

A Better Approach to Calculate Approved Yield-Indexing

January 3, 2006

**By
Dr. Myles Watts
Dr. Joe Atwood**



Executive Summary

Producers are concerned that their ability to obtain meaningful crop insurance is eroded after suffering poor yielding years. RMA (USDA) bases the yield guarantee used to trigger indemnity payments on a producer's actual production history (unless the producer has less than four years of yield history). RMA uses the average yield, calculated from four to ten years of farmer provided yield history, to develop the yield guarantee. Because four to ten years is a small sample, unusually low years will dramatically reduce the average yield. Reducing the average reduces the guarantee and the likelihood and level of indemnity payments.

RMA has attempted to address the issue by ad hoc cups or plugs. If the average declines by more than a prespecified percentage, the decline is cupped or limited when calculating the yield guarantee. A plug can be used to replace a low yield if the yield is sufficiently low. Both plug and cups are ad hoc, lacking statistical and actuarial validity. Furthermore it is nearly impossible to incorporate these ad hoc measures into the rates. (The primary problem is that these approaches nearly maintain the yield guarantee in bushels but the actual average yield has declined, so the effective coverage level increases possibly to 100% or more.)

Indexing is a statistically and actuarially sounder method of stabilizing the basis for the yield guarantee. The indexed yield also is a more accurate estimate of the expected yield than the simple four to ten year average. Indexing incorporates both the producer's four to ten years of data and longer term regional (county) yield data. Regional data is often available for 50 or more years. The longer data set provides additional information on the probability of unusually low or high yields. Incorporating this added information reduces the variability of the yield guarantee through time, making the yield guarantee less sensitive to a recent unusually bad year. (Of course, the yield guarantee is also less susceptible to a recent good year.)

An additional problem is that the simple average does not appropriately incorporate technology increases. If technology is increasing yields about 1% per year, and ten years of yield history are used, at the end of the ten years the simple average is roughly 5% below a more accurate estimate of the expected yield. (If the producer buys 75% coverage a more accurate estimate of the effective coverage is 75% times .95, or about 71% coverage.) The indexed yield incorporates technology in estimating expected yields.

More Detailed Discussion

The approved yield is multiplied by the coverage level to determine the yield guarantee, which also serves as the indemnity trigger and the basis for calculating the premiums.

The objective should be to develop an approved yield, which is as accurate as possible point estimate of the expected yield (and is administratively feasible).

(“Expected value” is the “best” point estimate of a random variable, which is usually a weighted average of all possible values that can be taken by the variable. In this case it would be the “best” point estimate of the yield in the year of the insurance contract.)

The commonly used indicators of accuracy are bias and efficiency. If the method used to calculate the approved yield consistently (but not necessarily always) either over or underestimates the true expected value of the parameter, the approved yield would be a biased estimator of the expected value. (More correctly, the estimator of the expected value is unbiased if the mean value of the estimator, calculated from a large number of samples, equals the expected value.) Efficiency is the volatility or the variability of the estimated value. In this case alternative methods of calculating the approved yield may be more sensitive to the “yield sample” drawn (provided actual production history) than other methods of calculating approved yields and therefore be more inefficient.

Following standard statistical criteria of bias and efficiency, the current method of calculating approved yields will be compared to an alternative method referred to as indexed yield.

The simple average currently used by RMA to compute approved yields is based upon four to ten years of producer supplied actual production history. Figures 1 and 2 are historic U.S. wheat and corn yields illustrating the persistent yield increases, commonly attributed to technological progress (broadly defined). Obviously if yields over some period are averaged and used as the approved yield, generally the approved yield will be less than the future yield and the approved yield is a biased predictor of the future expected yield. Furthermore, calculating an average from only four to ten years of data results in an approved yield with substantial variability. The bias and inefficiency of the simple average should encourage the pursuit of a better method to estimate the approved yield.

An indexed yield, the proposed alternative method to estimate approved yield, eliminates the bias and increases the efficiency of the estimate. Longer-term regional NASS data, often with 80 or so years of observations, is used along with the actual production history to develop a more accurate indexed yield. It is this longer regional data set used in calculating the indexed yield that increases the accuracy of the approved yield. The indexed yield calculation begins by statistically estimating a trend line from the long-term regional data set and then forecasting the expected regional yield for the year of the insurance contract. Using the four to ten years of available farm data, both the farm and regional average yield is calculated. The difference between these averages is added (subtracted) to (from) the expected regional yield generating the indexed yield. (Later this linear process is slightly modified to a multiplicative form. However the linear form is analytically more tractable.)

To facilitate later exposition about the accuracy of both the simple average and indexed yield method of estimating the approved yield, let:

T_I = number of farm yield observations

T_r = number of regional yield observations

b = annual increase (slope) of yield trend line

σ^2 = farm level variance

σ_e^2 = variance of trend line residuals or errors

$\sigma_f^2 = \sigma^2 - \sigma_e^2$ = farm level variance beyond regional variance

Bias

The indexed yield predicted from a linear regression equation has no bias. Using non-linear regression to estimate the trend may or may not be biased, however the linear approximation is unbiased and so the bias is considered negligible. The simple average is biased because of technology and the bias is

$$-b \frac{T_I + 1}{2} .$$

Efficiency

The variance of the simple average is well known as

$$\sigma_Y^2 = \frac{\sigma^2}{T_I} .$$

The variance of the indexed yield has two components (which can be shown to be orthogonal and so are additive) – the variance of the regional expected yield (σ_r^2) and the variance of the difference between the farm level and regional average yield, σ_f^2 .

However both of these are well known if the trend line is linear (where time, t , is the only explanatory variable). The year of the insurance contract/time of the needed approved yield is $T_r + 1$, and \bar{t} is the average of t .

If the trend line is nonlinear, then there is no known closed formed solution to the variance of the expected forecasted value. However using the linear counterpart should sufficiently illustrate the focus of this discussion. The variance of the expected regional yield in year $T_r + 1$ is

$$\sigma_{\hat{f}}^2 = \sigma_e^2 \left[\frac{1}{T_r} + \frac{(T_r + 1 - \bar{t})^2}{\sum_{t=1}^{T_r} (t - \bar{t})^2} \right] = \sigma_e^2 \cdot \frac{4 \cdot T_r + 2}{T_r (T_r - 1)}$$

and the variance of the difference of the averages is

$$\sigma_{\bar{f}}^2 = \frac{\sigma_f^2}{T_l} .$$

The variance of the indexed yield is

$$\sigma_{Indexed}^2 = \sigma_{\hat{f}}^2 + \sigma_{\bar{f}}^2 = \sigma_e^2 \cdot \frac{4 \cdot T_r + 2}{T_r (T_r - 1)} + \frac{\sigma_f^2}{T_l} .$$

In the earlier definitions $\sigma_f^2 = \sigma^2 - \sigma_e^2$. Statistically it can be shown that $\sigma^2 > \sigma_f^2 + \sigma_e^2$ if $b \neq 0$, since the regression line reduced the variance that is not captured by either σ_f^2 or σ_e^2 . It turns out that this uncaptured component is small (since T_l is small) relative to σ_f^2 and σ_e^2 , and so is ignored here.

Obviously the efficiency gain in using the indexed approach is due to the large number of observations available from the regional data. On the other hand, the use of regression increases the variance since at least two regression parameters must be estimated.

Of concern is the relationship between $\sigma_{\bar{y}}^2$ and $\sigma_{Indexed}^2$ and how that relationship is affected by the number of observations available at the farm and regional level. Setting

$$\sigma_{\bar{y}}^2 = \sigma_{Indexed}^2$$

and solving for T_l while recognizing $\sigma^2 \cong \sigma_f^2 + \sigma_e^2$ yields,

$$T_l = \frac{T_r (T_r - 1)}{4 \cdot T_r + 2} .$$

If the above equation is satisfied, then the variance of the simple average and the indexed yield is the same. If the right hand side is larger (caused by a relatively large value for T_r , or a longer time series of available regional yields), the indexed yield has lower variance. From the above equation, the indexed yield will provide a more efficient estimate of the approved yield, if the length of the regional data series is greater than approximately four times the length of the farm data series.

To illustrate the reduction in variance assume the following situation. Let $\sigma^2 = 900$, $\sigma_f^2 = 450$, and $\sigma_e^2 = 450$ (farm variance twice the regional variance). The resulting simple average and indexed yield variances are in Table 1 for various time series lengths of farm and regional yield data. Substantial efficiencies are gained by using the indexed yield and the longer regional long-term data set, particularly when the lengthy regional data sets are available.

Table 1. Efficiency of the Simple Average and Indexed Yield

Length of Farm Data (years)	Simple Average Variance	Length of Regional Time Series (years)						
		20	30	40	50	60	70	80
		Indexed Yield Variance						
4	225	210	176	159	150	143	139	135
6	150	172	138	122	112	106	101	98
8	113	154	120	103	94	87	83	79
10	90	142	108	92	82	76	71	68

An alternative approach to evaluate the accuracy of the simple average and indexed yield estimates is to calculate the probability of the approved yield being lower than a selected value. The previous example continues to be used with the added assumption that the increase in yield is 1 ($b=1$), the “true” expected yield is 100 (in the contract year), and the regional trend line residuals and farm yields are normally distributed. Table 2 presents the probability of an approved yield being less than a prespecified value.

Table 2. Probability of an Approved Yield Less Than a Specified Value Based on 20, 50, or 80 Years of Regional Yield Data and a Yield Trend

	Years of Producer History			
	4	6	8	10
Expected Simple Average Yield	97.5	96.5	95.5	94.5
Expected Indexed Yield	100	100	100	100
Probability of Approved Yield < 75				
Simple Average	0.067	0.040	0.027	0.020
Indexed Yield 20 years	0.042	0.028	0.022	0.018
Indexed Yield 50 years	0.021	0.009	0.005	0.003
Indexed Yield 80 years	0.016	0.006	0.002	0.001
Probability of Approved Yield < 80				
Simple Average	0.122	0.089	0.072	0.063
Indexed Yield 20 years	0.084	0.064	0.054	0.047
Indexed Yield 50 years	0.051	0.029	0.020	0.014
Indexed Yield 80 years	0.043	0.022	0.012	0.008
Probability of Approved Yield < 85				
Simple Average	0.202	0.174	0.161	0.158
Indexed Yield 20 years	0.150	0.126	0.113	0.104
Indexed Yield 50 years	0.110	0.078	0.061	0.049
Indexed Yield 80 years	0.098	0.065	0.046	0.034
Probability of Approved Yield < 90				
Simple Average	0.309	0.298	0.302	0.318
Indexed Yield 20 years	0.245	0.223	0.210	0.201
Indexed Yield 50 years	0.207	0.172	0.151	0.135
Indexed Yield 80 years	0.195	0.156	0.130	0.113
Probability of Approved Yield < 95				
Simple Average	0.434	0.451	0.481	0.521
Indexed Yield 20 years	0.365	0.352	0.344	0.337
Indexed Yield 50 years	0.342	0.318	0.303	0.290
Indexed Yield 80 years	0.333	0.307	0.287	0.272
Probability of Approved Yield < 100				
Simple Average	0.566	0.612	0.664	0.719
Indexed Yield 20 years	0.500	0.500	0.500	0.500
Indexed Yield 50 years	0.500	0.500	0.500	0.500
Indexed Yield 80 years	0.500	0.500	0.500	0.500

Table 2. Probability of an Approved Yield Less Than a Specified Value Based on 20, 50, or 80 Years of Regional Yield Data and a Yield Trend (Continued)

	Probability of Approved Yield < 105			
Simple Average	0.691	0.756	0.815	0.860
Indexed Yield 20 years	0.635	0.648	0.656	0.663
Indexed Yield 50 years	0.658	0.682	0.697	0.710
Indexed Yield 80 years	0.667	0.693	0.713	0.728
	Probability of Approved Yield < 110			
Simple Average	0.798	0.865	0.914	0.949
Indexed Yield 20 years	0.755	0.777	0.790	0.799
Indexed Yield 50 years	0.793	0.828	0.849	0.865
Indexed Yield 80 years	0.805	0.844	0.870	0.887
	Probability of Approved Yield < 115			
Simple Average	0.878	0.935	0.967	0.985
Indexed Yield 20 years	0.850	0.874	0.887	0.896
Indexed Yield 50 years	0.890	0.922	0.939	0.951
Indexed Yield 80 years	0.902	0.935	0.954	0.966
	Probability of Approved Yield < 120			
Simple Average	0.933	0.972	0.990	0.996
Indexed Yield 20 years	0.916	0.936	0.946	0.953
Indexed Yield 50 years	0.949	0.971	0.980	0.986
Indexed Yield 80 years	0.957	0.978	0.988	0.992
	Probability of Approved Yield < 125			
Simple Average	0.967	0.990	0.997	0.999
Indexed Yield 20 years	0.958	0.972	0.978	0.982
Indexed Yield 50 years	0.979	0.991	0.995	0.997
Indexed Yield 80 years	0.984	0.994	0.998	0.999

In Table 2, the expected simple average yield is the “true” expected yield reduced by the bias calculated from the earlier specified formula. The probability values in Table 2 were calculated using the normal distribution. Two example calculations may be illustrative. Note the value of .067, which is the probability of the approved yield being less than 75 if the simple average and four years of producer data is used to calculate the approved yield. Further note that the expected simple average is 97.5 and that the variance of the simple average is 225 (from Table 1). The standard deviation of the simple average is $\sqrt{225} = 15$ and so the usual Z value can be calculated

$$Z_s = \left| \frac{75 - 97.5}{15} \right| = 1.5 \quad .$$

Referring to the cumulative normal distribution table and a z = 1.5 results in a probability of .067. Similarly using an indexed yield approach, the probability of an approved yield

value less than 75 using four years of producer data can be calculated. The indexed expected yield is 100, the variance of indexed yield is 135 (Table 1), so the Z value is

$$Z_{indexed} = \left| \frac{75-100}{\sqrt{135}} \right| = 2.15$$

or a probability of .016.

If the approved yield is to be an efficient and unbiased estimation of the producer's actual expected yield, it needs to be as stable as possible with the distribution concentrated as much around the "true" expected value of 100 as possible. Therefore the probabilities associated with approved yield outcomes in Table 2 need to be as low as possible for approved yield outcomes below 100. For example in the case of the approved yield outcome of 75 or less, the probability of this outcome using the simple average .067 is less desirable than .016 probability if the indexed yield is used.

Analogously the probability associated with being below the observed approved yields above 100 needs to be as high as possible. Notice that one minus the probabilities in Table 2 are the probabilities of being above the observed approved yield value. It is desirable to minimize the probability of observing approved yields above the true expected value of 100. However minimizing the probability of being above an observed approved yield is identical to maximizing the probability of being below an observed value.

In Table 2, when the observed approved yield is below 100, the indexed yield substantially reduces the probability of observing low approved yields. The combination of eliminating the bias and increased efficiency results in this large reduction in the probability of a low observed yield. In Table 2, the results associated with observed approved yields greater than 100 are more mixed. This outcome is not surprising since the expected simple average yield is lower but the variance is higher than the indexed yield. The probabilities above an observed approved yield of 110 are quite similar. These results are robust across different yield distributions and example assumptions, particularly if yields are trending upward.

A critical component in calculating indexed yields is the estimation of a yield trend lines. Our experience supports the following approach. The yield equation estimated as a function of time is some form of (α_i are parameters)

$$\hat{Y}_t = \alpha_0 + \frac{\alpha_1 t^{\alpha_2}}{\alpha_3 + t^{\alpha_4}} .$$

The model form is highly flexible and allows the data to "speak". However, the data should not be allowed "to speak too much" so as to fit the trend line not only to technological change but also to short term phenomena such as drought. The "over fitting" is guarded by the following approach. The simplest form of the model is always chosen over more complex forms, and complex forms chosen only with compelling statistical evidence or other information.

The simplest form of the base model is where $\alpha_1 = 0$ and so $\hat{Y}_t = \alpha_0$. In this case, there is no upward or downward trend, the expected regional yield is constant through time and α_0 is the mean yield. If $\alpha_2 = 1$, $\alpha_3 = 0$, and $\alpha_4 = 0$, then the base model becomes $\hat{Y} = \alpha_0 + \alpha_1 t$ or is linear (and $\alpha_1 = b$ as defined earlier) as illustrated in Figure 3. Figure 3 also illustrates situations in which $\alpha_2 \neq 1$.

If $\alpha_2 = \alpha_4$ then $\hat{Y}_t = \alpha_0 + \frac{\alpha_1 t^{\alpha_2}}{\alpha_3 + t^{\alpha_2}}$ and (assuming $\alpha_i > 0$) \hat{Y}_t has a sigmoid form as

demonstrated in Figure 4. The predicted yield approaches the asymptote (maximum) of $\alpha_0 + \alpha_1$ as t becomes large. The level of α_2 affects the steepness of the yield trend. Figure 5 demonstrates that α_3 affects the length of the initial “flat” portion of the yield trend as well as the steepness of the trend line.

Figure 6 demonstrates the effect of $\alpha_2 \neq \alpha_4$. The original models with all five parameters can take on various shapes including both declining yields and permanently upward sloping values. A variety of these shapes are illustrated in Figure 6.

From a statistical perspective, testing restrictions helps choose the model form. From a practical perspective the following four models are estimated by regression (OLS).

Model	Form
A	$Y_t = \alpha_0$
B	$Y_t = \alpha_0 + \alpha_1 t$
C	$Y_t = \alpha_0 + \frac{\alpha_1 t^{\alpha_2}}{\alpha_3 + t^{\alpha_2}}$
D	$Y_t = \alpha_0 + \frac{\alpha_1 t^{\alpha_2}}{\alpha_3 + t^{\alpha_4}}$

Note that Model A is a restricted form of Models B, C, and D where $\alpha_1 = 0$. A test for restrictions (with OLS estimation) is an F test (we note that F tests are only a linear approximation in non linear models) where

$$F[J, T_r - k] = \frac{R^2 - R_r^2}{1 - R^2} \frac{T_r - k}{J}$$

where: R^2 = unadjusted R^2 for the unrestricted model
 R_r^2 = unadjusted R^2 for the restricted model
 T_r = number of regional observations
 k = number of parameter - unrestricted
 J = number of restrictions

If the restriction ($\alpha_1 = 0$) is insignificant, Model A is chosen. If the restriction is statistically significant, then Model B, the linear model is considered. If Model D (unrestricted) is restricted with two restrictions $\alpha_3 = 0$ and $\alpha_2 - \alpha_4 = 1$, then the model becomes the linear Model B (restricted). If the restrictions are insignificant, Model B is chosen. If the restrictions are significant, then Model C and D are compared by testing the restriction of $\alpha_2 = \alpha_4$ and selecting Model C or D accordingly. Our experience is that either the sigmoid (Model C) or the linear (Model B) will almost always be chosen.

Before the final selection of a model, the predicted and actual yield observations are graphed and inspected for previous undetected anomalies and other concerns.

As an illustration, all wheat regional yields and the predicted values from the trend line for Petroleum County, Montana, are presented in Figure 7. The estimated trend line yield equation is

$$\hat{Y}_t = 8.487 + \frac{14.106t^{9.656}}{988392 + t^{9.656}} .$$

Note that Models C and D are intrinsically nonlinear in the parameters. As such the parameters must be estimated with a search routine to find the best parameters (those parameter values which minimize the sum of the squared regression errors). Such routines for nonlinear estimation are notoriously fickle. However since the trend lines are nearly always monotonic, the estimation is less problematic. We have been most pleased with using Shazam to estimate the regional yield trend line equation (although we have experiences with several statistical packages with non linear estimation options). Secondly the independent variable, t, should be scaled so t does not become large (so as not to exceed internal format limits of the statistical package).

The wheat yield trend line for Petroleum County was estimated for the years 1925 through 2004. These year numbers are too large and so the time variable in the Petroleum trend line equation was transformed by

$$t = \frac{year - 1918}{10} .$$

Using the transformation equation and the earlier Petroleum yield trend line equation, the expected regional indexed yield is 22.6 for 2005.

Before completing the approved yield example in Petroleum County, there is an adjustment to the indexed yield method we have found useful. The indexed yield as discussed earlier is calculated as

$$\text{Indexed Farm Yield} = \text{Indexed Regional Yield} + (\text{Farm Average} - \text{Regional Average})$$

Table 3 presents recent regional (county) yield data for Petroleum County as well as hypothetical farm yields data. Recently the wheat yields have been low and have depressed the approved yield as demonstrated in Table 4. Table 4 demonstrates the volatility of the simple average and the stability of the indexed yield.

Table 3. Petroleum County Farm and County Wheat Yield Data

Year	Farm Yield (Hypothetical)	Regional Yield
1995	30.00	20.50
1996	35.00	26.00
1997	40.00	32.80
1998	30.00	36.30
1999	30.00	28.60
2000	15.00	9.10
2001	0.00	7.50
2002	10.00	5.60
2003	26.00	21.40
2004	25.00	21.40

Table 4. Petroleum County Approved Wheat Yields Calculated by a Simple Farm Average and Indexed Yield

Number Of Years*	Farm Average	Regional Average	Difference of Averages	Indexed Yield
4	15.25	13.98	1.28	23.88
6	17.67	15.60	2.07	24.67
8	22.00	20.34	1.66	24.26
10	24.10	20.92	3.18	25.78

*Most Recent Years

The use of the indexed yield for the approved yield reduces (but does not eliminate) “strategic” reporting. “Strategic” reporting of historical production history occurs when producers fail to report low yield years and thereby avoid depressing their approved yield and indemnity trigger. However if the regional yield is also low, there is no reason to avoid reporting low yields if the approved yield is based upon the indexed yield.

A useful adjustment to the procedures discussed previously is to calculate the approved yield as a proportion rather than as a difference, or as

$$\text{Indexed Yield} = \text{Expected Regional Yield} \times (\text{Farm Average} / \text{Regional Average})$$

The adjustment entirely precludes the highly unusual possibility of a negative indexed yield.

Figure 1. Wheat All, United States 1919-2005

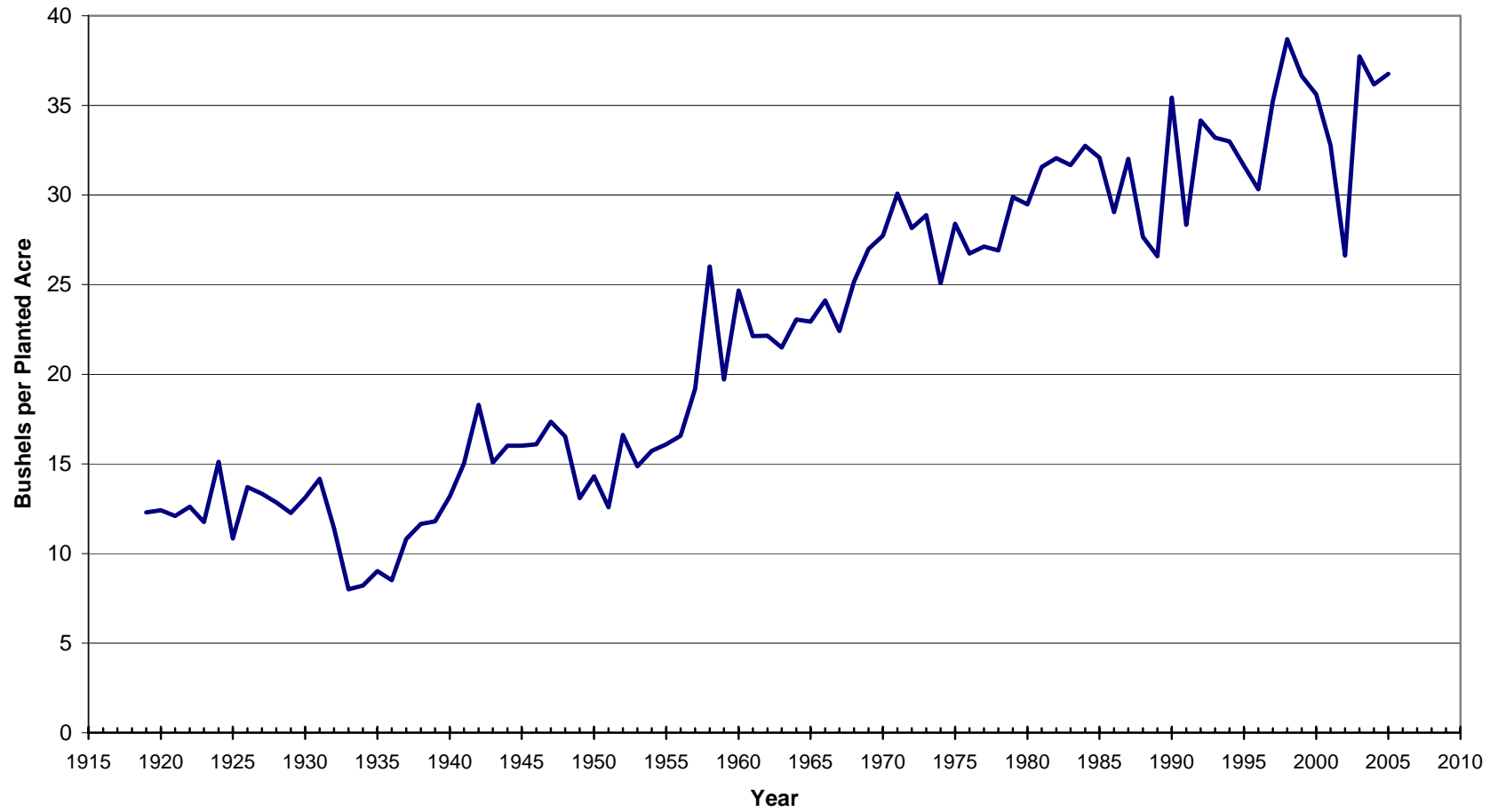


Figure 2. Corn for Grain, United States 1926-2005

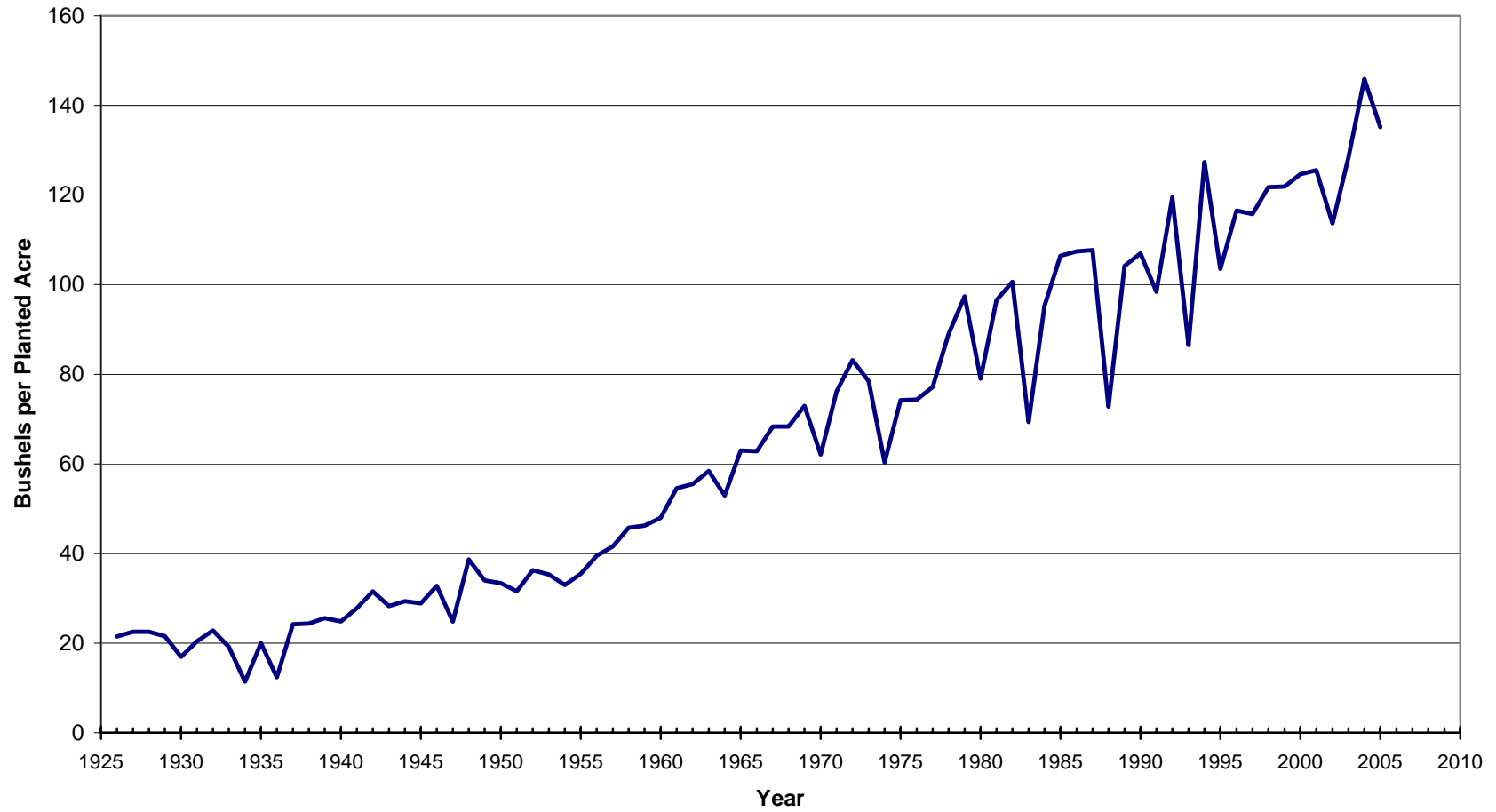


Figure 3. Illustration of the Effect of α_2

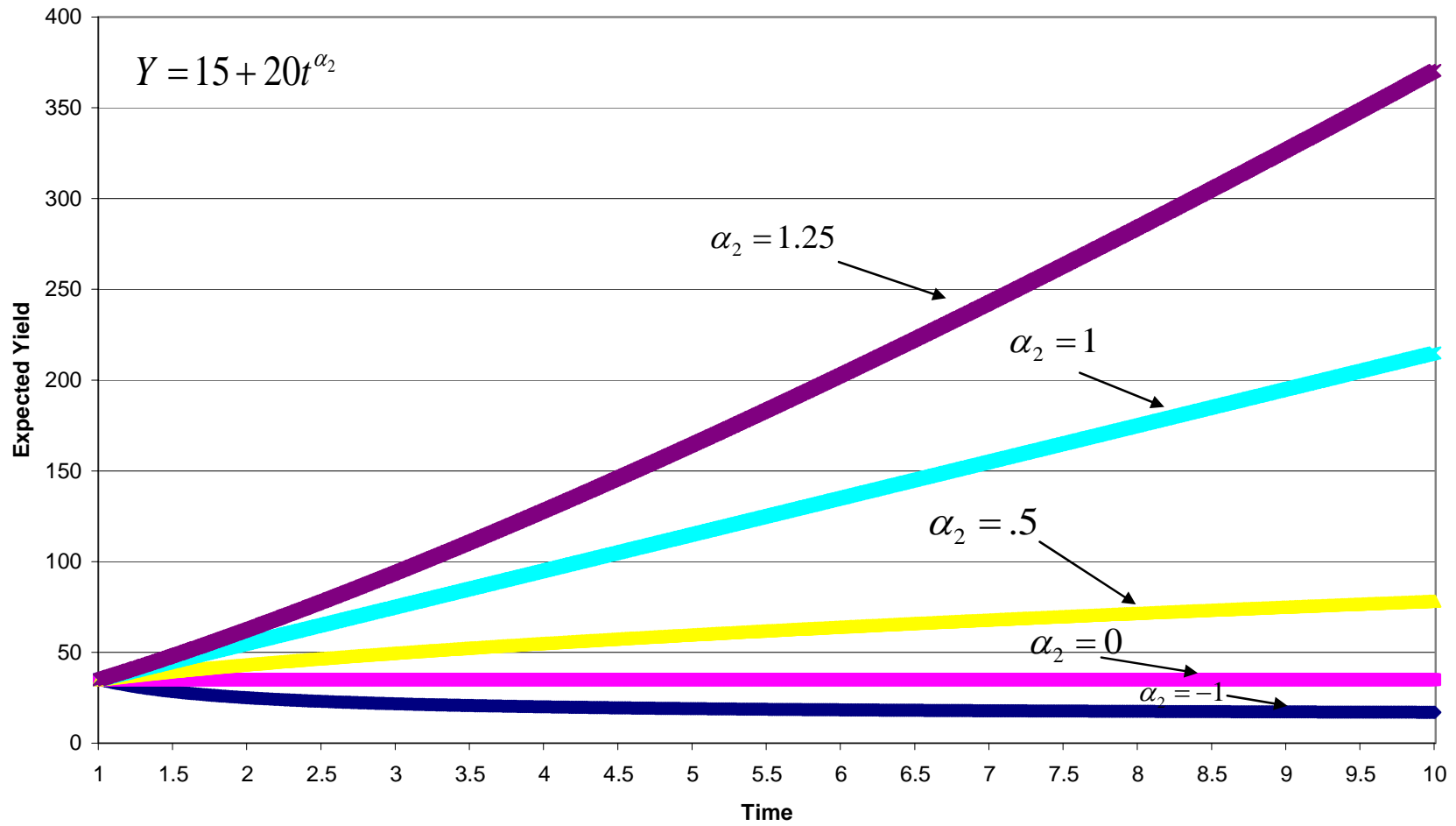


Figure 4. Illustration of the Effect of α_2

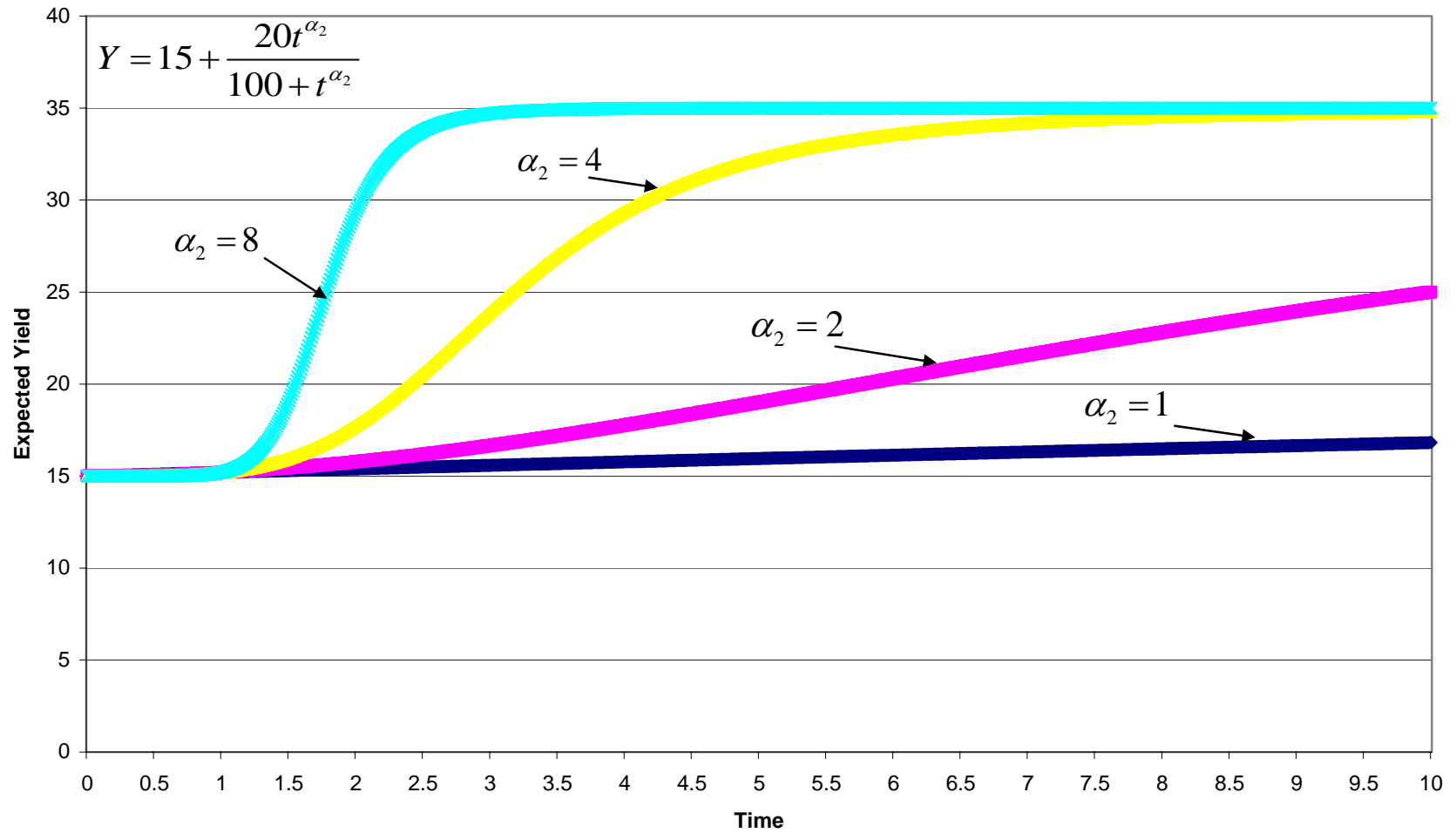


Figure 5. Illustration of the Effect of α_3

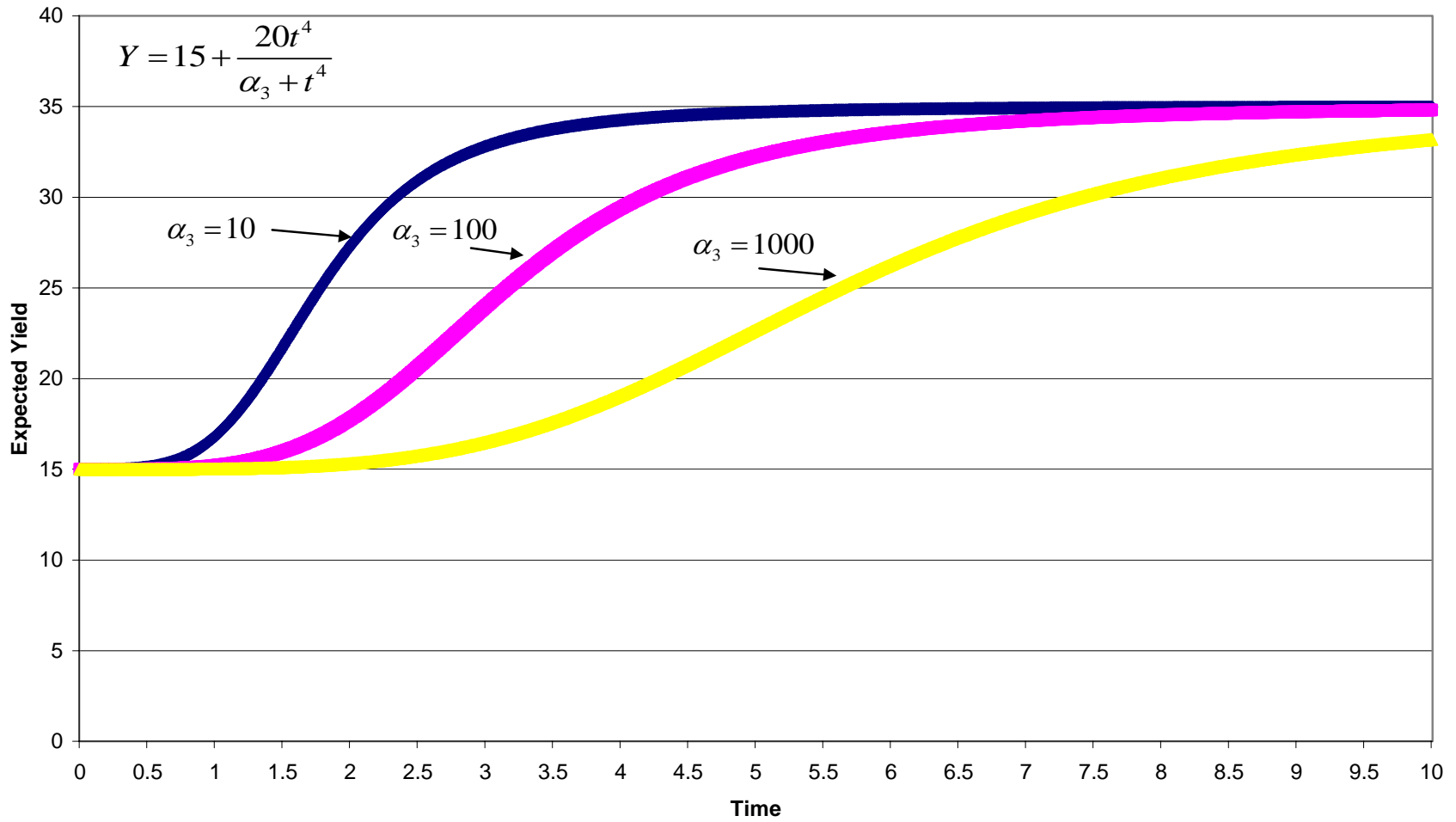


Figure 6. Illustration of the Effect of α_4

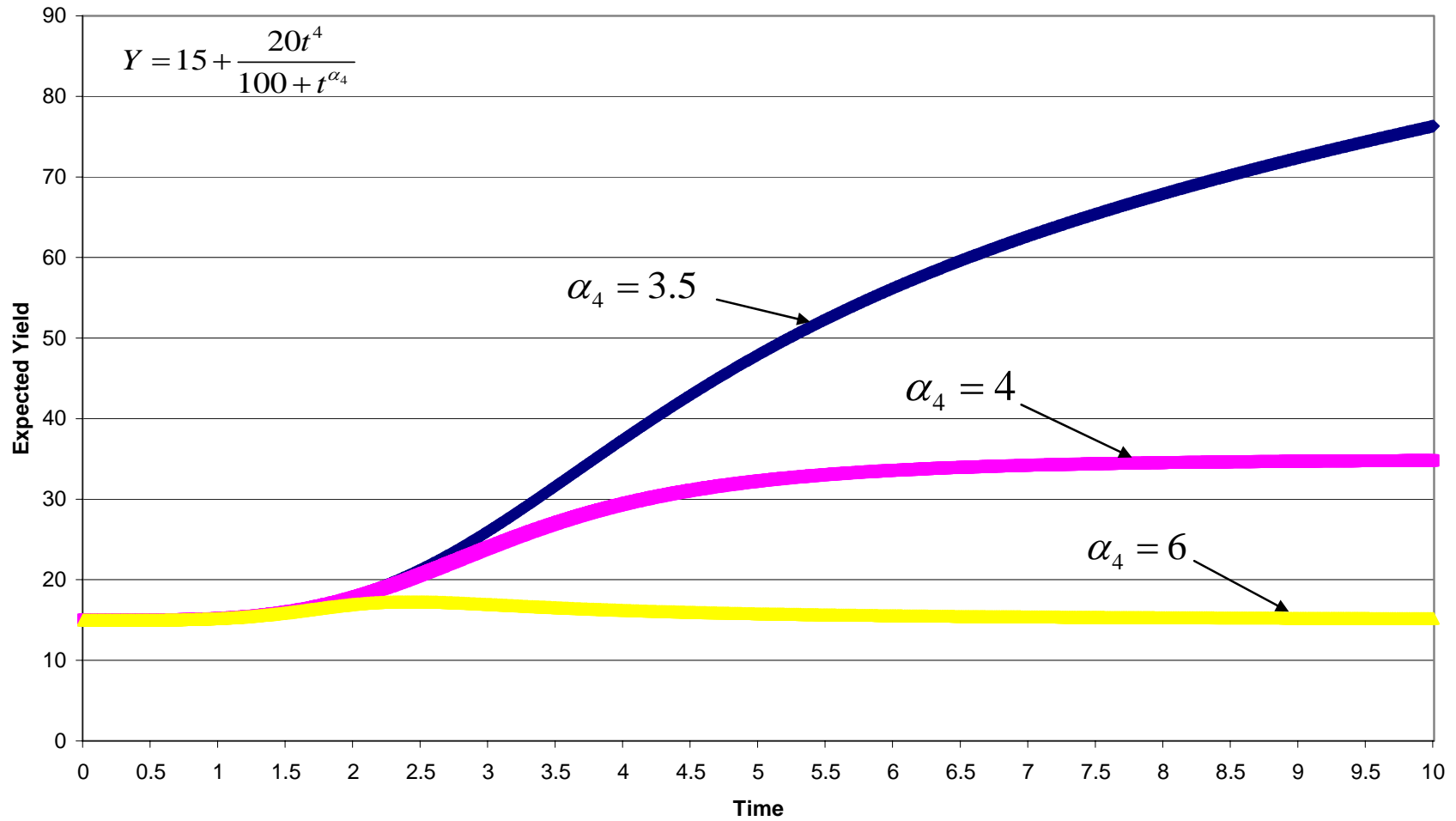


Figure 7. Wheat All, Petroleum County, Montana 1925-2004

