UN Crop Data

Joshua Winchester

2022-10-24

Our Purpose

Our goal is to examine all available UN crop data between 1970 and 2020 for the four major crops: Potatoes, Rice, Wheat, and Maize. We are going to attempt to answer three questions: 1) What can we infer about the nature of the production of these 4 crops in the past 50 years, 2) did COVID-19 impact agricultural output in any way?, and 3) how well can we predict the yield, production and area harvested for each year?

The source for this data can be found here: http://data.un.org/Explorer.aspx

The Data

```
str(orgcropdata)
```

```
## 'data.frame':    26308 obs. of  6 variables:
##  $ Crop          : chr  "Potato" "Potato" "Potato" "Potato" ...
##  $ Year          : num  2018 2018 2018 2018 2018 ...
##  $ Country       : chr  "Afghanistan" "Albania" "Algeria" "Angola" ...
##  $ Production    : num  615684 254543 4653322 458217 2403193 ...
##  $ Yield         : num  190026 261714 310916 71770 318836 ...
##  $ Area.Harvested: num  32400 9726 149665 63845 75374 ...
```

This data includes UN FAO (Food and Agricultural Organization) data from available nations between 1970 and 2020. Not all countries are represented (for various reasons: they don't submit to the UN, political situations, etc), and there are also additional difficulties because some countries change names, become parts of others, etc, such as Sudan. Nonetheless, most nations have had the same borders in the past 50 years, and also we will not be examining this on a national level. However, the vast majority of nations are represented, and especially all the largest ones agriculturally speaking.

After cleaning, munging, and some simple subsetting, we also had to reorganize the data because it was not in a friendly format. After this, we are left with 6 variables: Crop, Year, Country, Production, Yield, and Area Harvested. Here is a brief explanation of each of them:

Crop: Which of the 4 crops is this? Potatoes, Maize, Rice, or Wheat?

Year: Which year is it, between 1970-2020?

Country: Which country is this data from?

Production: Total crop production, in tonnes, for that given year.

Yield: Production per Area Harvested, in hg/ha, for that given year.

Area Harvested: Total area harvested from, in ha, for that given year.

It's important to note that not each year/nation/crop combination will have non na or 0 values, but the vast majority are positive non zero values.

Our Process

In order to accomplish this goal, we are going to use K-nearest neighbors to create predicted values for each given year. We will remove the 2019 and 2020 data to use for comparison. We will create 12 separate plots, 1 for each crop and output statistic. We will compare each plot briefly, with some commentary of what the plots mean.

Reading the plots

Green Triangles : Represents an individual data point from a nation and year

Red X : Represents a knn predicted value for that year

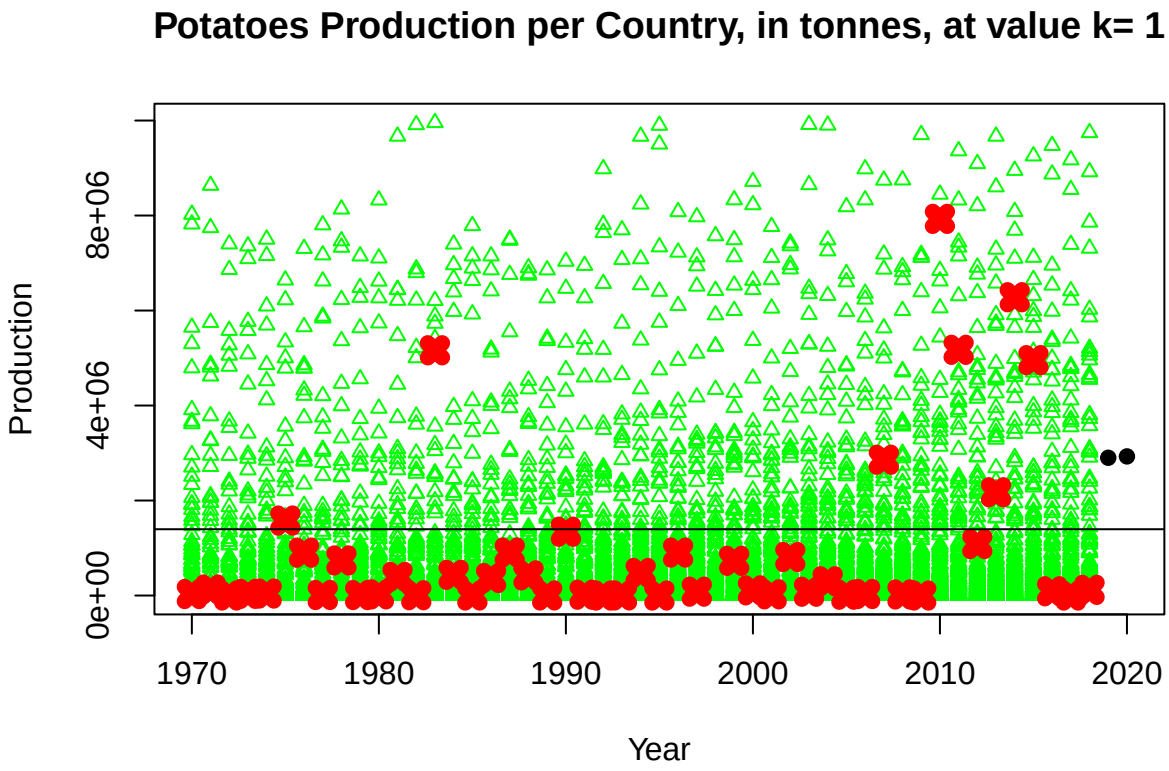
Black Bar : Represents the mean knn predicted value between 1970-2018

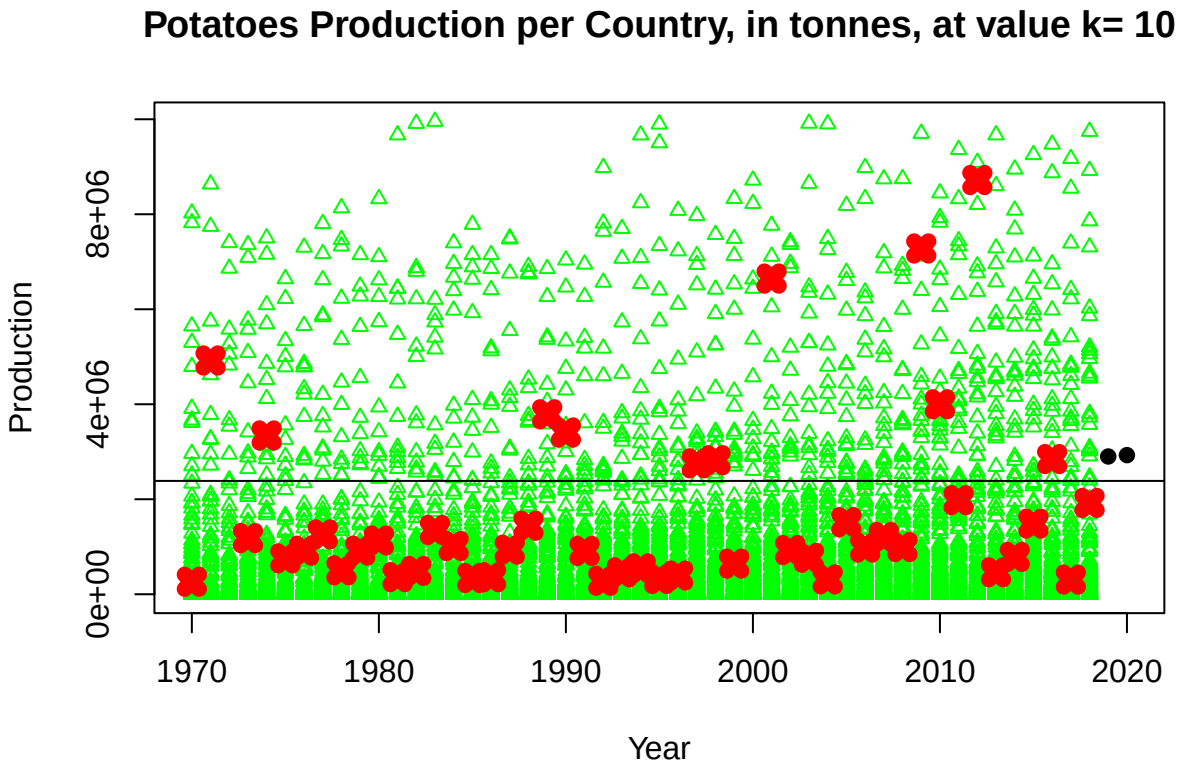
Black Dots: Represents 2019 and 2020 mean data

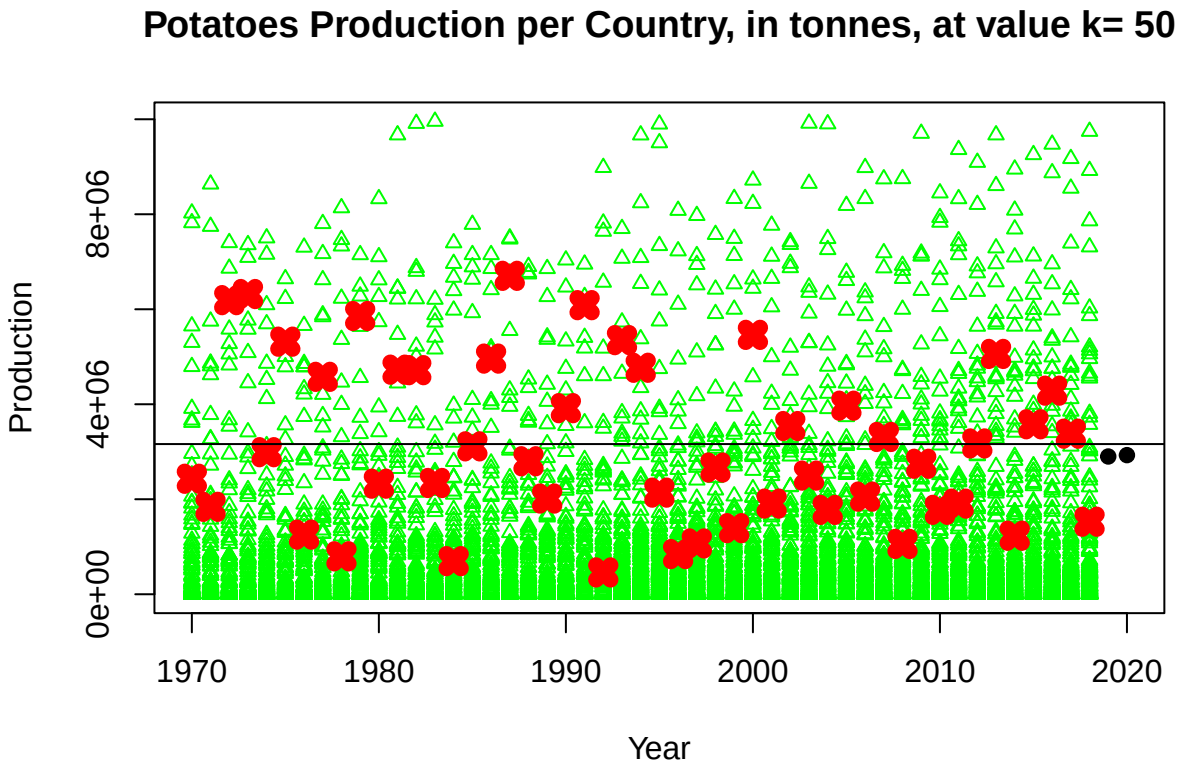
It's important to know too that Production and Area Harvested statistics have outliers removed, mostly because the outliers made the graphs hard to read.

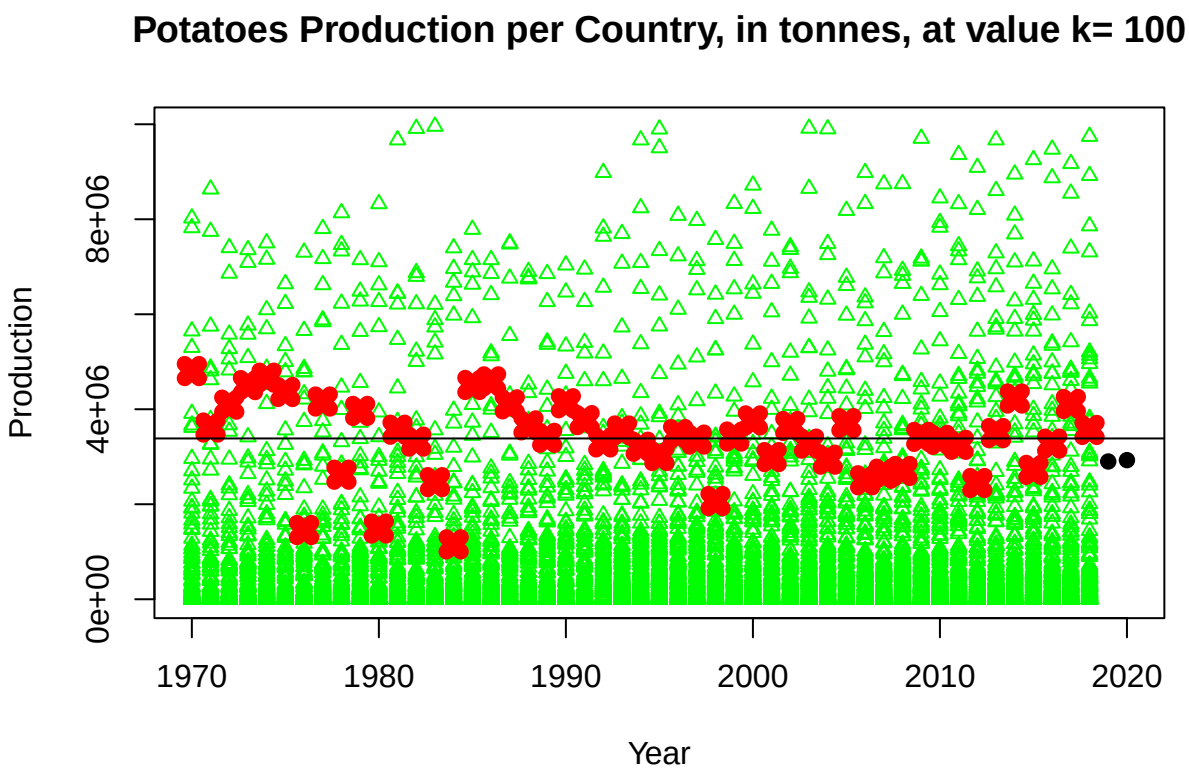
Testing Different Values of K

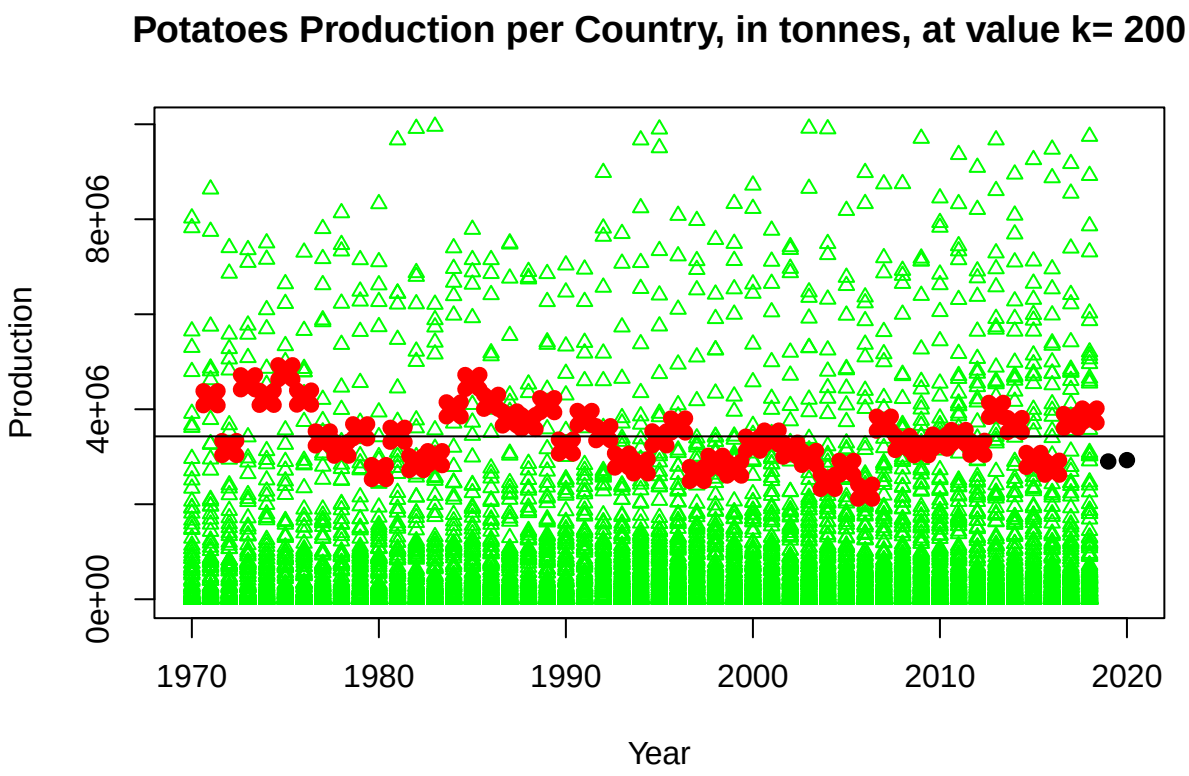
What follows is the output of multiple different values of k. The value tested is labelled at the top; you can quickly scroll through them to see what the output looks like. We are going to test this on Potato Production only, and see if we can find an approximately good k value.

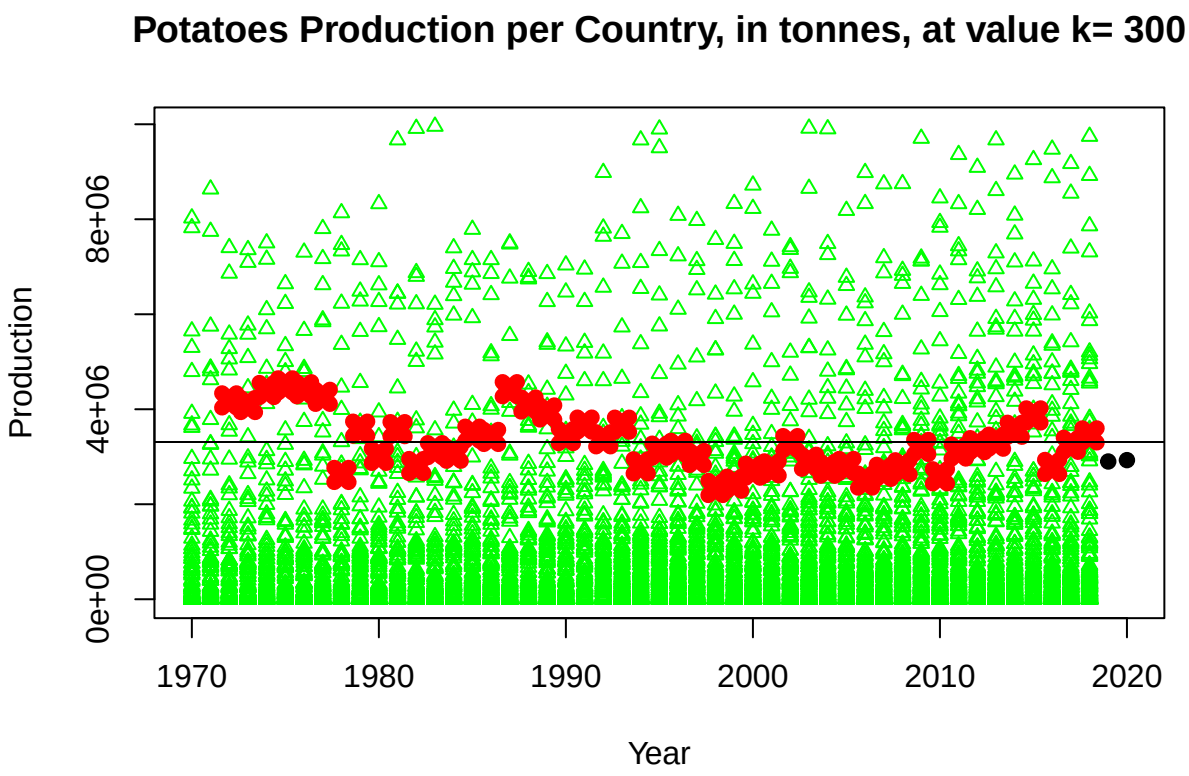


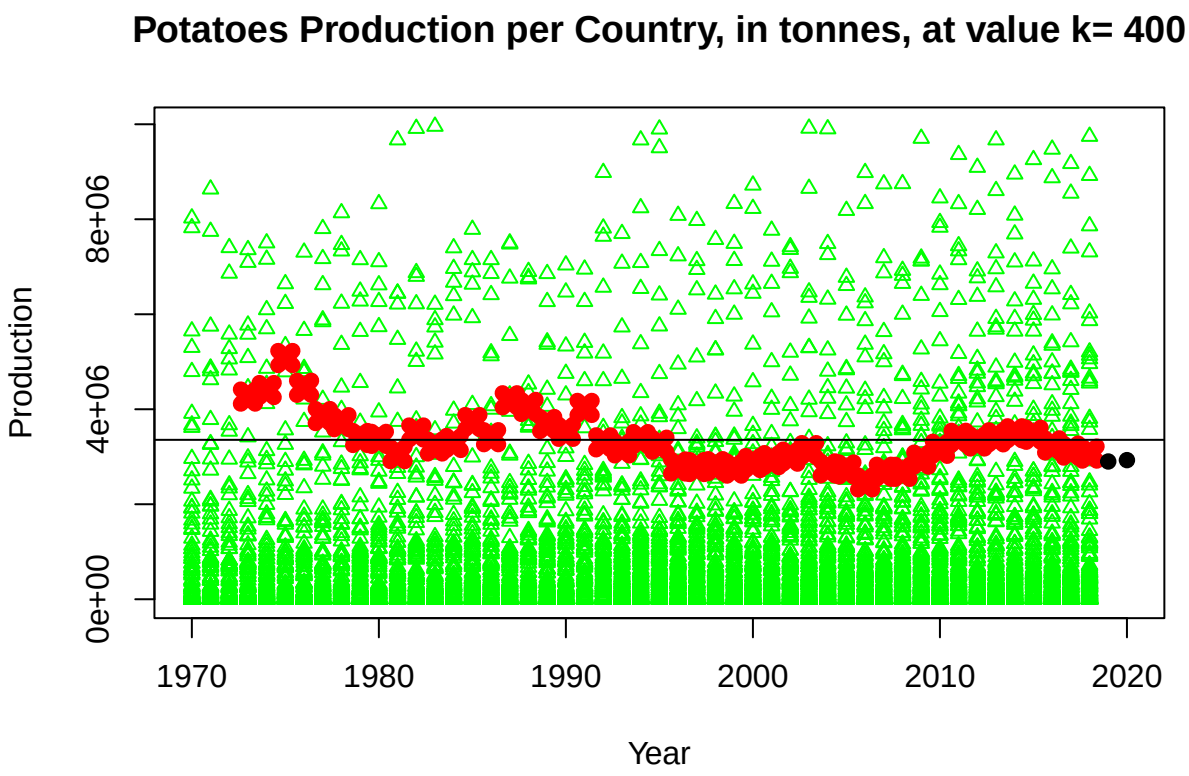

Potatoes Production per Country, in tonnes, at value k= 10
Production
0e+00
4e+06
8e+06
1970
1980
1990
2000
2010
2020
Year

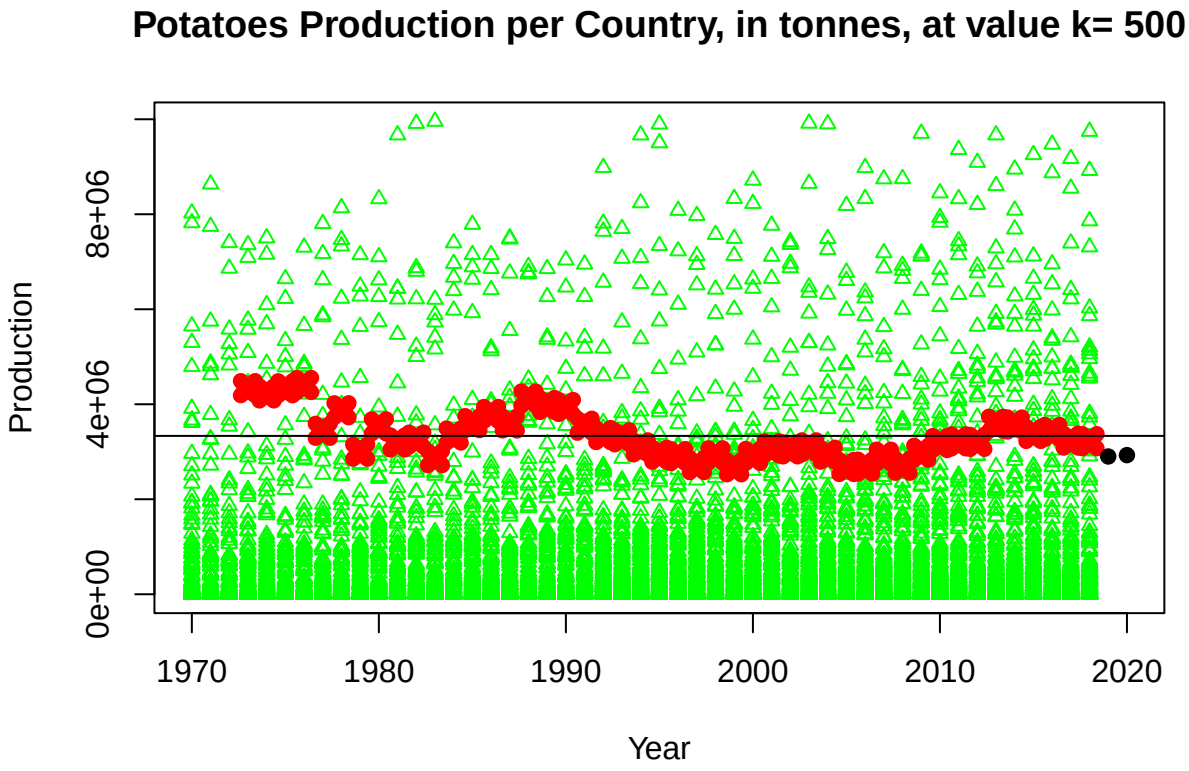

Potatoes Production per Country, in tonnes, at value k= 50
Production
0e+00
4e+06
8e+06
1970
1980
1990
2000
2010
2020
Year

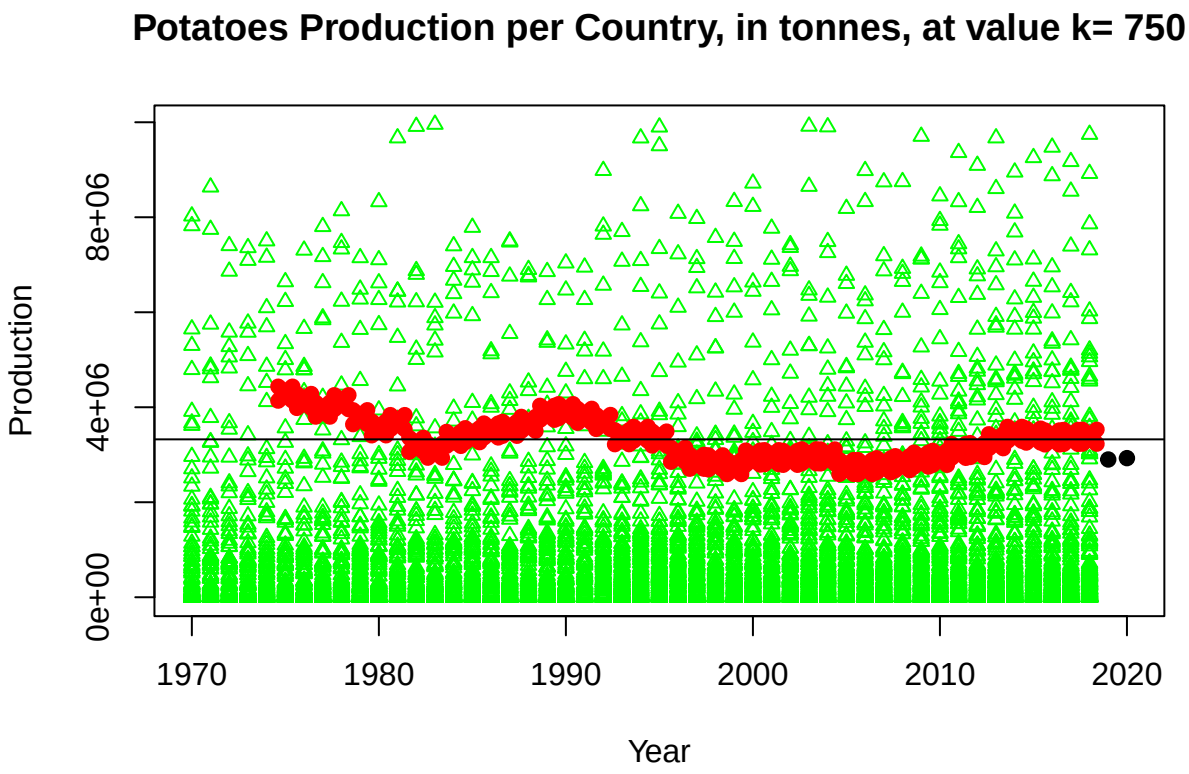

Potatoes Production per Country, in tonnes, at value k= 100
Production
0e+00
4e+06
8e+06
1970
1980
1990
2000
2010
2020
Year

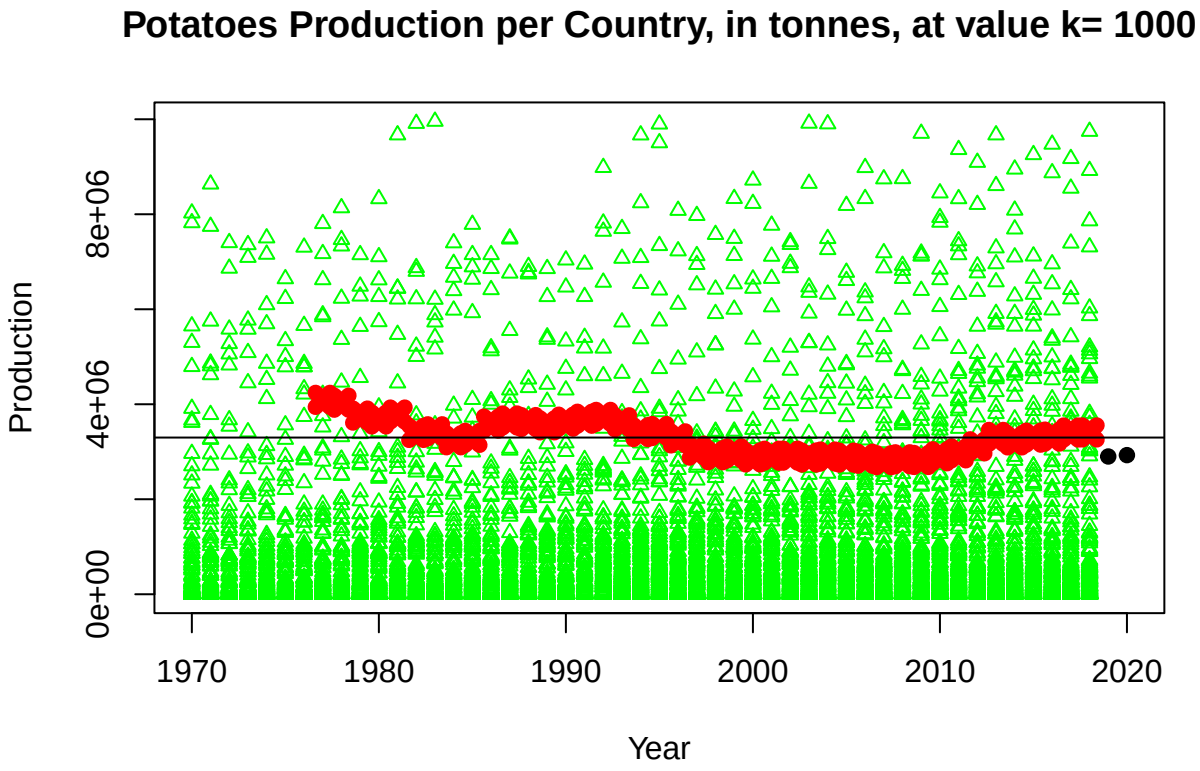

Potatoes Production per Country, in tonnes, at value k= 200
Production
0e+00
4e+06
8e+06
1970
1980
1990
2000
2010
2020
Year

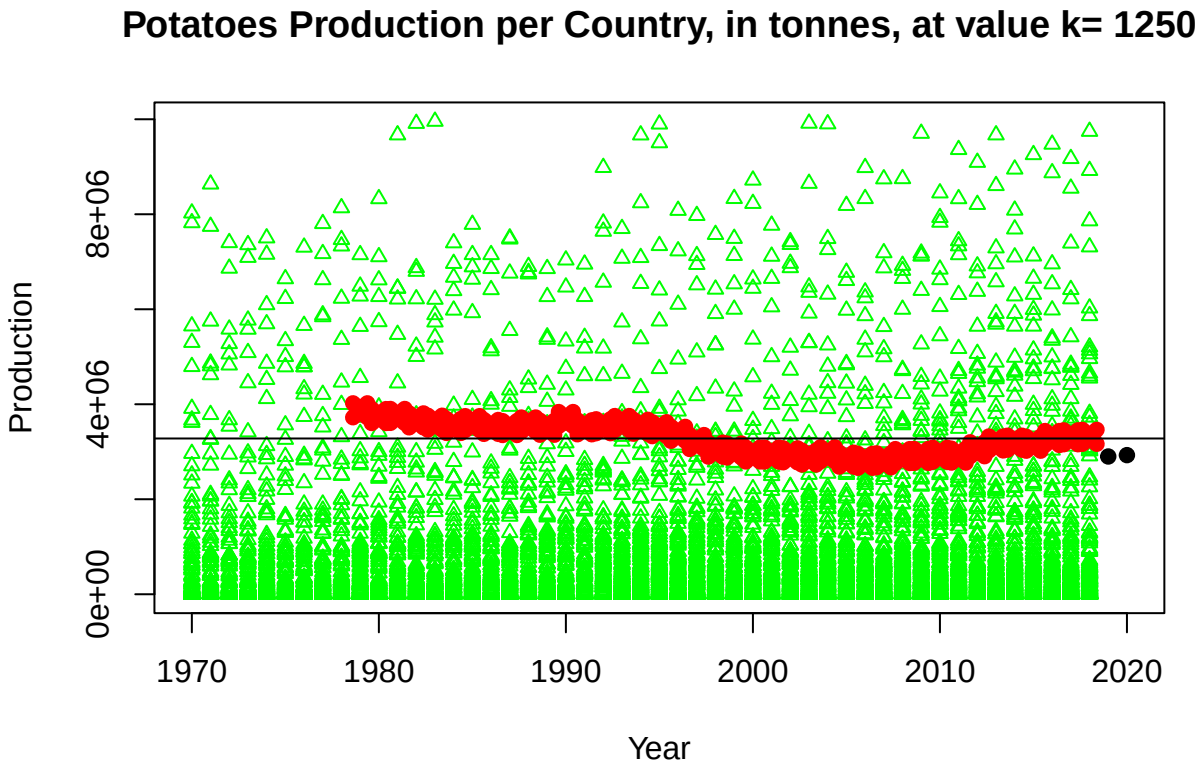

Potatoes Production per Country, in tonnes, at value k= 300
Production
0e+00
4e+06
8e+06
1970
1980
1990
2000
2010
2020
Year

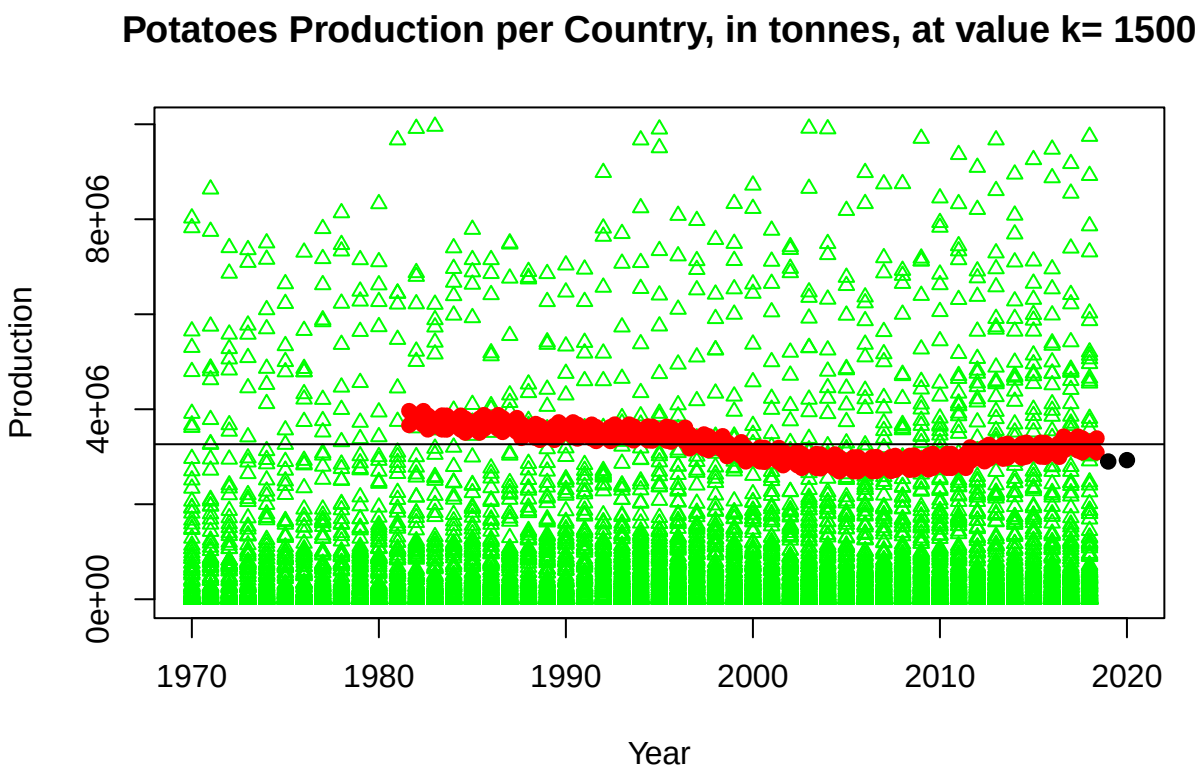

Potatoes Production per Country, in tonnes, at value k= 400
Production
0e+00
4e+06
8e+06
1970
1980
1990
2000
2010
2020
Year

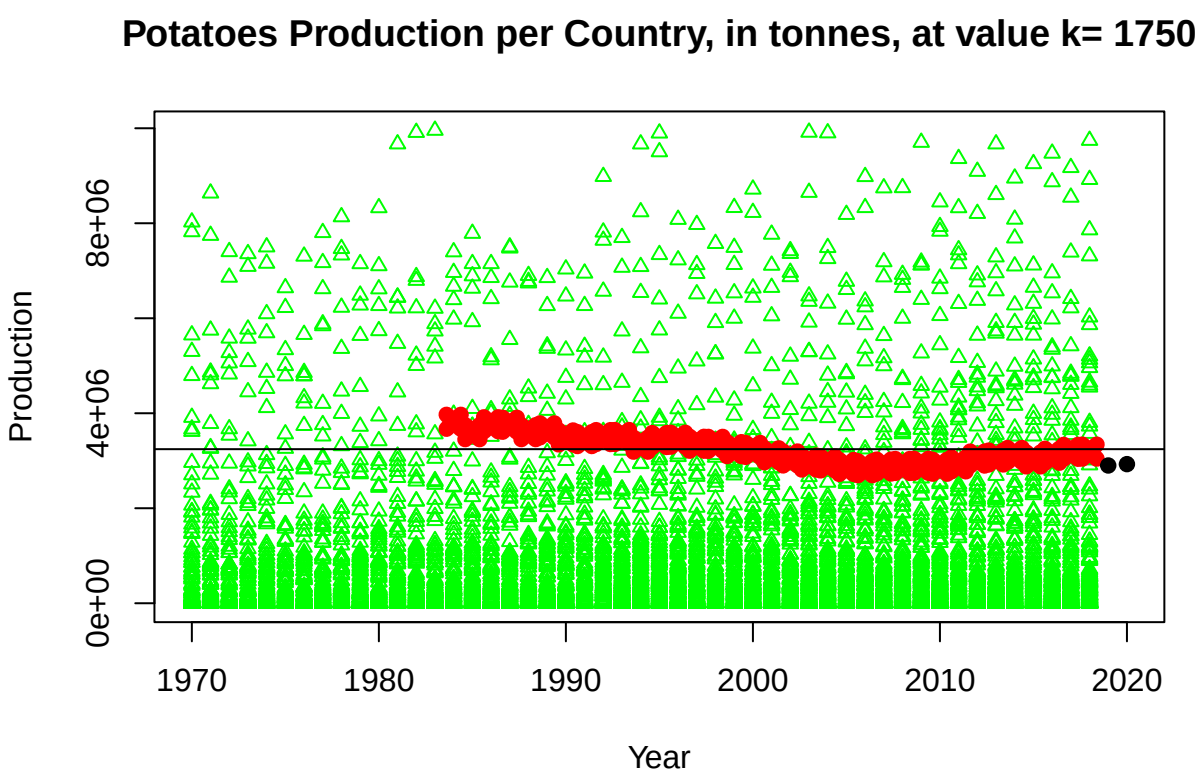

Potatoes Production per Country, in tonnes, at value k= 500
Production
0e+00
4e+06
8e+06
1970
1980
1990
2000
2010
2020
Year

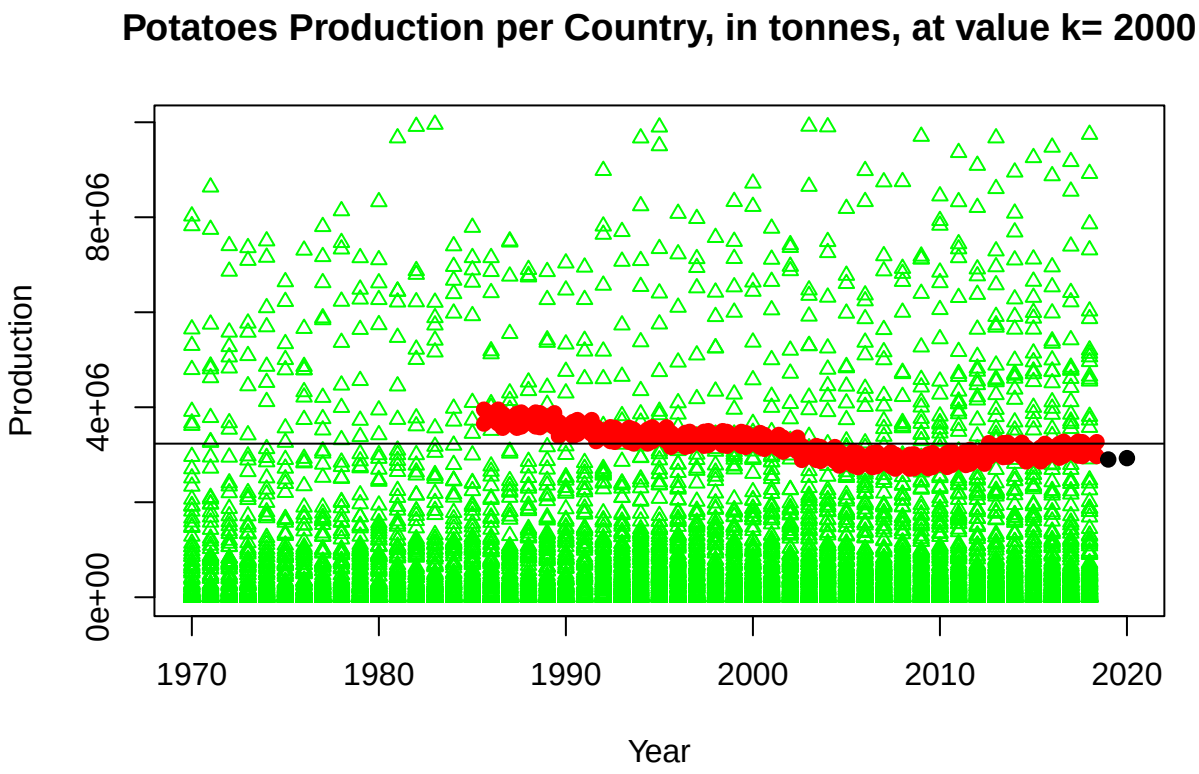

Potatoes Production per Country, in tonnes, at value k= 750
Production
0e+00
4e+06
8e+06
1970
1980
1990
2000
2010
2020
Year


Potatoes Production per Country, in tonnes, at value k= 1000
Production
0e+00
4e+06
8e+06
1970
1980
1990
2000
2010
2020
Year


Potatoes Production per Country, in tonnes, at value k= 1250
Production
0e+00
4e+06
8e+06
1970
1980
1990
2000
2010
2020
Year


Potatoes Production per Country, in tonnes, at value k= 1500
Production
0e+00
4e+06
8e+06
1970
1980
1990
2000
2010
2020
Year


Potatoes Production per Country, in tonnes, at value k= 1750
Production
0e+00
4e+06
8e+06
1970
1980
1990
2000
2010
2020
Year



Summary

What we saw was that, when our k=1, the results were chaotic and don't tell us much about the data. This is because our error is high, as a result of bias variance tradeoff: in this case, our variance is way too high and complex, as it is matching the data perfectly. If we scale up a bit, we find that about 300-500 there is an approximate sweet spot, and this makes sense: on average each of the crop/production statistic is between 6-8k total variables, and just as a quick approximation, if we divide that by 50 we get about 120-160, which is about how many data points we can expect for each year. A k value of 300-500 gives us about 2-4 years worth of data averaged around each point, roughly. This means that bad and good years (production wise) would be offset by previous stability slightly. As k approaches 2000, we see clear bias trending towards the mean. It seems that 300-500 gives us the most flexibility without being too chaotic in its results. From preliminary testing, I've found that the default guess of sqrt(nrows(dataset)) is good for yield here, but not good for production or area harvested. Going forward, we will use 400 for the production and area harvested, and sqrt(nrows(data)) for the yield.

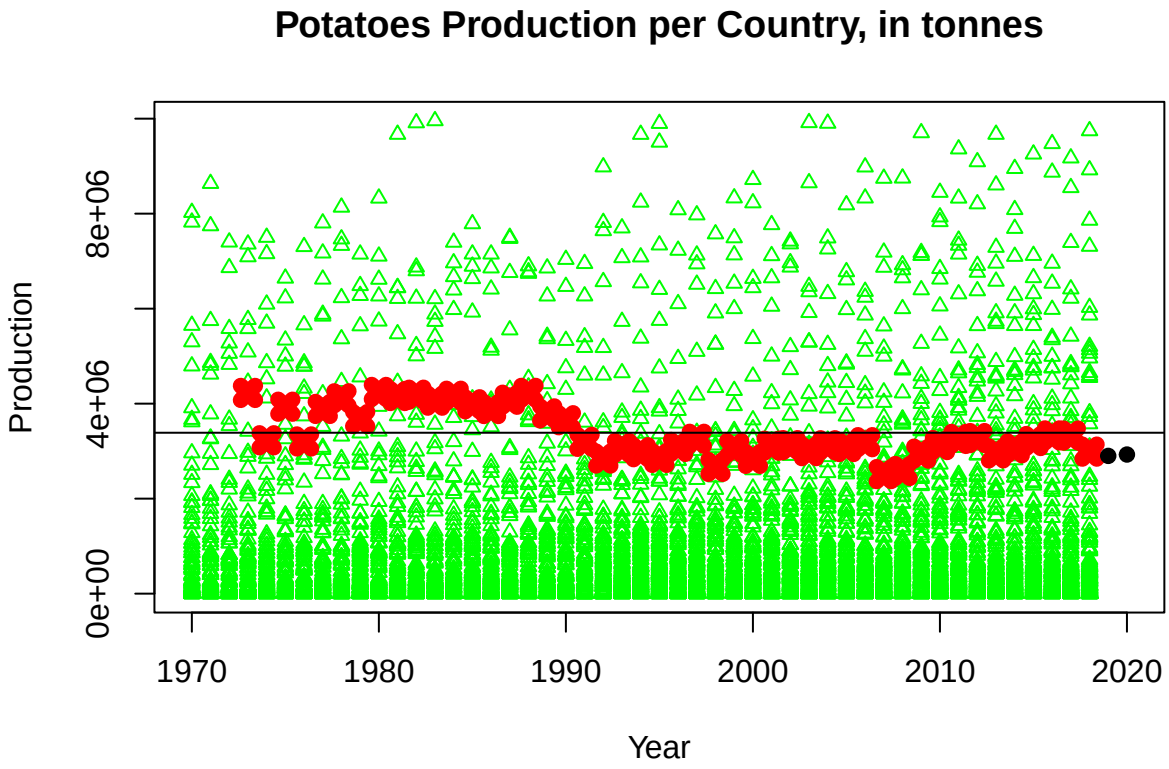
The Plots

Potatoes

Production

```
## [1] "10 K-fold CV results:"
```

```
##   intercept     RMSE    Rsquared      MAE    RMSESD RsquaredSD     MAESD
## 1      TRUE 13.99363 0.001017824 12.05012 0.2173661  0.0013107 0.2040866
```

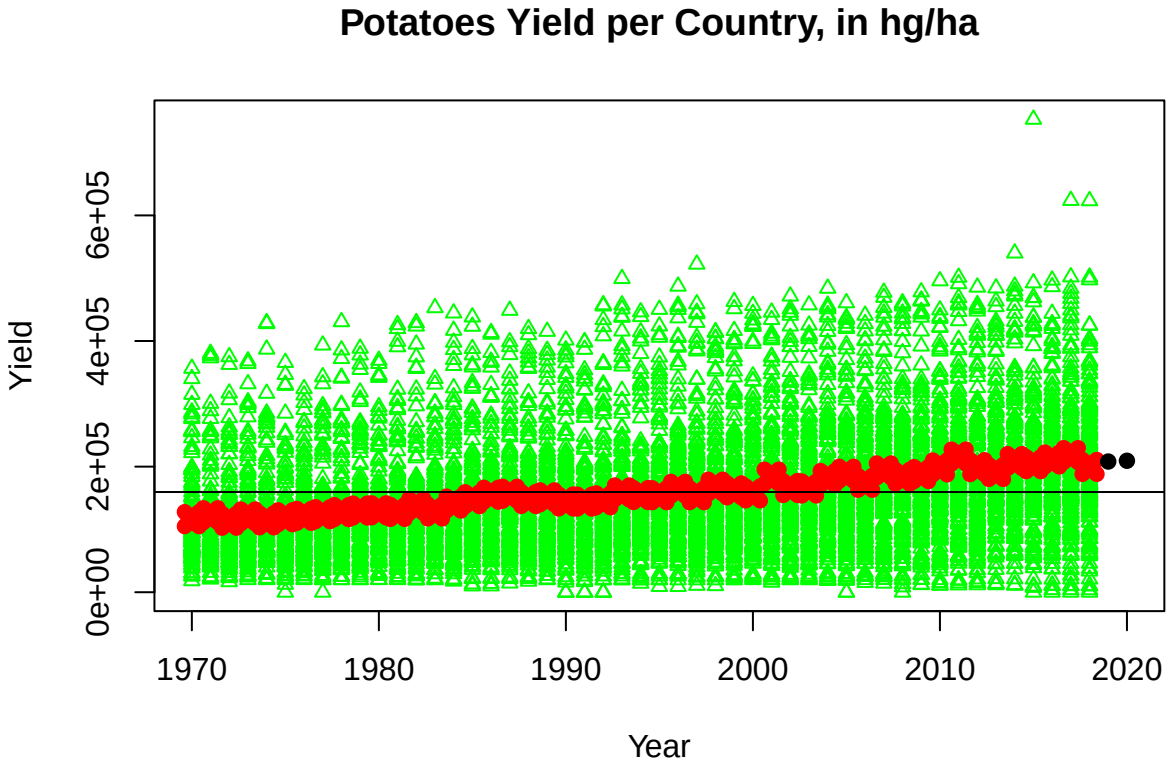


Keep note of the ~14 RMSE. We will see if others are similar. We see potatoes have leveled off in production, with a seeming peak in the mid 70's. 2019/2020 data is below the overall predicted mean, but thats not unexpected, as potato production seems to be going down slightly or staying flat as time goes on.

Yield

```
## [1] "10 K-fold CV results:"
```

```
##   intercept     RMSE   Rsquared      MAE    RMSESD RsquaredSD     MAESD
## 1      TRUE 13.44507 0.07864811 11.46851 0.3098805 0.02256711 0.2858214
```

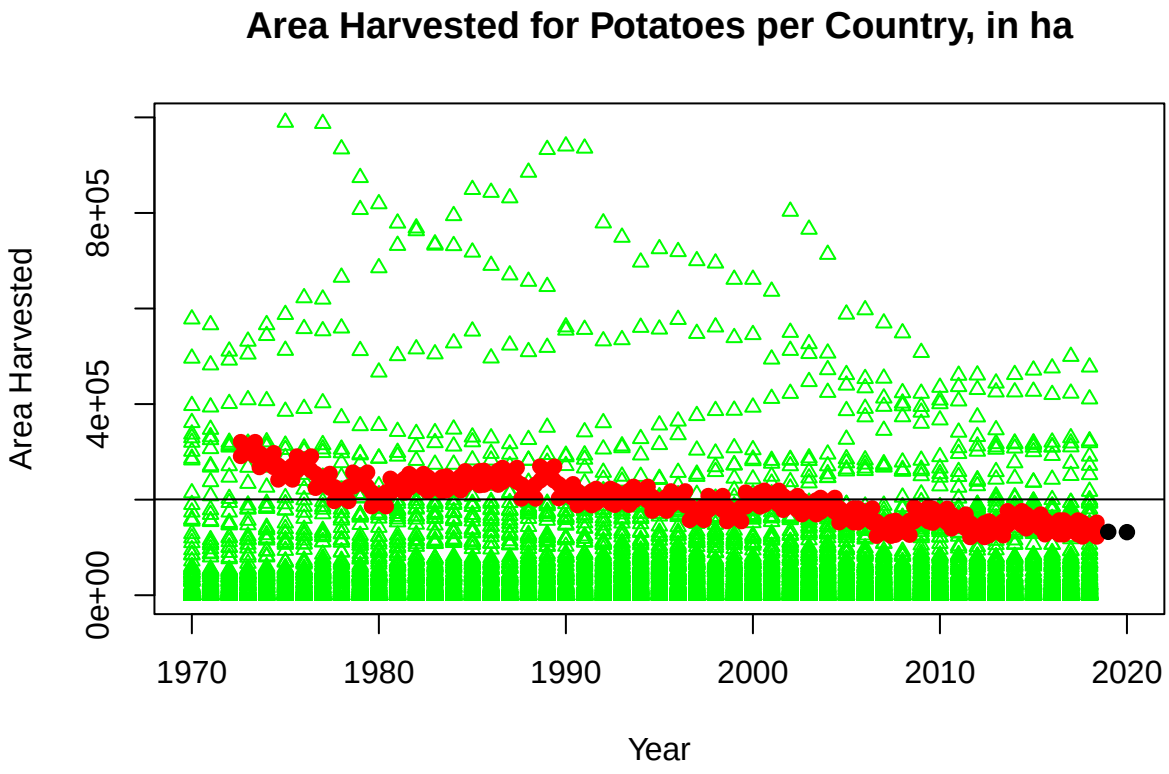


We see a steady rise in the predicted values of yield with respect to the mean, suggesting that as time goes on, its expected to see higher yields. We also see that 2019/2020 are right where would expect them to be. In general, we should expect yields to rise for potatoes.

Area Harvested

```
## [1] "10 K-fold CV results:"
```

```
##   intercept     RMSE    Rsquared      MAE    RMSESD  RsquaredSD     MAESD
## 1      TRUE 13.98486 0.002256207 12.03906 0.1158893 0.002262358 0.1368461
```



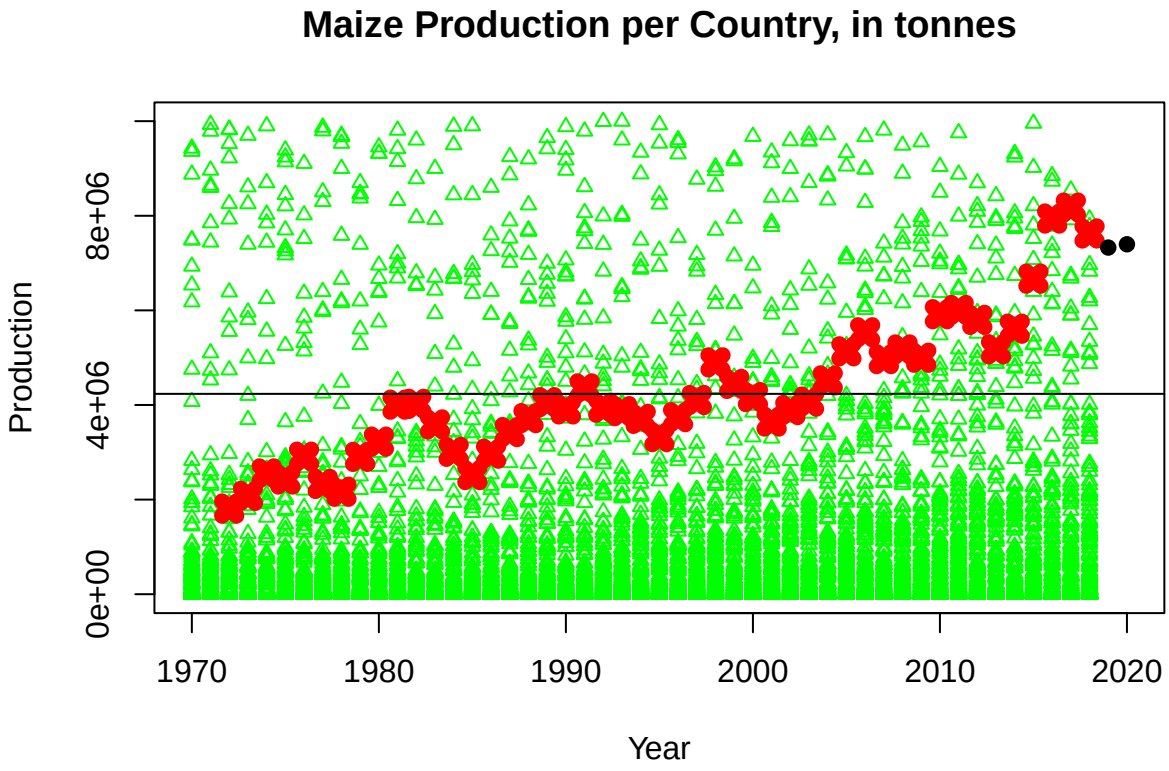
Area harvested is clearly going down for potatoes as time goes on. So yields are up, production is flat or slightly falling, but area harvested is going down, with 2019/2020 confirming this.

Maize

Production

```
## [1] "10 K-fold CV results:"
```

```
##   intercept     RMSE    Rsquared      MAE    RMSESD  RsquaredSD     MAESD
## 1      TRUE 14.07546 0.003688337 12.13595 0.1101725 0.003135334 0.1181294
```

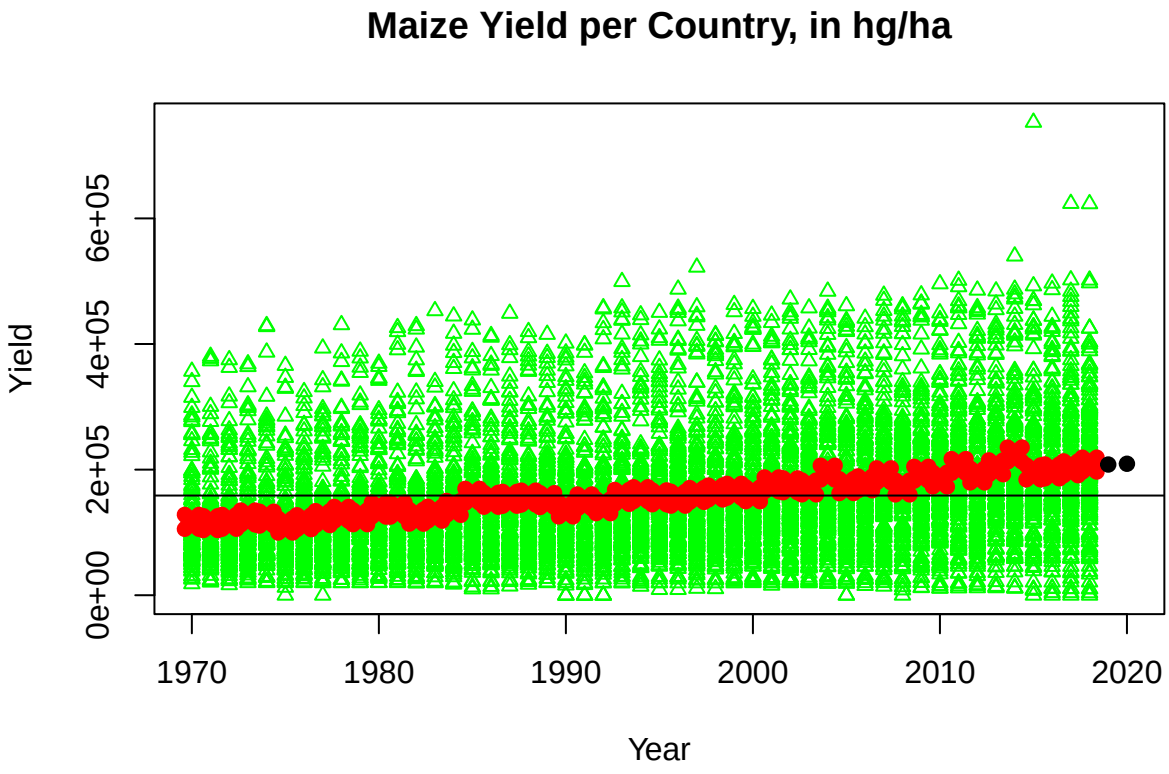


The k predictions clearly show a seeming overall rise as time goes on. 2019/2020 especially are fairly high, showing that average maize production in general has continued to rise in the past 50 years.

Yield

```
## [1] "10 K-fold CV results:"
```

```
##   intercept     RMSE   Rsquared      MAE    RMSESD RsquaredSD     MAESD
## 1      TRUE 13.44655 0.07816067 11.46791 0.1955585 0.01890138 0.2010174
```

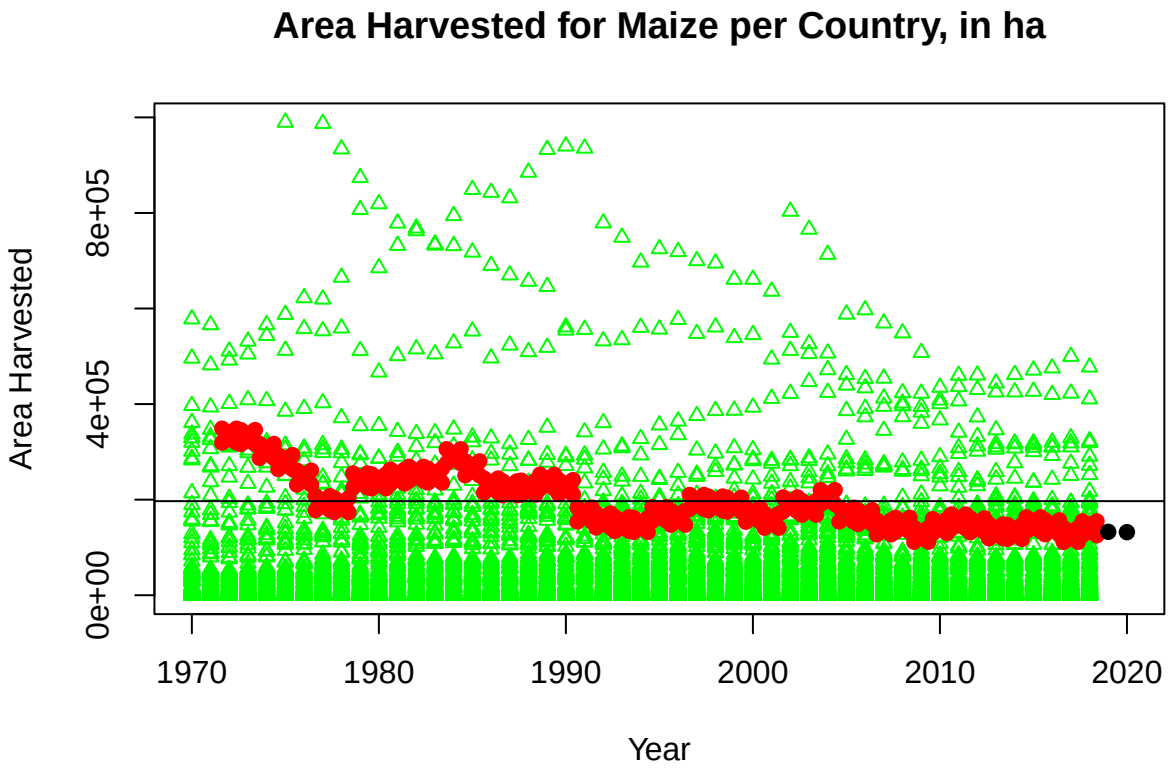


And maybe this is why production has gone up? Average yields have slowly risen, with 2019/2020 confirming this trend.

Area Harvested

```
## [1] "10 K-fold CV results:"
```

```
##   intercept     RMSE    Rsquared      MAE    RMSESD RsquaredSD     MAESD
## 1      TRUE 13.98691 0.003443009 12.04146 0.1199799  0.0021337 0.1147166
```



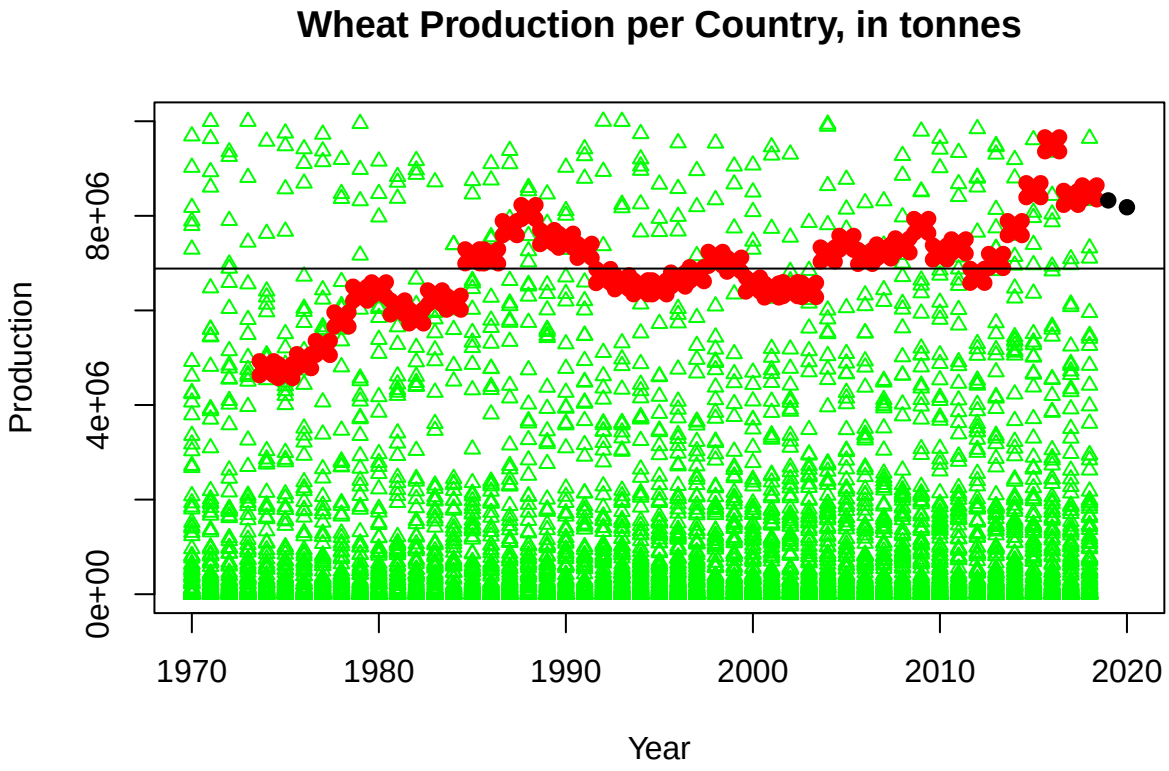
Area harvested is predicted to go down for maize, and 2019/2020 seem to be reaching towards the lowest means. But as we saw, production and yields have gone up for maize, suggesting much greater efficiency in farming maize in the past 50 years.

Wheat

Production

```
## [1] "10 K-fold CV results:"
```

```
##   intercept     RMSE    Rsquared      MAE    RMSESD  RsquaredSD      MAESD
## 1      TRUE 13.96762 0.001669898 12.01543 0.1206289 0.002215778 0.09156476
```

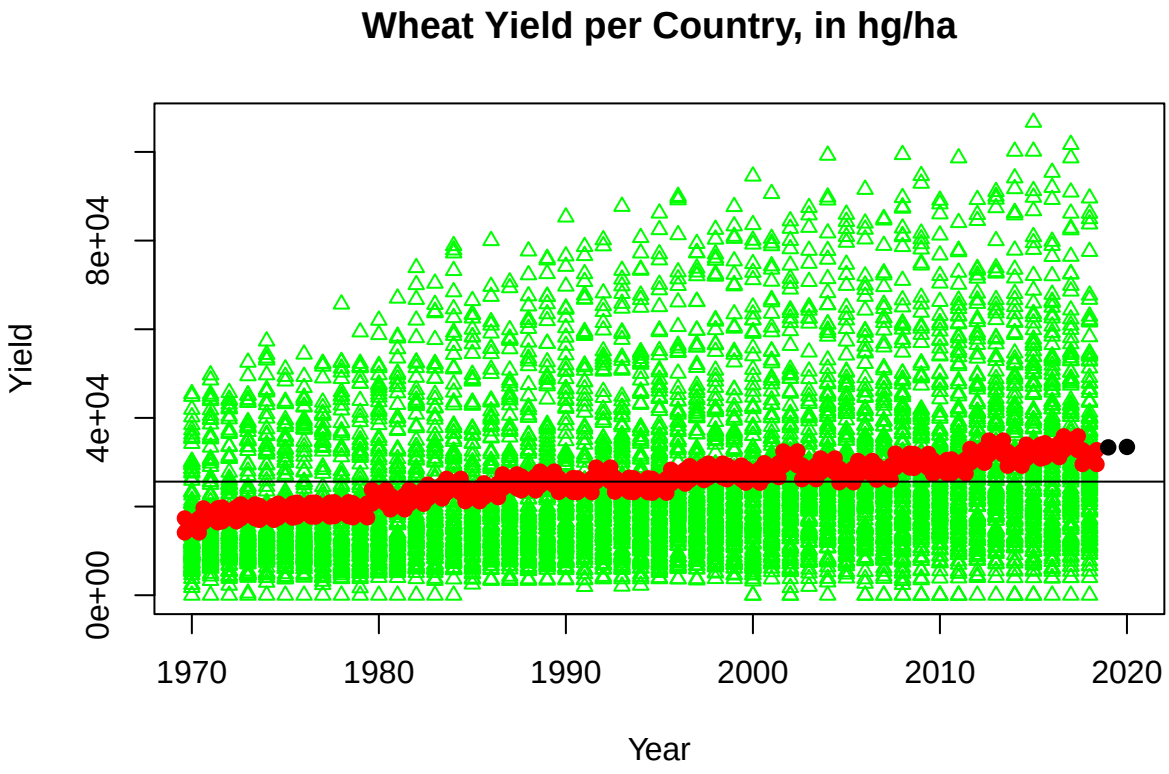


Production is seemingly unstable for wheat over time. Its possible our k value is distorting our view here a little. The 1980's peak is fascinating; maybe the break up of the Soviet Union caused troubles for wheat production in the early 90's? There was a resurgence in early 2010, and the 2019/2020 values seem to suggest wheat production is rising again.

Yield

```
## [1] "10 K-fold CV results:"
```

```
##   intercept     RMSE   Rsquared      MAE    RMSESD RsquaredSD     MAESD
## 1      TRUE 13.56453 0.05929193 11.61982 0.1479664 0.02128748 0.1664695
```

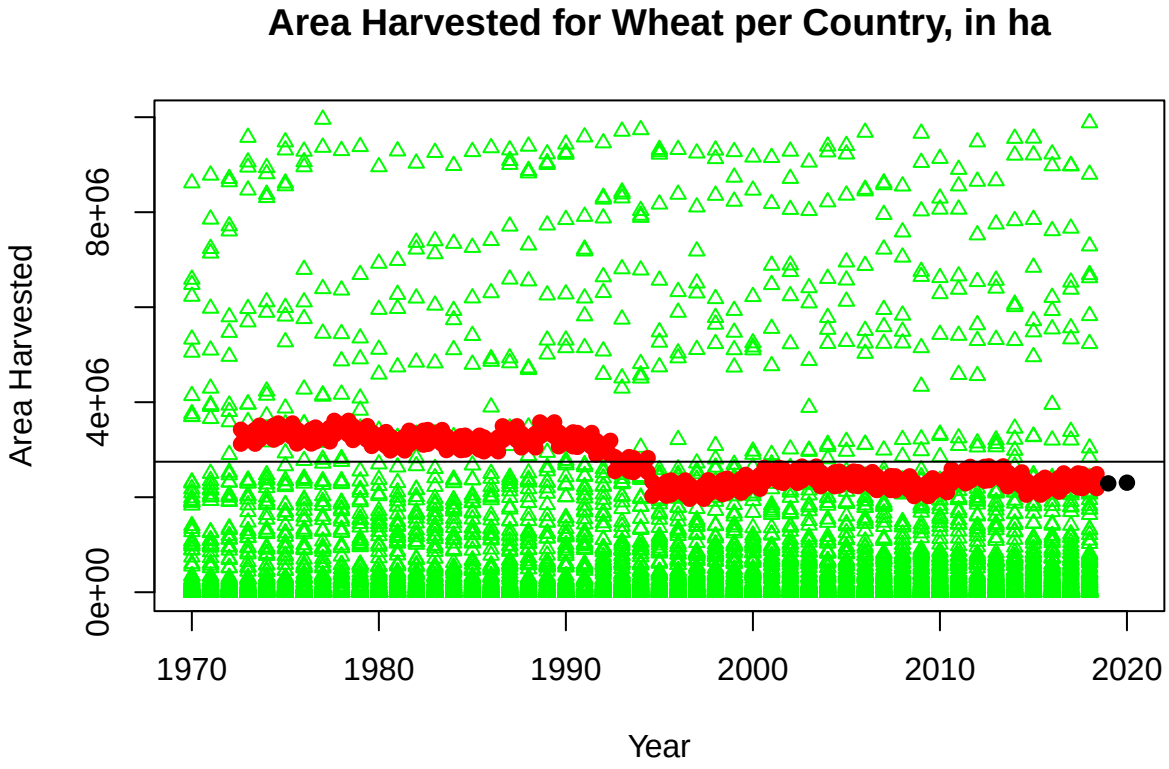


Wheat yields show a similar story to maize: improved efficiency steadily throughout time, with a 2019/2020 peak.

Area Harvested

```
## [1] "10 K-fold CV results:"
```

```
##   intercept     RMSE    Rsquared      MAE    RMSESD  RsquaredSD     MAESD
## 1      TRUE 13.95947 0.002701445 12.0015 0.1772785 0.003337413 0.1711659
```



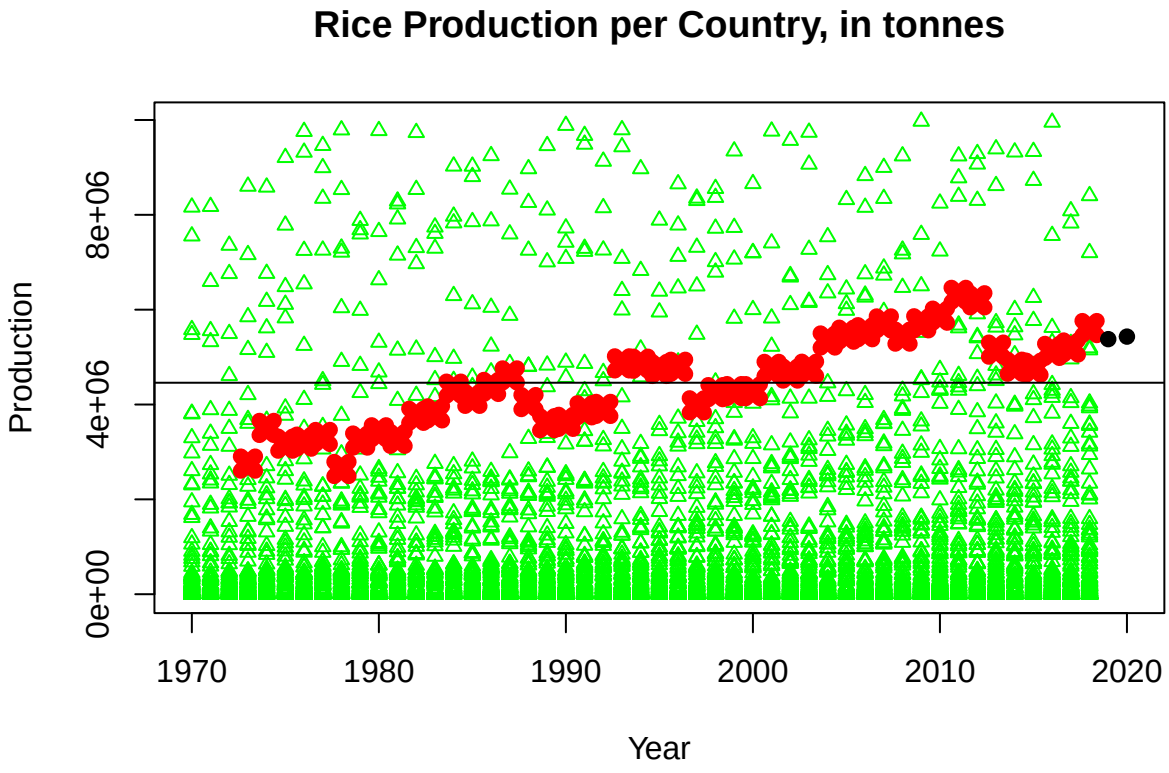
And again, similar to maize, harvested area for wheat is going down over time. Its clear that efficiency has improved for wheat too, regardless of overall production.

Rice

Production

```
## [1] "10 K-fold CV results:"
```

```
##   intercept     RMSE    Rsquared      MAE    RMSESD  RsquaredSD     MAESD
## 1      TRUE 14.10391 0.003736627 12.18078 0.1393434 0.003537431 0.1531489
```

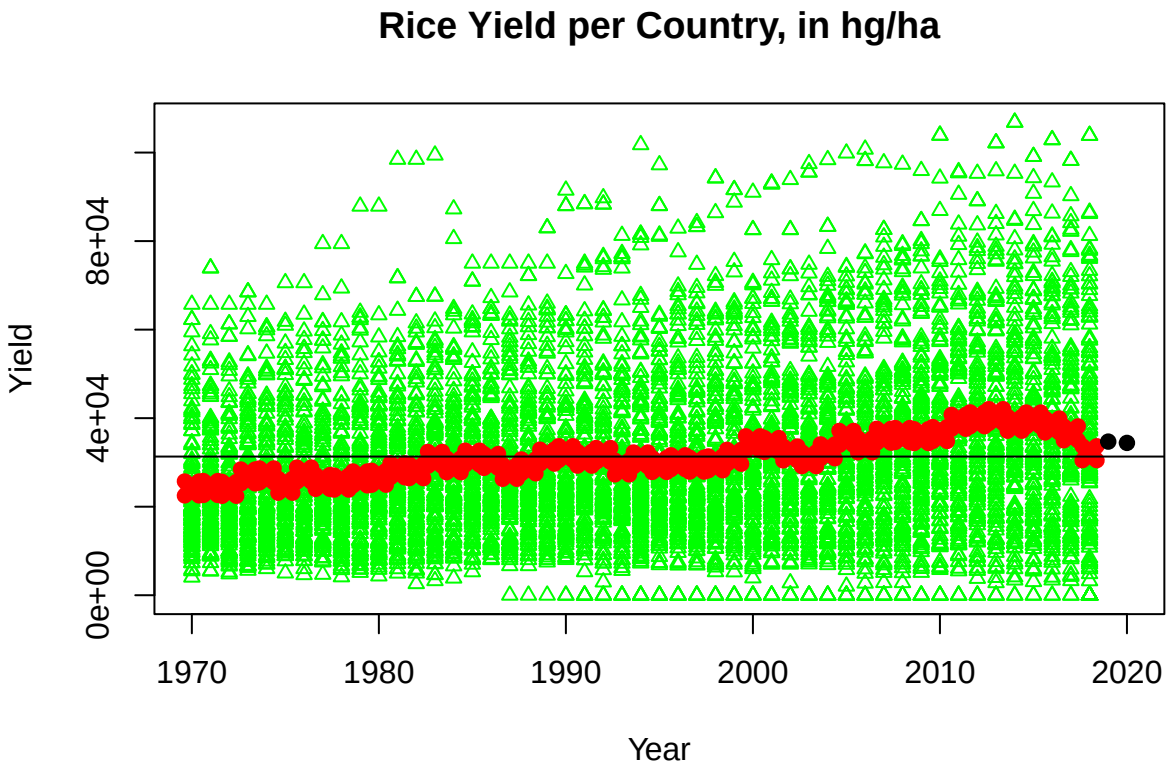


Much variability in the rice production predictions, but its clear that our knn is predicting rice is on a general upward trend too, with 2019/2020 maybe slightly lower than where they should be expected to be.

Yield

```
## [1] "10 K-fold CV results:"
```

```
##   intercept     RMSE   Rsquared      MAE    RMSESD  RsquaredSD     MAESD
## 1      TRUE 13.79246 0.04654593 11.86711 0.1020788 0.009908325 0.1093048
```

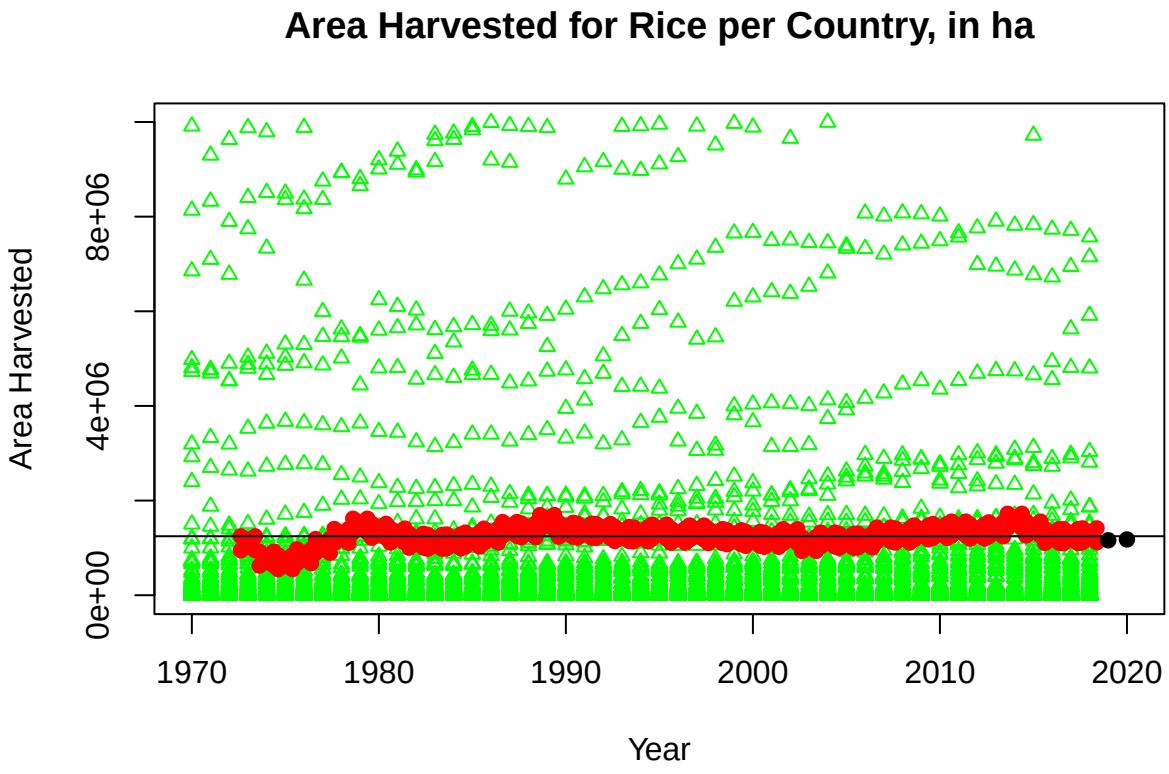


A steady rise in efficiency, however the drop towards the end is concerning. This might be a fluke in the data, or possibly a sign of serious disruption in rice yields; its clear that 2019/2020 yields are lower than where we might expect them to be as well.

Area Harvested

```
## [1] "10 K-fold CV results:"
```

```
##   intercept     RMSE    Rsquared      MAE     RMSESD  RsquaredSD     MAESD
## 1      TRUE 14.12227 0.001750331 12.19966 0.08647582 0.001752647 0.1163544
```



Area harvested for rice seems to be relatively constant, although the scale of our plot might be obscuring the ups and downs a bit. 2019/2020 are right on the mean for predictions. Overall, rice seems to be increasing in production and yields in general over time.

Conclusion

Overall, the impact of COVID on these 4 crops seem non existent.

The rise in yields is clear in all the crops. The rise in all 4 crops for overall production seems to be rising too, except for potatoes, where our k value might be obscuring this information. At the same time, area harvested is predicted to have either gone down, or remained relatively flat, showcasing at the bare minimum an improvement in efficiency for all crops. All of the RMSE for our k-fold are around 14, which is a sign that its fairly consistent at predicting throughout all of our data.