SPPPACY: Satellite Prediction of Per-Pixel Aggregate Crop Yield

Yuval Epstain Ofek and Richard Lee

Introduction

The goal of our work was to predict locations that, should corn be planted there, augment aggregate crop yield without being biased by current crop locations. The main benefit of such a project is to predict locations for new corn farms in order to maximize yield. Our tool can also be used to analyze current corn farm locations and predict how much yield the farm generates (in bushels/acre). This is important in order to identify under/over- performing farms for investments of any future analysis using yield.

One significant concern during a yield analysis is the lack of accurate high-resolution yield data. We were only able to find county-wise aggregate corn yield data predating the last decade. As such, we have developed a methodology that uses low resolution county data to predict high-resolution pixel-wise yield through the use of histograms.

Another concern is obtaining unbiased satellite data. Growing crops naturally changes the face of the land, and so current crops bias any spectral data that can be captured using satellites. In order to create a tool that is not biased by currently present crops, we opted to use only sensor data that is non-spectral and independent of anything on the ground, namely environmental and climate readings. These metrics have a natural correlation between them and crop production, while also not being dependant on current crops or anything present on the land.

The report will be structured as follows: in section 1 we describe the procedure for our experiments. In section 2 we display our results, compare our model to the baseline predictor, and discuss challenges faced throughout the work. Finally, in section 3 we conclude the work and provide ideas for future work.

1 Procedure

Herein we describe our application of using low resolution county-wise yield data to generate pixel-wise yield readings. The methodology takes advantage of the nature of histograms and uses large spatial-temporal histograms as a means of training a model to predict yield of any histogram input data of appropriate size, lending itself to any-scale pixel-wise prediction.

The method uses county-wise yield data found on the Iowa State University website. The data reports yield readings per Iowa county from 2011-2020. Some year-county pairs do not have yield values listed, and are treated as missing data. The yield data is reproduced in Appendix 1. This data is also visualized in figure 1.

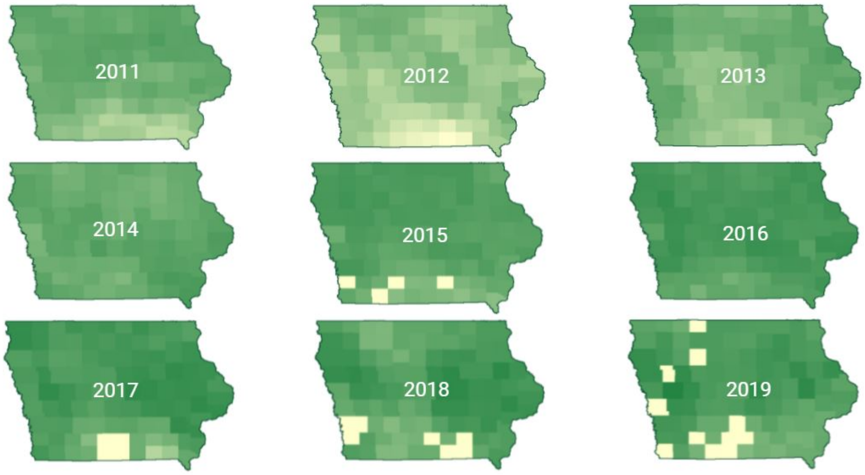


Figure 1: Visualization of corn yield data. Green corresponds to high yield and Yellow low yield. NaN values displayed as 0 yield.

We began the analysis itself by taking a environmental data from the 'Daymet V4: Daily Surface Weather and Climatological Summaries' dataset and the 'ERA5 Daily aggregates - Latest climate reanalysis produced by ECMWF / Copernicus Climate Change Service' dataset. We took data across the crop growing season, which we exaggerated to be from the beginning of April to the end of October, from 2013 up to 2020. The yield data began at 2011, but due to our original intention of using Landsat 8 data, we opted to only extract data from 2013-2019.

With data for each year, we proceeded to mask it using a mask generated from the USDA NASS Cropland Data Layers dataset. The mask was constructed by taking the pixels with 95% confidence of being crops, cultivated land, and only corn.

The masked images were then cropped to a given county and then aggregated band-wise both temporally and spatially into histograms of fixed bin sizes N, we set $N = 100$ bins for our experiments. The result of this was a set of B_i histograms for every year-county pair, where B_i is the number of bands in image collection i. In our experiment $B_{Daymet} = 7, B_{ERA5} = 9$. Once all of the image collections were converted into histograms, the bands for a given year-county pair (across all image collections) were concatenated together. The size of the resulting data for each year-county pair then became (N, B_{Total}), where $B_{Total} = \Sigma B_i$.

Taking into account missing yield data, the input for our tests in roughly of size $(650, 100, 16)$, where the first index refers to the number of year-county pairs. Once the histogram data was constructed, each histogram was normalized by dividing all element by the largest histogram bin count.

Once the data was ready, a deep learning regression model was created. The model can be seen in figure 2. The model used a 1D convolutional layer to take advantage of bin spatial relationship. This was followed by a spatial dropout for regularization, and a set of dense layers with relu activations to predict yield. The model was trained using the Adam optimizer, with a batch size of 8, and using mean square error (MSE) as a loss. We chose mean squared error as our loss even though our goal was to minimize mean absolute error (MAE), as the larger MSE led to more aggressive gradient updates each epoch and assisted with convergence.

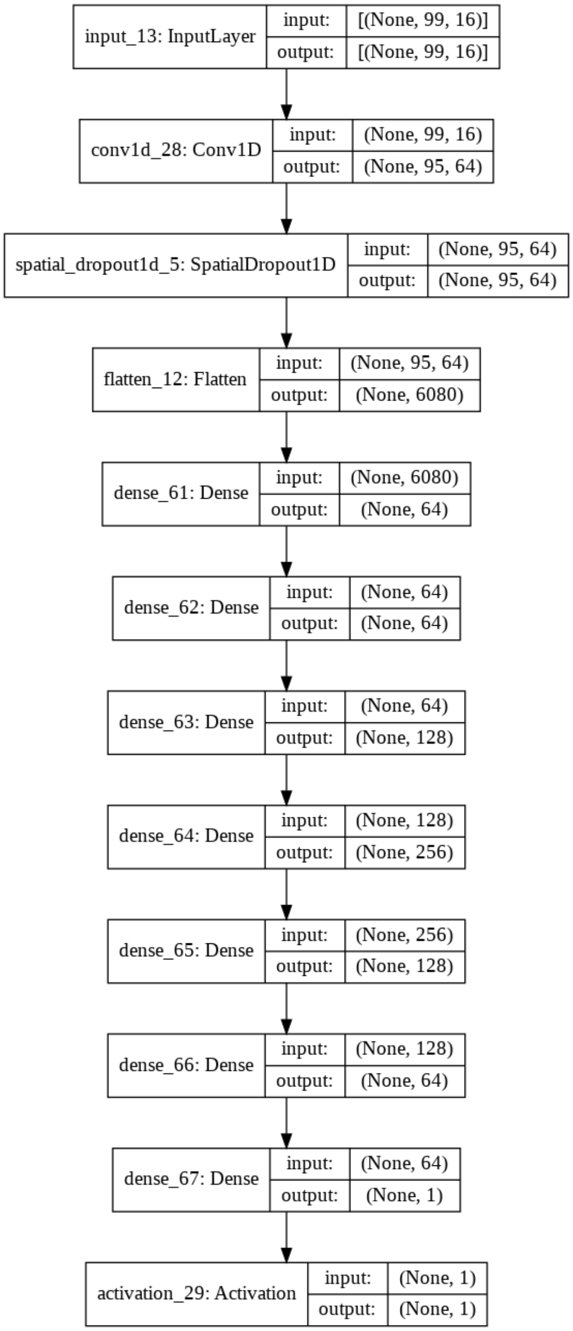


Figure 2: Deep Learning Model

To predict pixel-wise yield maps, temporal histograms at each pixel were calculated with the same number of bins N and total bands B_{Total} as the training data. The data should then be size $(Map_x, Map_y, N, B_{Total})$, where Map_x and Map_y are the map dimensions. The pixel wise data was fed into the model, and pixel-wise yield predictions were formed (in bushels/acre).

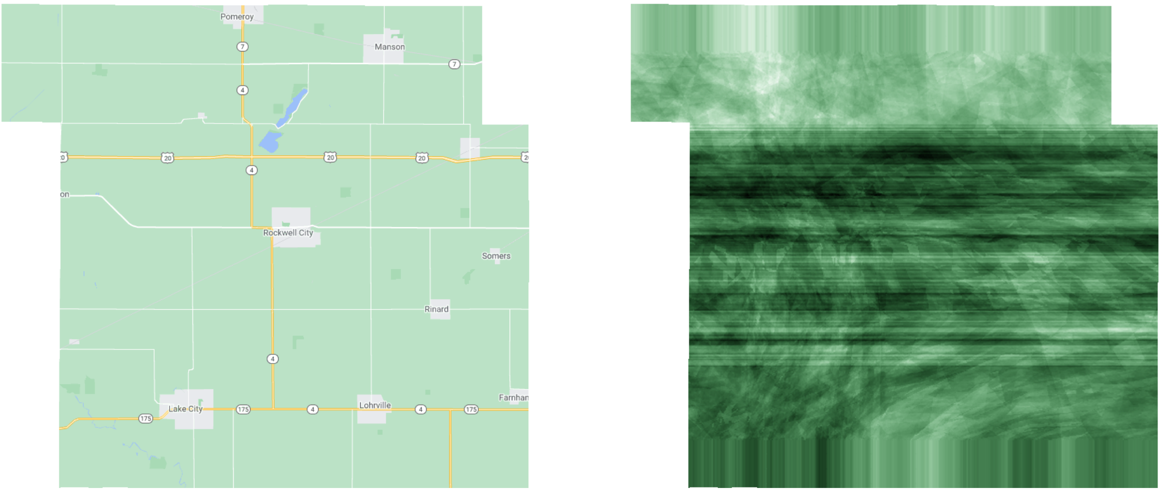
To summarize, we first compile a dataset of spatial-temporal histograms of across each of the input bands per county per year for all crop pixels. We then train a deep learning model using a these histograms after normalization and the yield data. Finally, we generate pixel-wise temporal histograms at each pixel and evaluate the model using those to generate a pixel-wise yield map.

2 Results and Discussion

We tested against a validation set that was held out during training. To evaluate the baseline performance, we predicted the county yield for each county in the validation set. This yielded a mean squared error of 734.297 and a mean absolute error of 20.637 bushels/acre. Using the same validation set and the predictions from our model, we obtained an MSE of 203.6 and an MAE of 10.7 bushels/acre, which indicates that the model was successful in capturing features of the data to perform yield prediction.

The goal of this project was to obtain pixel-wise masks of a region to predict yield. However, since our training data and aggregate yield are at the county level, the actual value predicted during inference for a pixel would not be valid. Instead, the model output can be used to show a relative yield compared to other pixels: the histogram of a high yielding county should resemble the histogram of a pixel that *could* have high yield, if crops (corn) were planted there.

Figure 3 shows a crop mask for Calhoun County, Iowa in 2015.



(a) Map of Calhoun County, Iowa.

(b) Model Output.

Figure 3: Visualization of Calhoun County and the predicted pixel-wise relative yield heat map. Higher brightness in 3b indicates higher predicted relative yield.

Looking at the generated model output, we can see that the there are a lot of horizontal and vertical artifacts present. We attribute this to the low resolution of the input data requiring interpolation to scale the environmental data to the resolution of the CDL data. However, some general trends can be seen as the model does predict some areas to be higher relative yield (lighter green) and some areas to be lower relative yield (darker green). However, this can be difficult to compare the results, so we use the Cropland Data Layer for corn and its inverse to clip the model output (Figure 4 and 5).

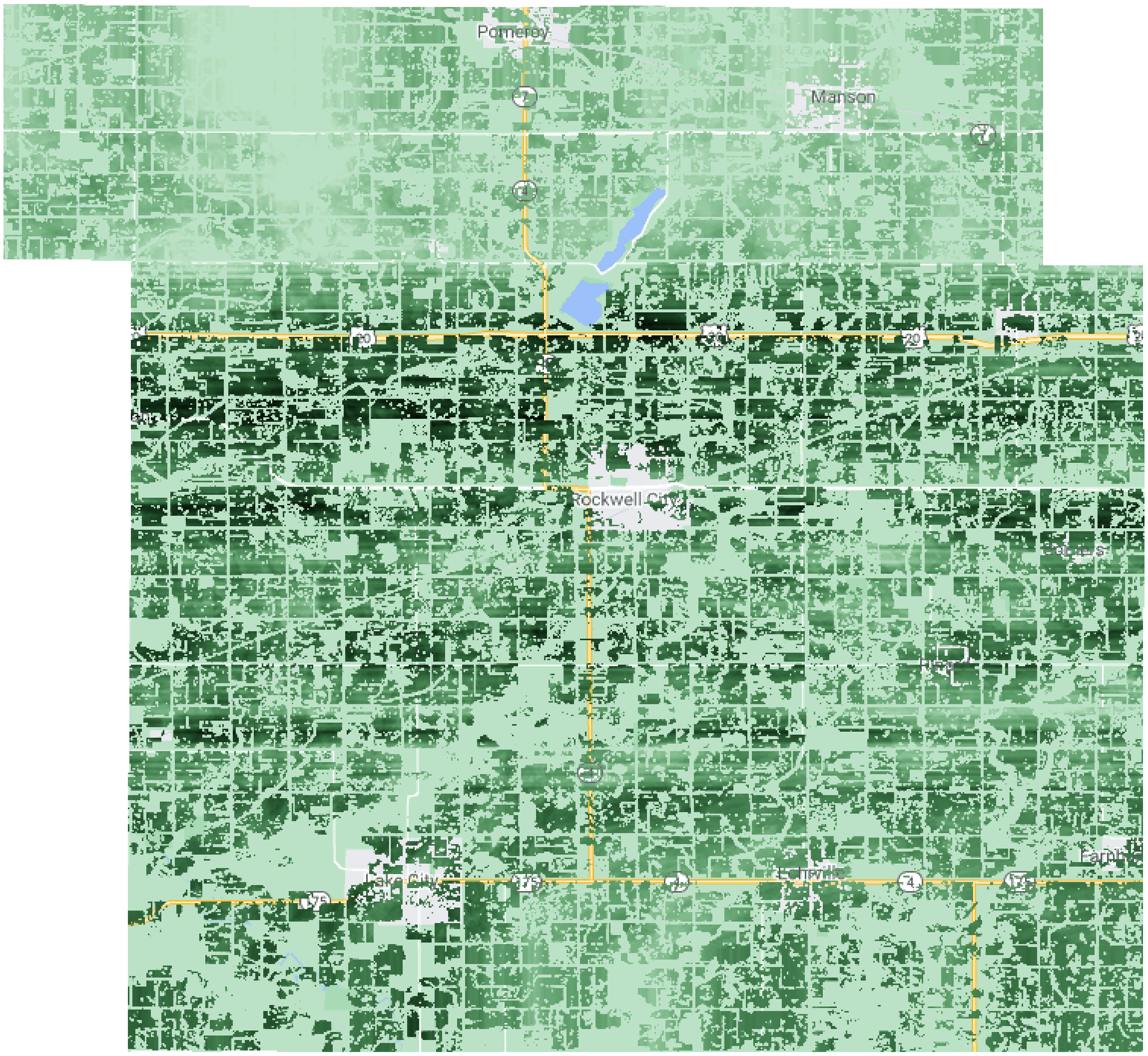


Figure 4: Clipping the model to areas with crops.

We can see that the region near the top was generally predicted to be very high yield, though the regularity of the region seems to suggest that this might have been another artifact. However, the region surrounding State Highway 20 does appear to be darker than the regions above and below it, which is intuitive, and seems to indicate that the model has learned.

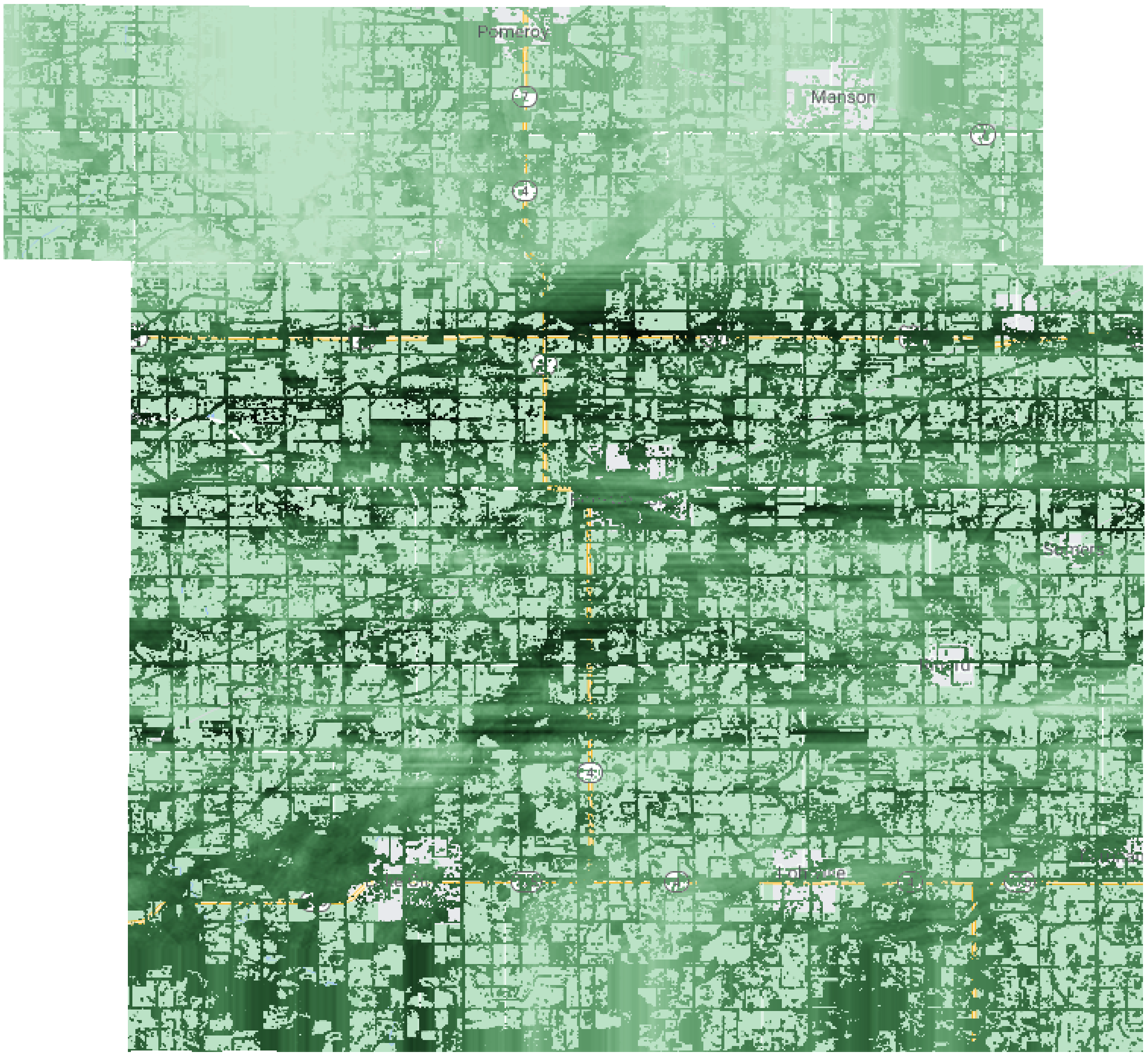


Figure 5: Clipping the model output to the areas with no crops.

Clipping to the areas with no crops shows that the pixels that are currently non-crop regions were generally predicted to have lower yield potential (large swathes of darker pixels). Again, the upper portion of Calhoun county seems to have been predicted to be particularly high yielding, which again, might be due to the low spatial resolution of the input data.

2.1 Limitations

Due to the time constraint of the project, we chose the minimal spatial resolution available from Earth Engine in order to maximize the chances of the download finishing. This resulted in counties of roughly 5x5 pixels (Figure 6). In order to match the Cropland Data Layers mask resolution, the image was interpolated using bicubic interpolation. Such low resolution interpolated up to such a high resolution, invariably resulted in noise in the training data and therefore the final output, as seen in the previous plots.

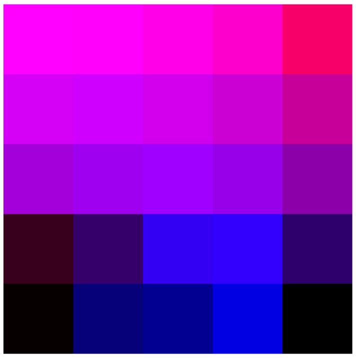


Figure 6: Adair County, Iowa; DaymetV4. April 1$^{\mathrm{st}}$, 2013.

Also, while histograms to represent distributions of each of the features is an interesting way of analyzing the data, the operation scales quadratically with respect to the image dimension (or linear with respect to the number of pixels), which means that generating histograms for high resolution data is also very expensive, but the massively parallelizable nature of it does seem to suggest that it would be a good candidate for GPUs.

In summary, while not being able to validate our pixel-wise predictions due to the inherent lack of yield-data, by the nature that the county-wise histogram used for training is an aggregation for pixel-wise temporal histograms, our methodology holds ground.

3 Conclusions

Our work shows promise at predicting yield on a pixel-wise basis while only being trained on county-wise data. Our county-wise model was able to outperform the baseline predictor significantly, resulting in roughly half the mean absolute error of the baseline. Furthermore, by the nature of the input dataset formation, yield estimates at any resolution can be found. The input data is also inherently independent of any surface structure or present crops, allowing our work provides an unbiased estimation of yield for any given pixel.

Future Work

Our work is an initial step towards creating an unbiased model to predict corn viability, yet more can be done to improve its effectiveness. We list a number of our ideas to do so as follows:

- Due to the limitations of using only environmental and climate data, the model proceeds to predict yield for places where corps cannot be grown at all due to geological limitations (such as water sources or cities). Supplementing the predictions with readings from a land cover classification will enable the model to predict yield for only places that *can* grow corn.
- Another challenge with using non-spectral data is that it forces an assumption that the only cause for crop growth is environmental data. This is a faulty assumption, as crops depend on the soil to grow as well. One way to include this aspect in the model is to create a soil/land segmentation map and use it as a skip connection for the model, enabling the model to use soil information in its prediction while still remaining unbiased by current crops.
- Due to computational limitations, our analysis used low-resolution data from a limited set of data sources and for one specific crop: corn. To provide a more robust system, we propose using higher resolution data, obtaining yield for more crops, and increasing the number of bands that the model has available for its predictions.
- Linked to the prior comment, conducting a rigorous band selection prior to training a model is also a topic of interest. For one, it will provide researchers a better understanding on how environmental factors affect yield and specifically for this work it will limit the amount of extraneous data the model sees and prevent overfitting to it.
- We also propose choosing a wider range of spatial locations for the analysis. This will allow the model to generalize better and might mean there would be more data to use for training.
- Lastly, we comment that prior works have shown success in using NDVI to predict yield. An alternative idea we propose is using NDVI as an intermediate variable for yield prediction, or in other words predicting NDVI from environmental factors and then yield from NDVI. The results of such analysis rely on prior works to support their pixel-wise yield predictions and might yield better results.

Appendix 1 - Yield Data

	2011	2012	2013	2014	2015	2016	2017	2018	2019	2020	Avg.
Adair	152.8	104.4	137.8	169.3	176.5	190.3	175.2	149.5	176.6	169.7	160.2
Adams	151.7	109.8	147.6	169.9	169.6	187.4	179.9	171.6	NaN	174.2	162.4
Allamakee	173.3	153.6	162.1	175.6	175.4	205.1	190.5	198.4	187.2	197.4	181.9
Appanoose	120.1	44.5	109.1	170.9	154.3	186.3	170.0	NaN	166.8	145.0	140.8
Audubon	176.6	114.8	157.1	185.5	189.8	206.1	211.2	180.0	216.0	171.7	180.9
Benton	157.2	129.3	163.7	185.4	187.5	199.9	215.3	214.6	197.6	145.6	179.6
Black Hawk	186.1	126.4	166.1	165.2	192.9	206.4	209.2	206.2	205.1	170.8	183.4
Boone	182.5	147.1	154.2	180.9	192.3	208.4	192.4	193.7	194.8	161.6	180.8
Bremer	196.5	133.3	175.1	159.0	203.1	210.6	212.7	212.4	212.5	188.7	190.4
Buchanan	187.0	140.4	168.8	169.3	191.6	205.6	221.3	207.1	218.9	189.3	189.9
Buena Vista	180.2	148.2	160.5	172.1	202.7	201.1	187.9	193.1	190.8	184.9	182.2
Butler	186.7	122.0	179.2	167.7	193.3	207.3	211.4	211.3	209.5	193.9	188.2
Calhoun	181.9	134.8	132.6	189.6	194.7	207.3	195.1	193.3	198.8	160.5	178.9
Carroll	180.2	105.3	136.5	188.6	195.4	203.1	211.3	208.4	219.7	151.4	180.0
Cass	175.2	120.4	159.5	173.7	193.7	195.9	197.4	169.5	200.3	178.5	176.4
Cedar	184.9	158.9	169.7	184.1	188.8	215.3	222.3	222.3	193.7	154.1	189.4
Cerro Gordo	169.3	124.8	163.1	168.3	196.9	193.6	208.9	184.4	191.8	183.6	178.5
Cherokee	176.7	158.7	185.1	183.6	209.6	219.7	203.7	213.3	208.9	186.8	194.6
Chickasaw	188.8	125.8	171.8	157.6	202.8	206.6	194.4	192.9	212.1	186.8	184.0
Clarke	107.7	76.7	126.5	162.2	152.4	163.9	NaN	161.8	NaN	156.1	138.4
Clay	185.4	170.1	167.2	166.3	202.8	202.0	197.3	158.5	185.5	169.1	180.4
Clayton	186.8	152.9	181.2	176.7	193.2	210.3	213.6	201.5	197.5	192.1	190.6
Clinton	180.0	139.6	185.6	197.0	188.2	213.8	213.3	210.9	196.1	179.4	190.4
Crawford	181.8	131.3	157.8	183.0	199.1	213.5	221.9	214.0	234.7	182.3	191.9
Dallas	172.1	129.0	145.3	195.5	182.4	204.0	187.2	174.0	195.0	140.5	172.5
Davis	97.8	52.4	146.7	184.2	142.4	188.7	107.3	NaN	160.0	170.9	138.9
Decatur	126.1	61.2	131.2	164.0	162.7	181.1	NaN	172.5	NaN	160.2	144.9
Delaware	186.5	129.8	184.7	174.9	195.8	202.3	218.0	224.8	204.6	191.5	191.3
Des Moines	146.6	143.7	165.4	199.0	176.0	210.0	200.1	195.4	176.9	191.3	180.4
Dickinson	169.3	160.9	162.4	157.9	194.5	188.0	196.2	160.4	170.0	183.2	174.3
Dubuque	185.1	149.4	196.0	184.9	197.0	211.5	215.1	211.8	216.4	198.3	196.6
Emmet	169.6	166.6	166.4	165.0	203.3	198.5	210.1	155.8	NaN	187.2	180.3
Fayette	186.6	151.0	181.2	170.9	191.8	202.5	207.6	200.1	200.5	181.1	187.3
Floyd	181.1	123.2	170.0	162.8	196.5	196.3	203.4	184.4	202.9	188.5	180.9
Franklin	189.7	150.0	179.6	165.8	200.2	204.8	206.1	204.3	202.7	181.1	188.4
Fremont	155.6	131.2	159.9	191.5	179.9	180.8	204.5	195.1	NaN	195.9	177.2
Greene	183.4	124.6	137.9	181.2	189.0	205.5	202.2	201.3	201.4	154.6	178.1
Grundy	172.8	161.5	180.8	182.4	199.9	198.5	217.3	225.2	209.8	183.7	193.2
Guthrie	169.4	113.4	130.1	176.1	179.2	198.2	197.6	192.2	196.7	152.2	170.5
Hamilton	180.6	138.0	135.0	173.1	198.1	209.0	194.6	188.9	198.2	167.2	178.3
Hancock	179.8	143.3	169.9	173.9	201.2	200.3	200.2	176.7	193.3	201.2	184.0
Hardin	186.7	163.6	156.6	166.3	199.3	208.0	213.7	216.6	200.2	156.6	186.8
Harrison	169.9	130.9	177.1	161.1	189.5	204.6	193.2	176.0	NaN	181.2	175.9
Henry	123.2	132.0	160.4	196.6	178.4	199.0	196.4	189.7	164.5	173.0	171.3
Howard	185.8	137.8	159.1	164.4	202.1	202.2	201.7	185.5	194.2	189.7	182.3
Humboldt	185.5	146.1	155.4	176.9	194.8	206.1	200.9	162.3	198.7	189.5	181.6
Ida	182.9	151.4	177.7	189.9	203.1	216.0	215.7	222.2	NaN	189.3	194.2
Iowa	171.6	135.7	165.9	192.1	199.3	210.6	216.7	211.7	190.6	151.9	184.6
Jackson	178.3	114.6	173.2	188.9	190.6	202.1	202.1	202.4	185.5	186.7	182.4
Jasper	171.1	150.9	158.8	191.5	198.9	214.7	204.6	218.6	209.3	NaN	190.9

	2011	2012	2013	2014	2015	2016	2017	2018	2019	2020	Avg.
Jefferson	112.4	102.8	145.5	200.9	168.2	197.0	165.5	198.8	162.7	168.3	162.2
Johnson	171.9	132.4	181.1	184.6	187.1	189.0	213.1	212.3	181.3	162.2	181.5
Jones	171.4	127.3	179.3	180.1	186.6	201.3	217.0	208.9	203.0	172.6	184.8
Keokuk	152.4	113.0	156.7	193.2	181.0	208.6	176.4	203.7	175.5	164.4	172.5
Kossuth	179.9	164.6	167.7	176.6	199.0	204.7	205.6	180.2	198.7	191.5	186.9
Lee	105.2	116.0	144.1	200.6	145.1	194.6	178.0	201.3	172.5	175.3	163.3
Linn	169.1	124.2	178.6	176.1	189.2	200.3	217.4	215.4	223.7	165.9	186.0
Louisa	148.9	162.9	158.5	188.2	181.7	200.2	216.6	196.5	174.6	182.6	181.1
Lucas	114.9	85.3	117.9	153.7	154.8	174.6	NaN	NaN	NaN	150.1	135.9
Lyon	185.4	161.6	190.5	172.9	196.0	206.9	218.9	193.0	186.6	198.1	191.0
Madison	145.1	101.4	140.9	171.4	169.6	185.5	173.1	167.8	175.7	172.9	160.3
Mahaska	165.0	133.0	168.6	189.9	190.3	204.7	177.2	207.7	189.9	181.8	180.8
Marion	152.6	115.1	156.4	171.6	178.1	198.0	185.7	190.2	170.7	NaN	168.7
Marshall	166.3	157.6	163.2	190.5	194.9	211.3	222.4	226.0	221.9	NaN	194.9
Mills	155.0	128.8	171.0	184.9	NaN	188.4	190.1	NaN	202.9	183.2	175.5
Mitchell	180.1	131.4	164.1	175.1	202.1	199.1	213.4	193.3	194.5	204.5	185.8
Monona	158.3	115.1	175.2	156.6	169.0	184.5	196.0	181.4	203.9	179.4	171.9
Monroe	112.5	67.3	115.3	175.7	NaN	180.0	169.9	167.0	162.3	156.0	145.1
Montgomery	153.2	132.0	154.0	163.1	185.2	185.5	199.1	191.5	204.8	185.6	175.4
Muscatine	164.6	152.6	158.3	188.1	177.7	203.0	206.8	209.6	173.2	173.8	180.8
O Brien	185.1	159.8	195.3	180.3	205.2	210.7	213.8	201.6	204.9	200.6	195.7
Osceola	184.4	169.6	182.0	175.4	204.7	203.7	211.6	180.2	182.5	186.3	188.0
Page	136.2	116.2	154.5	182.4	165.8	185.8	193.1	188.1	191.4	182.6	169.6
Palo Alto	174.6	170.6	168.8	164.6	197.5	196.0	195.4	157.9	198.3	182.1	180.6
Plymouth	164.5	110.1	188.9	187.3	203.6	209.2	193.0	213.8	217.8	NaN	187.6
Pocahontas	181.2	165.3	163.8	174.8	205.5	206.8	190.9	179.2	NaN	187.3	183.9
Polk	160.6	149.1	143.6	181.4	187.0	199.8	200.3	186.2	196.7	151.0	175.6
Pottawattamie	170.9	131.9	179.4	169.1	204.0	210.4	195.8	NaN	208.0	177.3	183.0
Poweshiek	179.0	152.3	158.5	188.4	197.8	213.2	215.1	218.7	200.3	139.0	186.2
Ringgold	118.7	71.6	121.1	166.4	147.8	179.5	167.8	153.3	NaN	160.0	142.9
Sac	183.4	130.7	142.6	179.8	204.8	214.3	194.5	213.5	220.6	161.9	184.6
Scott	174.2	131.7	167.8	194.5	195.9	216.0	220.0	214.0	189.3	182.3	188.6
Shelby	173.6	137.8	182.6	185.2	195.3	215.3	207.1	189.1	217.3	191.6	189.5
Sioux	177.6	143.2	194.1	187.4	201.9	212.2	220.7	211.9	192.2	203.0	194.4
Story	163.2	157.6	137.2	169.9	188.0	211.9	200.3	192.3	189.3	149.2	175.9
Tama	162.0	160.2	158.7	182.7	189.5	208.4	220.6	224.3	204.8	133.5	184.5
Taylor	140.5	91.7	136.2	178.8	NaN	159.9	175.0	154.4	161.7	167.2	151.7
Union	132.6	87.3	137.0	176.1	NaN	183.0	147.9	155.8	168.9	180.7	152.1
Van Buren	100.9	110.3	144.1	192.2	158.3	189.0	128.2	180.9	171.0	154.4	152.9
Wapello	127.3	92.6	150.4	195.2	159.3	196.9	153.0	NaN	176.5	171.6	158.1
Warren	129.1	103.5	146.7	169.9	170.3	184.6	183.2	163.6	NaN	169.0	157.8
Washington	156.3	132.1	159.0	191.3	195.0	208.2	220.4	214.3	178.9	186.5	184.2
Wayne	126.0	55.7	125.1	171.9	158.0	174.5	NaN	171.6	151.5	169.7	144.9
Webster	190.6	149.2	137.6	190.0	201.0	200.8	197.9	188.1	202.9	174.0	183.2
Winnebago	176.6	161.1	149.1	172.7	203.6	197.9	210.7	180.8	198.8	208.3	186.0
Winneshiek	186.2	140.6	181.8	174.0	185.6	210.7	199.7	198.1	200.4	179.2	185.6
Woodbury	173.3	134.8	164.2	177.5	202.3	203.7	200.5	219.3	218.2	192.1	188.6
Worth	173.0	143.1	150.7	181.5	202.1	191.5	198.5	180.0	196.5	206.9	182.4
Wright	189.9	153.7	168.8	169.4	198.4	200.2	206.9	180.5	197.0	187.1	185.2