inferentialist consulting

inference as unique as your data

Sparse Functional Models: Predicting Crop Yields

Dustin Lennon
Lead Statistician
dustin@inferentialist.com

Executive Summary

Here, we develop a general methodology for extracting optimal functionals mapping time series to real numbers. We refer to this as a "Sparse Functional" (SF) model and use it to predict annual soybean crop yields from precipitation data prior to harvest. We consider county level data in Iowa for the 14 growing seasons spanning 2000 to 2013.

Results are good. The SF model reduces the residual sum of squares over a null model by 66%. Moreover, computation, including cross validation for selection of model complexity, is fast. Fitting 1,108 crop-years worth of data takes 6.98 seconds.

	MSE	RMSE	MAD
Mean Model	12932.71	6.82	5.56
SF-Precip Model	4320.63	3.94	4.01

Table 1: Performance Measures on test set

Finally, the model is easily extended to handle other field data, such as temperature. It is straightforward to add features leveraging domain knowledge, say, for example, consecutive number of freeze days or cumulative rainfall in VC or R5; the latter requiring, in addition, known time intervals associated with soybean growth stage.

Introduction

Recently, I was asked to consider the problem of predicting crop yield given precipitation data. This turns out to be an interesting problem for a number of reasons.

First, the predictive variables are the annual historical time series, whereas the response variable is the yield, measured in bushels per acre, recorded at the end of a growing season. So, a function is to be used as a covariate for predicting a real number.

Second, precipitation data is inherently noisy. A big summer storm rarely repeats annually with any regularity. This is in contrast to temperature data which is, arguably, quite consistent across years. See Figure 1.

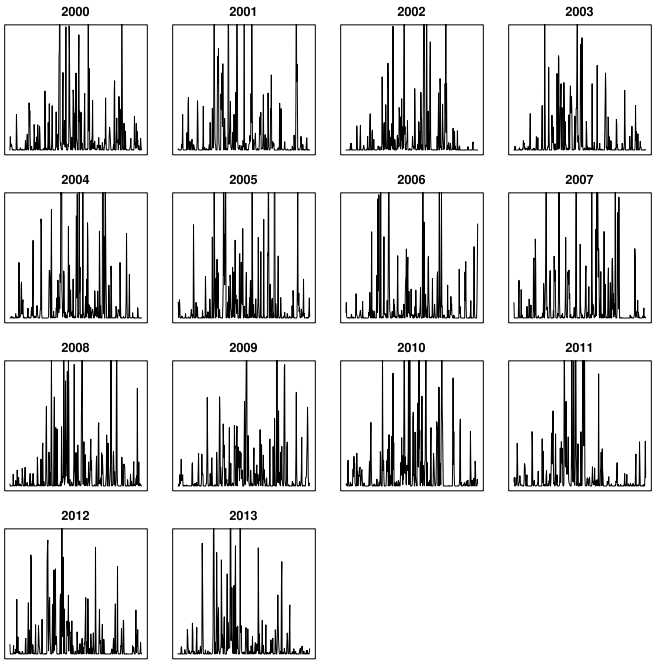

Figure 1: Annual precipitation data for Dickinson County, Iowa

Third, the data, at a county level, are available to the public through government databases. We obtained precipitation data from NOAA/NCDC and crop yield data from USDA/NASS.

Data Preprocessing

Daily historical precipitation data is available as point data from measurement stations. We need to transform this point data into field data and then aggregate it by county. The process used here is straightforward: krige the data onto a uniform grid and then, for each county, report the average over the subset of the grid coincident with the interior of each region.

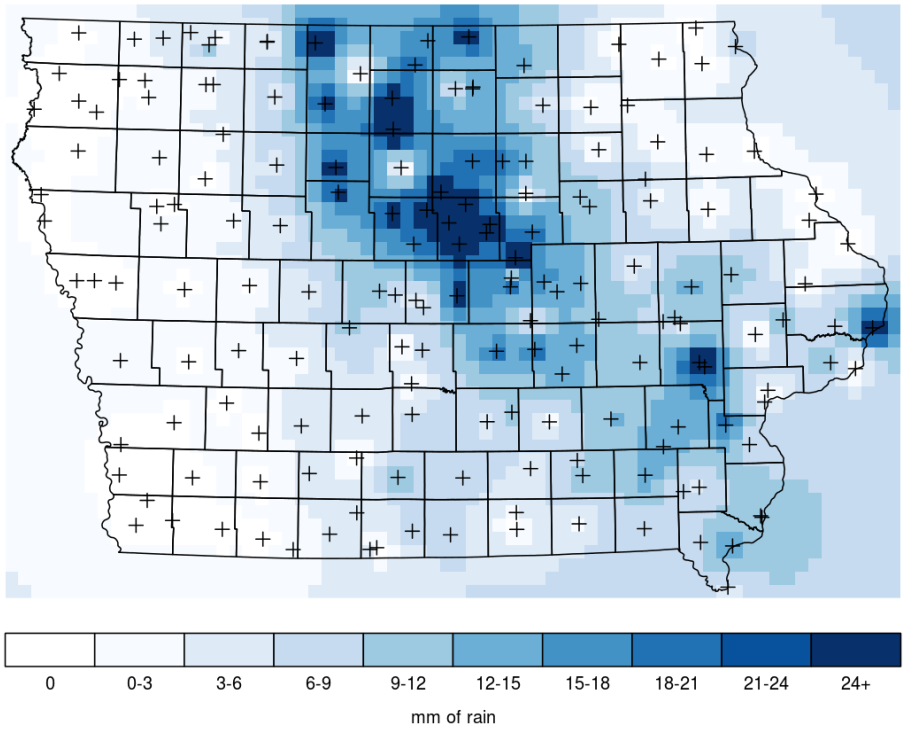

Figure 2: Precipitation kriging, May 26, 2001

Hence, for each (year, county) pair, we obtain a time series of precipitation data and combine this with the county-level response variable, yield in bushels per acre. There are 14 years and 99 counties for a total of 1,386 observations.

Methodology

Sparse Functionals

The key contribution of this paper derives from the following, simple observation. We need to reduce each time series of historical precipitation data into a feature vector. Figure 1 indicates that this should probably involve smoothing of some sort, but which time intervals and scales to choose?

Wavelet theory gives us a starting point. In particular, the simplest wavelets, Haar wavelets, use a hierarchical decomposition into averages and differences. Alternatively, this could be viewed as applying integral- or differential-operators on a specific collection of nested dyadic partitions. This pre-determined partitioning is where the wavelet paradigm falters. It would be great if our interest in precipitation aligned on dyadic boundaries. But that's not the case. Harvest season doesn't necessarily line up with powers of two.

The solution is to break the assumption of a unique wavelet representation. Instead, introduce shifts in order to allow finer grids to replicate signals represented on coarser grids. This is best exhibited by example. In Figure 3, we show the first four levels of the domain decomposition, shown here on a annualized scale.

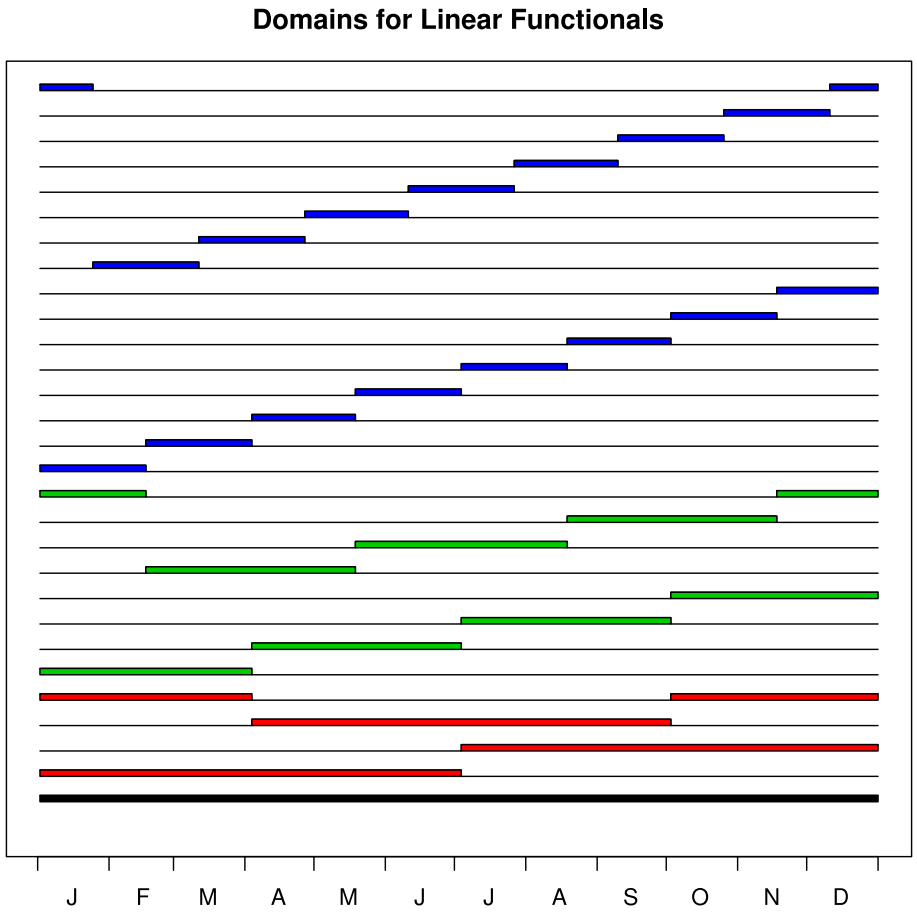

Figure 3: Hierarchical domain decomposition

Note that shifts are introduced so that any pairwise adjacent intervals on the finer scale can be equated with a single interval on the coarser scale.

Now define integro-differential functionals on the subdomains to obtain a feature vector. For our purposes, we used a simple averaging operator to obtain daily rainfall estimates over each respective interval.

Sparse Fitting

We optimize the following objective function, assuming a linear combination of the feature vectors defined in the previous section:

$$\min_{\mu,\alpha} \|y - \mu - G\alpha\|_2^2 + \tau\|\alpha\|_1 \tag{1}$$

where y is the vector of county-level crop yields, μ is an overall mean across year and county, and G is the associated model matrix comprised of the sparse functionals.

τ controls sparsity and is chosen, by cross validation, to lie within one deviation of the minimum. To do the fitting, we used the glmnet R package described in [1].

Visualizing the coefficients is, perhaps, of interest. See Figure 4. The black bars are the coefficients on their respective subdomains of definition. The red curve is the sum of the coefficients at each point in time. The effect of precipitation is, not surprisingly, tightly bound to the growing season which runs from April to September.

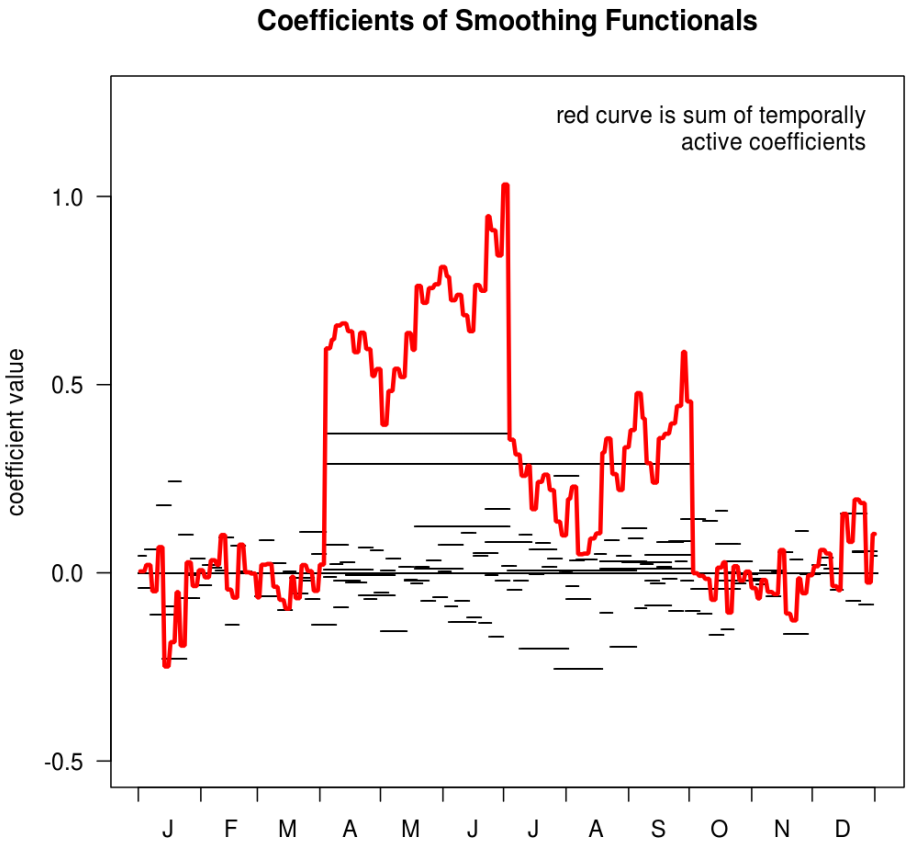

Figure 4: Fitted daily-precipitation smoothing coefficients

It appears to be the case that soybeans need relatively more water in the early months after planting. As summer progresses, this water requirement drops off before moving upward again as harvest approaches. In the off-season, average precipitation rates have a far smaller impact on crop yield.

Results

Table 1 in the Executive Summary reports the key results.

We also include, below, a plot of predicted yield versus true yield on the test set.

Figure 5: Predicted yield versus true yield

Figure 5 indicates that the model fits fairly well. The predictions are nicely distributed around the $y = x$ line, and there appear to be no wildly problematic outliers.

Further Work

Clearly, precipitation isn't the only covariate that effects soybean yields. Temperature is another field covariate that should also have a big impact on the harvest. It can be added to the model in a similar way as precipitation. However, other functionals may also be of interest.

With respect to temperature, big swings in daily temps could be a factor. This could be quantified with, say, a quadratic variation operator. Or, maybe, late spring freezes have a negative impact. To test this theory, one could include a functional that counts number of freeze days after planting.

Another variable is fertilizer levels, which can be introduced in a more controlled way. Fertilizer concentrations could be studied in this sparse functional context as well, perhaps by introducing a non-sparse parameter, like μ, in the objective function.

Of course, farmer insights are probably the best source of on-the-ground knowledge for generating additional hypotheses about crop yield. In many cases, these may be easily reduced to additional functionals that could be included in the feature vector.

References

[1] Jerome H. Friedman, Trevor Hastie, and Rob Tibshirani. Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software*, 33(1):1–22, 2 2010.