explained by these models was moderate to low, showing lack of linearity in some cases. However, for some counties, these trend models seemed to be a good fit. For those cases with lack of linearity, this adjustment would be not useful.

Another factor to consider is the length of the series of yield estimates available. In order to apply this approach, yield series should be several years long, ideally more than five years. The following graph shows the distribution of the number of counties against the length of the yield estimate series. About 11% of the county data have series shorter than 4 years.

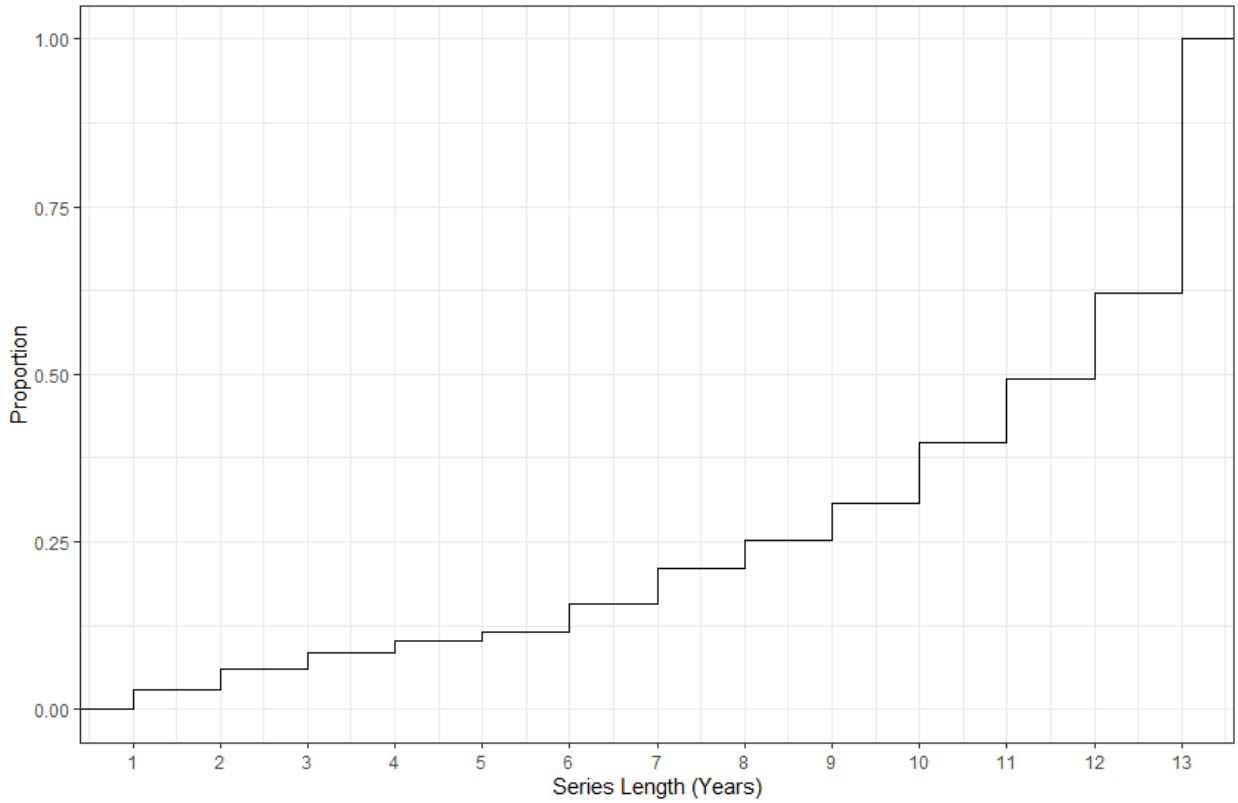

Figure 5. Distribution of series length of NASS yield data

Once the yields are corrected by the temporal trend, the yield-scaled metric scores were recomputed. The following graph shows the correlation between the indicators before and after the correction:

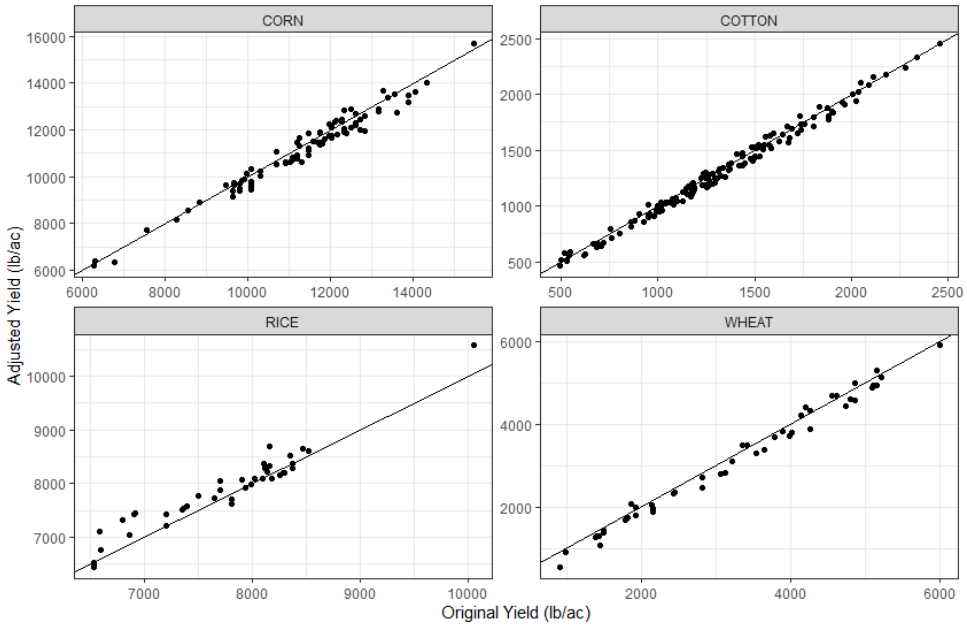

Figure 6. Comparison between the original and the adjusted yields using detrending approach 1

Except for Irrigation Water Use, the rest of the metrics showed that most of the adjusted values were close to the original ones. This result agrees with the lack of linearity mentioned above; for most cases, it fails to pick up the trend of county yields. In the case of Irrigation Water Use, the severe offset observed could be attributed to county-level estimates for Cotton including rainfed observations, while approx. 96% of cases in the dataset provided were irrigated.

APPROACH 2: NON-LINEAR TREND

Another option is to drop the linearity assumption and to fit more complex functions to capture localized trends. Several algorithms compute such functions, and one option is local polynomial regression, also called *locally estimated scatterplot smoothing* (LOESS). It consists in fitting local regressions using only neighbor observations by weighing them based on the distance to the target point.

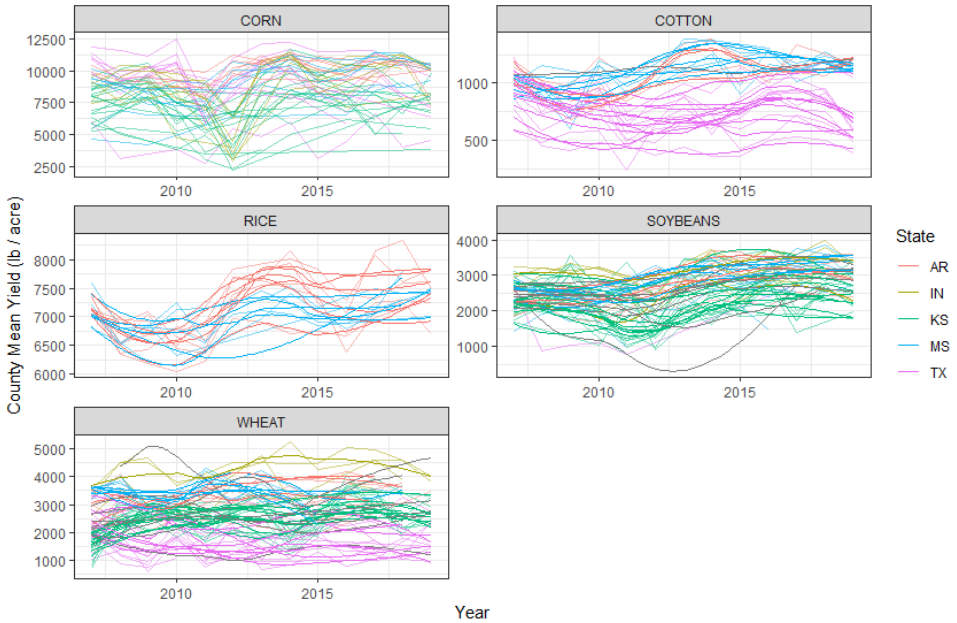

Figure 7. Evolution of yields across years and non-linear trend adjusted at the county level

This approach works better for series with enough length, which is the case for the Fieldprint dataset.

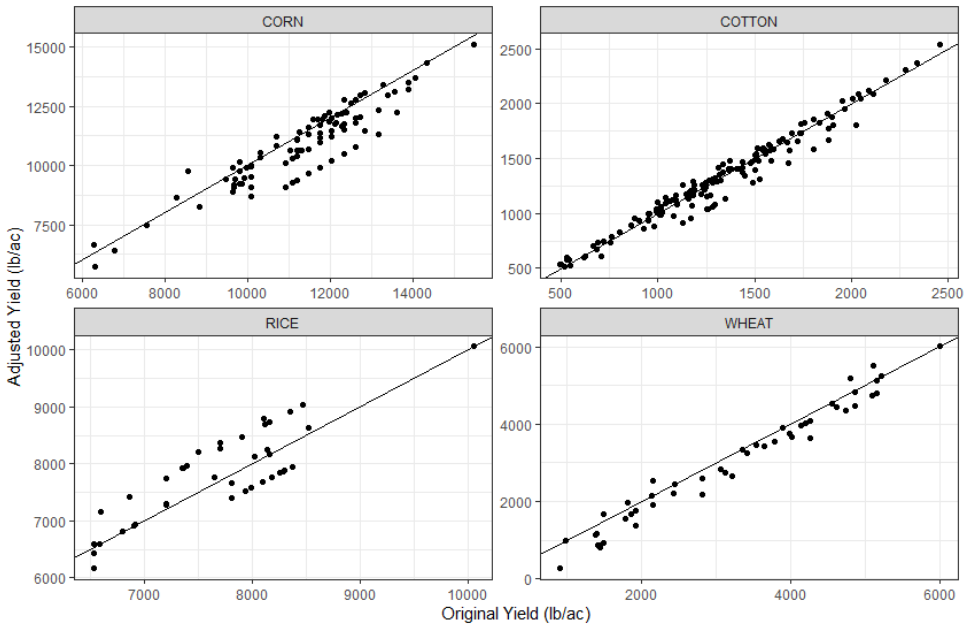

Figure 8. Comparison between the original and adjusted yields using detrending approach 2

In this case, more records were adjusted by the trend captured by the LOESS algorithm. A similar effect is observed for Irrigation Water Use in Cotton.

APPROACH 3: SUBTRACTING COUNTY YIELDS

A more flexible approach would be to adjust yields by calculating the difference between the individual field yields for each particular year and the overall county yield across years. Thus, the adjusted yield would be obtained as follows:

$$Y'_{ij,t} = Y_{ij,t} - (\overline{Y}_{i,t} - \overline{Y}_{i,\cdot})$$

where: $Y'_{i,t}$ is the adjusted yield for field j from county i in year t; $Y_{ij,t}$ the observed yield for a field j from county i in the year t; $\overline{Y}_{i,t}$ is the average yield for county i on year t and $\overline{Y}_{i,\cdot}$ is the mean yield for county i across the years in the series.

Although this method is more flexible as it does not involve any particular trend, the drawback is that it could be copying the general pattern and the noise in the data.

Once the yields are corrected by the temporal trend, we can recompute the yield-scaled metric scores.

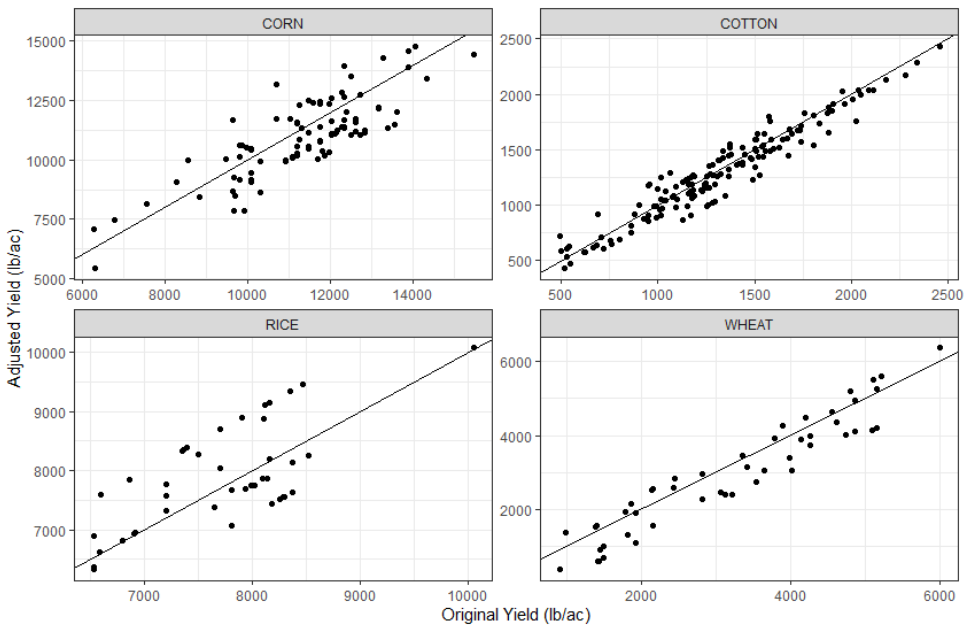

Figure 9. Comparison between the original and the adjusted yields using detrending approach 3

As expected, this correction method captures more variability. This outcome is clearly observed for Land Use, which is a direct quotient between yield and the area of the field. However, this approach cannot be partitioned into trends across years or any other source of variation.

APPROACH 4: YIELD INDEXING

Lastly, an approach similar to the previous method is to use the county yields to create a yield index, relating county yields to a particular year of the series, for example, the last year

available, like deflating prices series. In the previous approach, the reference yield was the overall mean yield. These indices are used to adjust the actual yields to recompute the yield-scaled metric scores using the deflation approach.

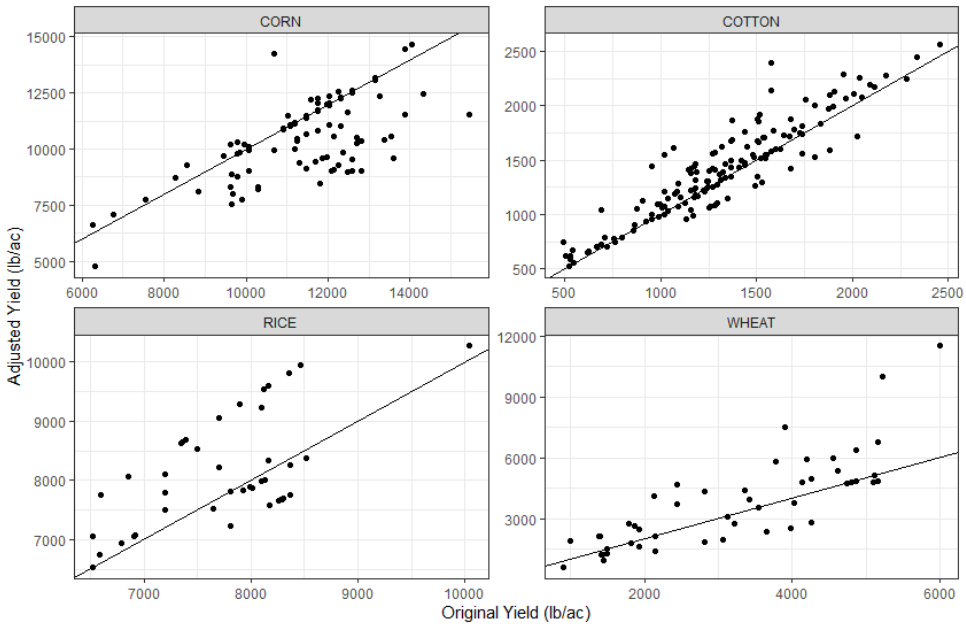

Figure 10. Comparison between the original and the adjusted yields using detrending approach 4

COMPARING RESULTS

To assess the effect of these adjustments, the change Δ of each metric Y between two successive years (t_1 and t_2), which can be consecutive or not, were computed as follows:

$$\Delta_{t_2-t_1} = \frac{\Delta Y}{\Delta t} = \frac{Y_{t_2} - Y_{t_1}}{t_2 - t_1}$$

where: ΔY is the change in a metric between two successive years and Δt is year difference.

The following graph shows the median difference or change in the metrics crop when the data for the same fields and crops are compared using metric scores without yield trend correction or with the four approaches detailed above. The median is a summary statistic that shows the value which divides the sample of fields in two halves based on the value of the Δ, so for a given median, 50% of fields would have a change of the metric above or below this value.

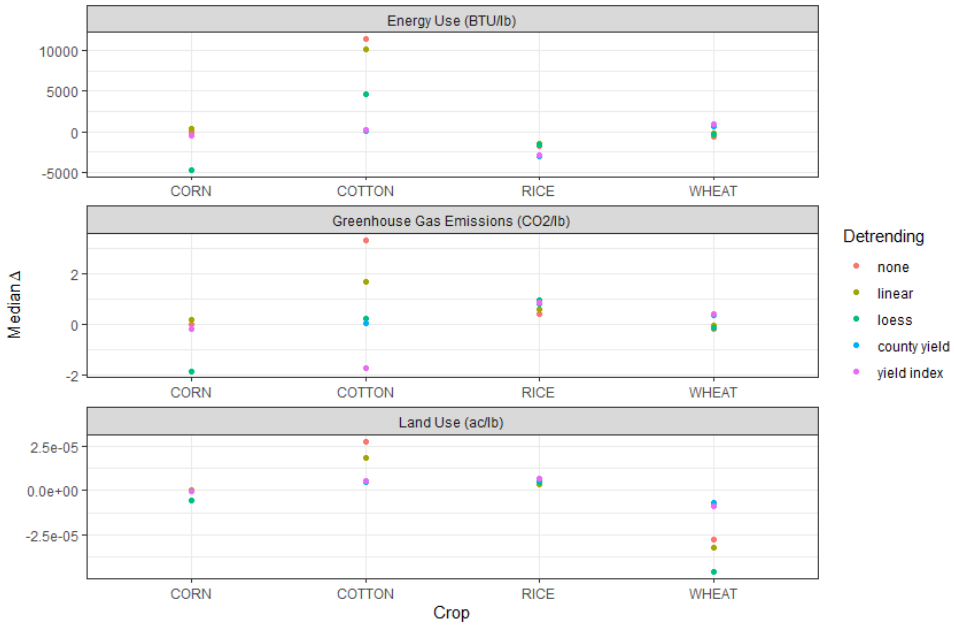

Figure 11. Effect of the different yield detrending approaches on the inter-annual variation of the metrics for each crop

It can be seen from the figure above that, depending on the yield correction method, the median change on some metrics varies substantially. For example, for Cotton, the median *Δ* was nearly 0 for the county yield or yield index methods, indicating that, the net change on this metric was null for most of the fields after adjusting the yields by the year effect. In contrast, the metrics computed on the original yields or those adjusted by the trend captured by the linear model or the LOESS algorithm indicated an increase on this metric.

Finally, the following plots show the evolution of the changes of the studied metrics across time for the different crops using the original and detrended data. In these plots, positive values indicate that the metric calculated in a particular year is higher than the year before for the same field and crop. In contrast, negative values mean that the metric for that particular year was lower than the one computed in the previous year for the same field/crop.

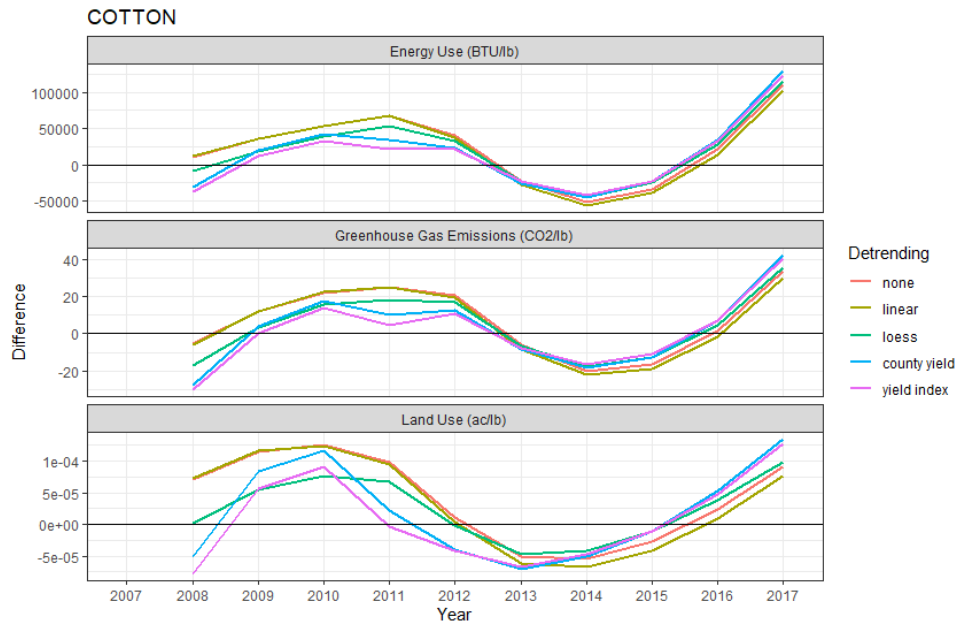

Figure 12. Evolution of the changes on metrics using raw and adjusted yield data for Cotton

For the case of Cotton, the evolution of the changes calculated using different approaches were closely to those observed using raw yields. The Land Use Metric, which is more sensible to changes in the yield estimate, reflected larger differences between approaches. For example, in years 2008-2011 the unadjusted yields or yields adjusted by Approach 1, resulted in changes with more positive values.

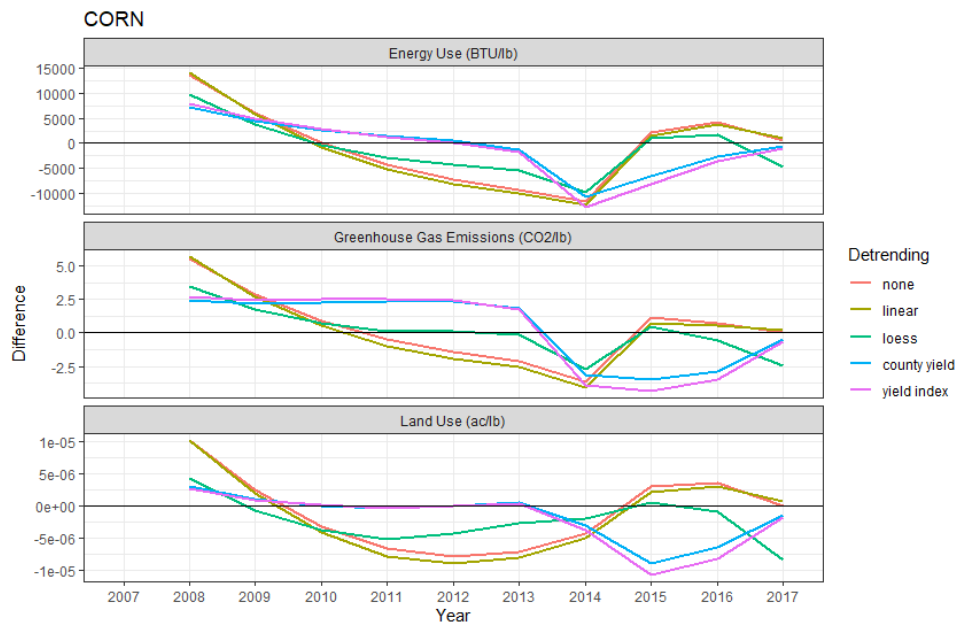

Figure 13. Evolution of the changes on metrics using raw and adjusted yield data for corn

For corn, the pattern shows that between the years 2008-2013, metrics based on unadjusted or poorly detrended corn yields tended to show a change from positive to negative values. In contrast, the county yield and yield index methods resulted in smaller values across time. Land Use showed a similar pattern but with values around 0, denoting no changes across years. After 2014, metric scores using adjusted yields from the county yield and yield index methods consistently resulted in lower scores than the other methods.

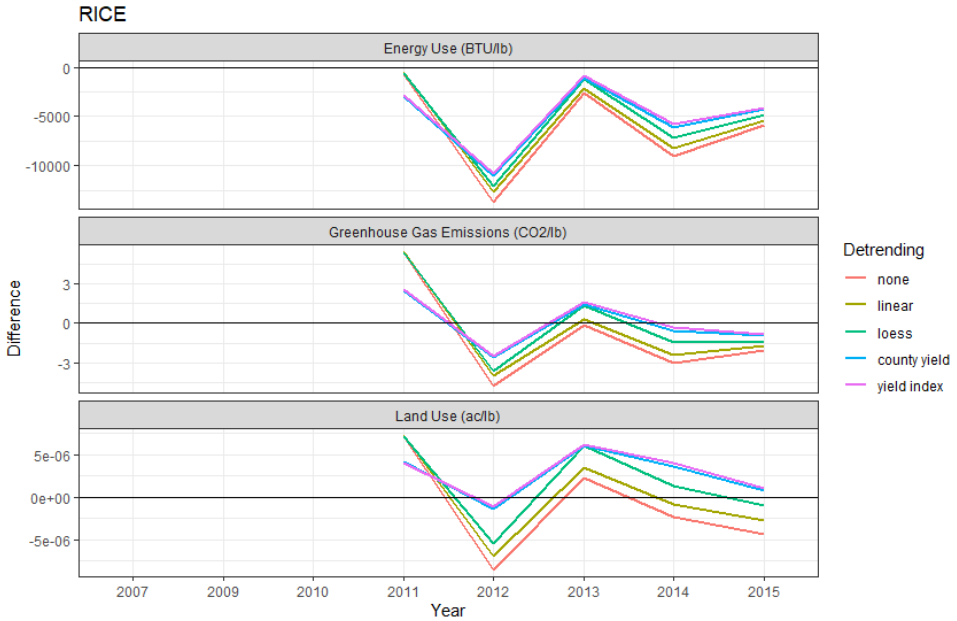

Figure 14. Evolution of the changes on metrics using raw and adjusted yield data for rice

In the case of rice, metrics from unadjusted or poorly detrended yields resulted in lower scores across the years analyzed. In some cases, changes in the directional effect are observed. For example, in the year 2014, which compares 2013 vs. 2014, original metrics or adjusted by approach 1, resulted in a negative value for Land Use, but the opposite effect was observed for detrended yields using approaches 3 and 4.

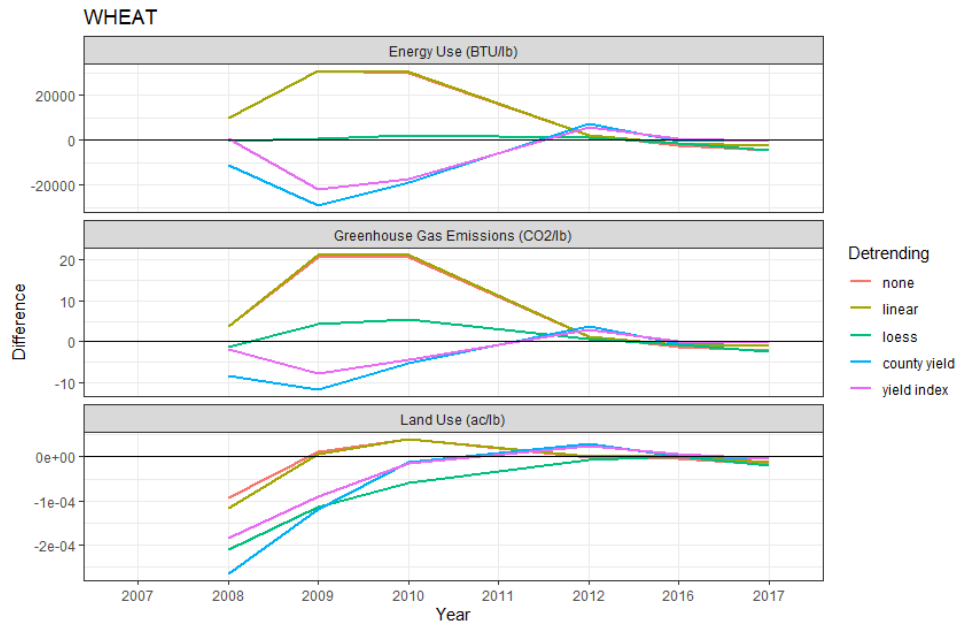

Figure 15. Evolution of the changes on metrics using raw and adjusted yield data for wheat

For wheat, discrepancies are observed for the years 2008 through 2012 for the detrending methods. Changes in the directional effect are observed for Energy Use and Greenhouse Gas Emissions Metrics. Such an outcome would warrant further investigation of the NASS county yield trends, and the yields entered in the Fieldprint Platform to find sources of variabilities that may explain the discrepancies.

CONCLUSIONS

- The detrending approaches based on the county mean yield or the county yield of a reference year are straightforward to apply and captured more variability than approaches 1 and 2 based on linear and non-linear (loess) trend models.
- Yield-scaled metrics need to account for yield variability in order the make fair comparisons between years with different weather conditions.
- This analysis was based on a subset of longitudinal records (same field/crop across years). Adding more longitudinal data to the analysis may result in better detrending outcomes.
- Cotton yield data showed more variability than other crops, and the statistical models performed poorly; it is likely that localized conditions are influencing yield, and explanatory variables not included in the current dataset could improve model performance.