PPD 557 Project: Bean Farming

Mose Wintner

The scenario: It's 2012, and I have decided to leave the lucrative world of research mathematics to become a bean farmer. I've saved $\$150,000$ to get myself started. In addition, the government has offered me a five-year interest-free loan to do this in one of three counties in Washington state named for a powerful early American: any of Adams, Franklin, and Grant counties (really this is just where I could find abundant, recent data). I have found three seemingly suitable properties and I want to decide which property to purchase and move to based on a multiattribute utility calculation utilizing several criteria, including the size and price of the properties; a "cool factor" measure of natural beauty and proximity to a city, using data from a 1999 study by the USDA's economic division; price of planting beans per acre; whether or not I would have to take out a loan to purchase the property and begin planting; and a prediction of yield and resulting profit for the next five years based on bean crop yield data from the USDA.

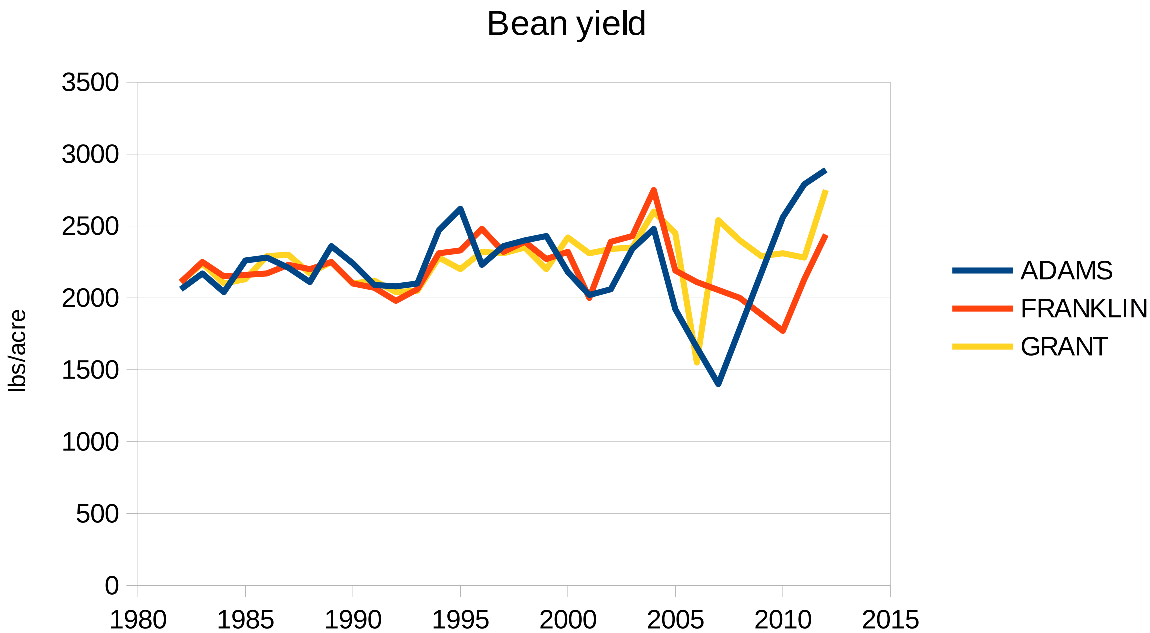
The data: I assume that the initial cost of starting a farm is uniform across counties and factored in to the cost of the property.

The USDA data I used to determine the natural beauty and "cool factor" is from a study by the USDA's Economic Research Division [1]. From the abstract:

> The natural amenities scale is a measure of the physical characteristics of a county area that enhance the location as a place to live. The scale was constructed by combining six measures of climate, topography, and water area that reflect environmental qualities most people prefer.

Included with the natural amenities scale are previously published measures of proximity to a city or metropolitan area and type of terrain, all of which are important to me and my wife. The study uses a holistic scale, so I used data from the study rather than the complex model they used for their rankings. I also supposed that the individual properties came with their own natural idiosyncracies in my estimations.

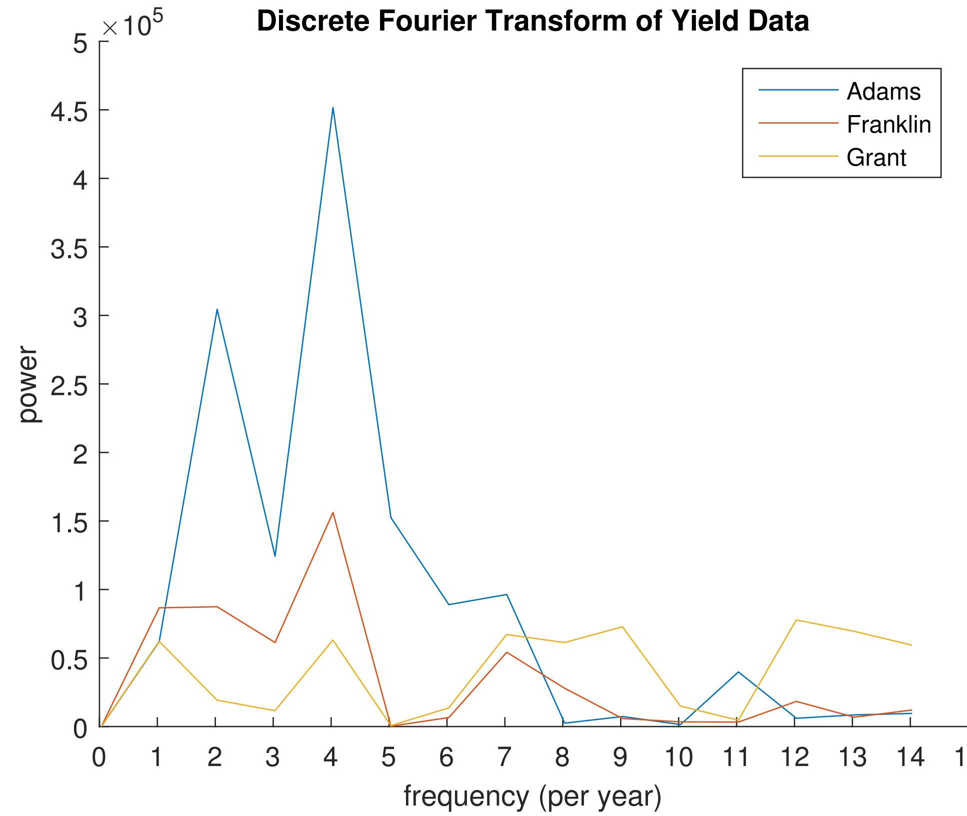
The agricultural data I used also come from the USDA, but I used yield data for bean crops from each of the three counties dating $1982 - 2012$, which was the largest range on crop yield data by county I could find for Washington [2].



The analysis: The bulk of my analysis went into predicting bean yield. First, I attempted a linear least squares/regression analysis on the (normalized) bean yield data, which for all three counties had a very small slope and R^2, suggesting no significant linear trend.

	slope	intercept	correlation	R^2
Adams	5.94	-9626	.173	.030
Franklin	1.85	-1482	.087	.008
Grant	9.00	-15702	.397	.157

Next, I wanted to test for years-long seasonal patterns in the data, so I ran a discrete Fourier transform on it in MATLAB. Since Fourier analysis with missing data is difficult, there were a few data points for which I had to estimate for missing yield data using linear interpolation. The Fourier analysis revealed that in Adams and Franklin counties, there were strong 4-year trends in the data. In Grant county, there were many relatively weak frequency components to the data and none too strong, although there was a *relatively* significant frequency component of four years. I'm not able to explain why, because all three counties share a border in Northern Washington and therefore shouldn't necessarily be subject to widely varying weather patterns, but this is what the data showed. Because of this, I decided that I would try to predict near-future bean yield using a combination of variants of the 4-year moving average and simulation, rather than attempt to subtract from the data frequency components corresponding to four-year patterns.



Before selecting a method to use to predict near-future bean crop yield, I ran error analyses on the four-year moving average and several variants for the known data. One of the variants was an 8-year moving average, which I figured should give results similar to and/or smoother than a 4-year moving average, though perhaps too smooth given there was no strong 8-year frequency component in the data. I also tried an exponentially smoothed 4-year moving average ($\alpha = 0.3$) and a four-year weighted moving average, with weight vector $(0.1, 0.2, 0.3, 0.4)$. The 4-year exponentially smoothed prediction barely edged out the simple prediction in minimizing total error. Note that I had to average over a smaller number of data points in some counties and prediction methods because of missing data.

Errors	MAD				Cumulative			
	Adams	Franklin	Grant	Total	Adams	Franklin	Grant	Total
4yr moving	258.9	183.7	146.9	589.4	35.6	15.7	30.9	82.2
4yr weighted	366.1	265.8	142.4	774.3	144.9	145.9	28.4	319.2
8yr moving	320.3	204.6	156.4	681.4	33.8	2.7	45.2	81.8
4yr exp. (α=0.3)	163.4	137.0	157.2	457.6	-17.3	32.3	25.4	40.4
simple	163.9	146.3	161.3	471.5	-12.2	25.8	21.3	34.9

I measured the sample mean and sample standard deviation for the bean yield by county for the known data. Then I verified via a χ^2-test (not included) that the yield by county was pretty much normally distributed and simulated yields for 2013-2017. Then I smoothed the data using the 4 year exponentially smoothed moving average.

Next, I attempted to predict the price per acre for 2013-2017 by using both linear and exponential regressions on the yearly price per pound for beans in Washington state [3]. Both regressions performed about the same, with R^2 just exceeding 0.5, suggesting some significance. However, after examining a graph of the data, I noticed that the price of beans appeared to grow exponentially beginning sometime in the mid-90s. An exponential regression on the price data since 1997 had a much stronger correlation coefficient of .7558; I used this to predict the price for the next five years.

I suspected that perhaps some of the same factors which affect price also affect yield, so I measured the correlation coefficient between price and yield for the three different counties. They revealed only very weak correlations, so my suspicion was ostensibly wrong, and I proceeded assuming they were independent. I couldn't find a history of the cost of growing beans per acre, so I simplified my assumptions by assuming the price data (per pound) had already factored in the cost of planting beans.

Finally comes the (multiplicative) multiattribute utility analysis. I chose profit to have a scaling constant of 0.5, which is easy to scale other attributes against. The property cost in itself doesn't have any bearing at all on my decision, so its scaling constant is 0. Rather only its effect on my profit margin and loan status is important.

Similarly, since the size of the property *in itself* doesn't matter much in comparison it has a much lower scaling constant of 0.1. Of course, it has a larger effect on the profit margin, since a larger property means more area on which to plant beans. (The numbers given represent the properties' farmable areas.) A larger property is generally nicer than a smaller one, but if the property is too large, it becomes difficult to manage. Therefore I modeled the utility function as an inverted parabola with vertex at an optimal size of 30 acres, i.e. as $1 - (\frac{x}{30} - 1)^2$.

Now, to consider the risk-free five year loan. If my projected profit plus my $\$150,000$ savings were not to meet the cost of the property, I would have to take out a bank loan at an interest rate of 20%, plus added stress. The interest alone would give the loan amount's scaling constant a value of 0.6, but the stress of operating a farm under a loan I estimate to be significant, at an additional 0.1. The stress should be considered to increase exponentially with the amount of the loan, in part because the amount of time required to pay off the loan increases exponentially with the amount. The exponential utility function should therefore have the same instantaneous rate of change at $x = 0$ as the profit function, which is simply the slope of the profit utility function, or the range of the three five-year projected profits. Of course, the loan should actually count as *negative* utility, since it represents debt, so it will have the form $1 - e^{x/\text{range(profit)}}$.

Finally, since the natural beauty of the property and county as well as the population of the surrounding areas is comparably important to happiness as money, but not quite as important as avoiding the stress of a loan, it gets a scaling constant of 0.6. All unspecified utility functions are risk-neutral. Via MATLAB's `solve()` and `root()`, the master scaling constant came out to $K = -0.9227$. Below is a list of the results coming from one typical simulation; the bottom is the result of the multiattribute utility calculation, which I programmed into my Excel worksheet.

Attribute		Adams		Franklin		Grant	
Name	k_c	Value	Utility	Value	Utility	Value	Utility
Property cost	0	$387,400	0	$262,100	0	$345,600	0
Size (acres)	0.1	39	.9100	23	.9456	35	.9722
5-yr Profit (proj.)	0.5	$204,710.97	1	$113,684.98	0	$185,683.62	.7910
Loan+stress	0.7	$32,689.03	-.4321	$0.00	0	$9,916.38	-.1151
Coolness	0.6	0.243	.226	0.857	.857	0.561	.561

Adams:	0.49180
Franklin:	0.56389
Grant:	0.61971

The simulation tends to prefer the property in Grant county to the Franklin county property in a not-terribly-large majority of the simulations, but when it favors the property in Grant county, it tends to do so by a large margin. Therefore, it would be wisest to choose the property located there.

References

[1] David McGranahan. Natural amenities scale (including the 6 components) for U.S. counties. Technical report, United States Department of Agriculture Economic Research Service, September 1999. http://www.ers.usda.gov/data-products/natural-amenities-scale.aspx.

[2] United States Department of Agriculture National Agricultural Statistics Service. http://quickstats.nass.usda.gov/.

[3] United States Department of Agriculture National Agricultural Statistics Service. http://quickstats.nass.usda.gov/.