Journal of Agricultural and Resource Economics 25(2):411–431

Zone-Based Group Risk Insurance

H. Holly Wang

Current county-based group crop insurance, i.e., Group Risk Plan (GRP), is not an effective risk-reducing tool in counties where natural conditions are different across the area. Using only the historical yield information, a statistical approach is developed to group farmers by their yield similarity rather than linking them based on their association with a particular county. The cases of Washington State wheat farms and Iowa corn farms are the focus of this investigation. Sub-county or cross-county zones (clusters) are identified, and each farm is classified into a cluster where individual farm identification remains unknown. To improve risk-management and cost effectiveness of the crop insurance instrument, we propose implementation of zone-based GRP as a substitute for county-based GRP, where homogeneous zones rather than county boundaries are used for indemnifying yield.

Key words: classification, cluster, crop insurance, GRP, mixture distribution, yield risk, zone-based GRP

Introduction

Recent changes in the farming environment, including the elimination of government price support programs, increased global competition and price variation, and tightened environmental and natural resource concerns and policies, have exposed U.S. farmers to higher risks. Purchasing insurance is an effective way of reducing farmers' income risks.

The U.S. Department of Agriculture's (USDA's) Risk Management Agency (RMA), previously the Federal Crop Insurance Corporation (FCIC), has provided two types of agricultural insurance: traditional yield insurance programs and new revenue insurance programs. Indemnities for both programs can be based either on realized individual farm yield or on county average yield. Therefore, we categorize these designs here as individual-based versus group-based insurance programs.

Some forms of Multiple Peril Crop Insurance (MPCI) have been available since 1938, and a revised form of MPCI was instituted in 1980 for most major crops in the U.S. The current version of MPCI, the Actual Production History (APH) program, is an individual-based yield insurance, i.e., if a grower's own farm yield falls below the preselected coverage level, the grower will be paid the difference. However, moral hazard, adverse selection,[1] and high administration costs have prevented FCIC from providing APH at

Holly Wang is assistant professor, Department of Agricultural Economics, Washington State University, Pullman. The author gratefully acknowledges the helpful comments of Walter Butcher and two anonymous journal reviewers. This research is supported by the Agricultural Research Center, Washington State University.

[1] "Moral hazard" describes a circumstance where farmers alter their management plan after purchasing MPCI in order to save their production cost and claim more indemnity. "Adverse selection" occurs when only farmers with high risk tend to buy insurance. Both may cost insurers losses when the insurance premium is based on actual average risk.

a low cost to the government. The government paid $4.2 billion to support this program between 1981 and 1990 (U.S. General Accounting Office).

The alternative to APH is to base indemnity on the average yield of an area (Miranda). In any particular year, an insured grower will receive an indemnity payment only when the area average yield is lower than his/her preselected coverage level. Under this insurance, moral hazard and adverse selection have no basis, and the administrative cost is greatly reduced. RMA is currently providing a county-based crop insurance program, Group Risk Plan (GRP), for a limited number of crops to farmers in certain areas in an attempt to reduce the agency's financial losses.

The risk-management effectiveness of group-based insurance to a particular farmer depends heavily on the correlation between the individual farm's yield and the group average yield (Miranda; Wang et al.). Generally, yields are correlated across space, but only when all farmers follow similar production practices and when natural conditions (such as precipitation and soil quality) are homogeneous for the entire county will every farm's yield be highly correlated to the average county yield.

Wang et al. have estimated that the yield correlation between typical farms in southwest Iowa and their county average is about 0.8, when the area considered has relatively homogeneous natural conditions. Typically, geographical and natural conditions vary from one area to another within a county, which results in different farming practices and different yields even in the same county. An example is where a portion of a county is hilly and loses moisture to runoff and drainage, while the remainder of the county is flat and receives moisture from the hilly lands. In a dry year, the hilly fields generate poor yields, yet the flat fields may produce good yields. In a wet year, the opposite will be true, but county average yield may be nearly the same in both wet and dry years. In such cases, the whole-county-based GRP cannot effectively protect farmers from low yield.

Much of the current literature on area crop insurance focuses on improving the risk-reducing effectiveness of the existing GRP program and/or reducing the government's costs (e.g., Wang et al.; Skees, Black, and Barnett; Williams et al.; Miranda). Although it has been stated that a reduction in area size will improve the risk-reducing effectiveness, no alternative has been found to the county boundaries of the GRP. Finding such an alternative is the thrust of this research.

Dividing a heterogeneous county into smaller, more homogeneous zones on which to base GRP average yields can help improve the effectiveness. In such a case, a sub-county zone-based GRP (ZGRP) will be more effective for farmers than the county-based GRP without losing the advantages over APH. In contrast, some Corn Belt counties are small, homogeneous, and have natural conditions and farming practices similar to adjacent counties. In that case, the ZGRP insurance zone can be expanded across counties to further reduce administration costs without losing much effectiveness.

There are two essential difficulties in the design, evaluation, and implementation of ZGRP: (*a*) identifying the different zones, each of which must consist of farms with homogeneous farming conditions, and (*b*) classifying/assigning each farm into an appropriate zone. These two problems are complicated by the fact that farm identifications are removed from databases available to the public. For example, RMA keeps yield records of individual farms in each county for 10 years, but to protect the confidentiality of farmers, farm identifications (such as names, addresses, etc.) are unavailable. In this investigation, we consider a model-based approach to these two problems when each farm's yields are available for several years without farm identifications. The technique of using the pooled yield data to identify the groups is called "clustering" in statistics.

Because the zone is referred to as a geographical region covering all farms with similar yields, and because the most important factors affecting yields are assumed to be the agronomical conditions rather than farm operators' characteristics, once a farm is classified into a particular zone based on its historical yield, it stays there without being reclassified in the near future. Or, more likely, the farm is classified by the insurer based on its location rather than its historical yield when the zone locations are identified. In either case, moral hazard won't be a problem—producers' efforts in changing their yields for one year or two won't affect the cluster parameter, their classification, or their indemnities.

The remainder of the article proceeds as follows. In the next section we describe the statistical methods for clustering and classification. These methods are then applied to Whitman County (WA) wheat yield data and corn yield data of several Iowa counties. Conclusions are drawn in the final section.

Statistical Analysis

This section starts from a basic linear model specifying a yield variable based on its cluster, time, and farm. A statistical model is developed that can be used to estimate the cluster parameters based on the data set, and three classification methods are then provided to classify farms into each cluster. A misclassification measure is introduced to evaluate these methods.

The mixed-effects model is defined as:

$$X_{ijt} = \mu_i + f_{ij} + y_t + \varepsilon_{ijt}, \quad i = 1, 2, \ldots, I;\ j = 1, 2, \ldots, n_i;\ t = 1, 2, \ldots, T, \tag{1}$$

where X_{ijt} denotes the yield of farm j in cluster i at year t, μ_i is the mean yield of cluster i, f_{ij} is the random farm effect with mean zero and standard deviation σ_{if}, y_t is the random year effect with mean zero and standard deviation σ_y, and ε_{ijt} is the random error with mean zero and standard deviation σ_ε. The number of clusters is denoted by I, n_i is the number of farms in cluster i, and T is the total number of years. The mean yield μ_i is a fixed cluster effect, the random variable f_{ij} accounts for the effect caused by the unmeasured farm-specific natural conditions and farming practices,[2] and the random year variable y_t accounts for the effect caused by precipitation, temperature, and other yearly factors. We assume all random variables are independently and normally distributed, which can be relaxed as seen later.

Technology improvement over time can be modeled by a common deterministic trend for all clusters. For example, a linear trend, αt, can be added into equation (1) if the time series is long enough to reveal the trend. In this case, the mean yield for cluster i will be $\mu_i + \alpha t$ for any i and t.

[2] Although some farming practices and farm-specific natural conditions, such as managerial expertise and soil type, can be treated as nonrandom, they are not explicitly modeled for three reasons. First, after taking the temporal average later in the analysis, each farm is represented by one observation and can be viewed as a random draw from the pool of unidentified farms. Second, modeling those factors for each farm would be too costly to implement in a national-level ZGRP program. Finally, the data are not available for this study.

The goal here is to cluster and classify farms. In particular, we would like to define a cluster, identify all appropriate clusters, and then classify each farm into one of them. With the linear mixed-effects model, a cluster can be defined as a group of farms with the same yield distribution. Therefore, yields in cluster i must have the same mean $(\mu_i + \alpha t)$, and the same variance $(\sigma_{if}^2 + \sigma_y^2 + \sigma_\varepsilon^2)$.

Model (1) is not estimated directly. Because it is unknown from the data set to which cluster a farm belongs, the likelihood function for a sample from model (1) cannot be evaluated. Therefore, a mixture distribution model is adopted later, using the temporal average yield of each farm for clustering and classification. Let $\overline{X}_{ij}$, $\overline{\varepsilon}_{ij}$, and $\overline{y}$ be the averages over time. Then, in the absence of a time trend, $\overline{X}_{ij} = \mu_i + f_{ij} + \overline{y} + \overline{\varepsilon}_{ij}$ describes the normal random variables with the following covariance structure:

$$(2) \qquad \mathrm{Cov}(\overline{X}_{ij}, \overline{X}_{kl}) = \begin{cases} \sigma_{if}^2 + \dfrac{1}{T}\sigma_y^2 + \dfrac{1}{T}\sigma_\varepsilon^2 & \text{if } i = k,\ j = l, \\ \dfrac{1}{T}\sigma_y^2 & \text{otherwise.} \end{cases}$$

The correlation coefficient between $\overline{X}_{ij}$ and $\overline{X}_{kl}$ $(i \neq k, j \neq l)$ approaches zero when T approaches infinity, implying they are approximately independent for a large T. The presence of the linear trend affects the mean only by a constant, which can be estimated easily by the same method as in the no-trend case. Therefore, it is omitted in the discussion below without losing generality.

Clustering

The T-year average yields $\{\overline{X}_{ij}\}$ consist of I sets of random variables: $\{\overline{X}_{1j}, j = 1, \ldots, n_1\}$, $\{\overline{X}_{2j}, j = 1, \ldots, n_2\}, \ldots, \{\overline{X}_{Ij}, j = 1, \ldots, n_I\}$, that are normally distributed with mean μ_i and standard deviation σ_i $(i = 1, 2, \ldots, I)$, respectively, where $\sigma_i^2 = \sigma_{if}^2 + (\sigma_y^2 + \sigma_\varepsilon^2)/T$ from equation (2). If the variables are all independent, as they approximately are when T is large, then $\{\overline{X}_{ij}\}$ can be viewed as an i.i.d. sample from the following mixture distribution:

$$(3) \qquad f(x; \theta) = \sum_{i=1}^{I} p_i \varphi_i(x \mid \mu_i, \sigma_i),$$

where the parameter vector is $\theta = (p_1, p_2, \ldots, p_{I-1}, \mu_i, \sigma_i, i = 1, 2, \ldots, I)$; the proportion of each normal component is $p_i = n_i/n$, $p_I = 1 - \Sigma_{i=1}^{I-1} p_i$; and $\varphi_i(\cdot \mid \mu_i, \sigma_i)$ denotes the probability density function of the normal distribution with mean μ_i and standard deviation σ_i.

The mixture distribution of (3) has been extensively studied and applied in many fields including agriculture, economics, fisheries, and medicine (Everitt and Hand; Titterington, Smith, and Makov; McLachlan and Basford). According to this application, the sample here is pooled from I normal distributions with p_i percent of the sample coming from the ith distribution. Parameters in the mixture distribution (3) can be estimated via the maximum-likelihood method, once the number of clusters (I) is known. The likelihood function for an i.i.d sample $(x_1, x_2, \ldots, x_n)$ is written as:

$$(4) \qquad \textstyle\prod_i f(x_i; \theta).$$

Although the average yields are not exactly independent, and consequently expression (4) (with x_i being replaced by $\overline{X}_{ij}$) is not the exact but an approximation of the likelihood function of our sample from model (1), we still can maximize (4) for approximate parameter estimation (Wang and Zhang). This procedure is referred to as approximated maximum-likelihood estimation, or AMLE. Note that it does not give estimates for σ_{if}, σ_y, or σ_ε in model (1), but these parameters are not needed in this approach to clustering and classification.

Although the clustering technique has been applied extensively in spatial statistics, use of a temporal average is rarely seen because most of those studies include only cross-sectional observations. Given this situation with both time-series and cross-sectional data, there are at least three advantages of using the temporal average yield rather than the annual yield data.

First, there are often missing values in the annual yield data due to crop rotation and fallow. Despite these anomalies, the average yields still can be regarded as a sample from the mixture distribution. Second, the yields are not independent, but averaging over time will reduce the correlation across farms to make the sample closer to independence. Third, crop yield distributions are usually skewed to the left, because extremely bad weather can totally destroy crops, while even the best of good weather can improve yield by only a small margin above more frequently occurring normal weather. As a result, beta distributions and other nonsymmetric distributions have been used by some agricultural economists when studying yield risk and crop insurance (Nelson; Hennessy, Babcock, and Hayes). Nevertheless, the average yields over the years will approximately follow a normal distribution, based on the Central Limit Theorem, so that we can relax the strong assumption of y_t being normally distributed.

One problem we may encounter in estimating the mixture yield distribution for crop insurance purposes is that the farm-level or less aggregated-level data are kept for no more than 10 years. The limited size of T makes the average yield less independent and less normal. Fortunately, results from Wang and Zhang's paper using data simulated from short time series and beta distributions indicate the AMLE based on equation (3) performs reasonably well in clustering.

Another estimation difficulty is determining the number of clusters (I) for which AMLE cannot give a reliable estimate. The shape of the pooled sample histogram can give some suggestions on I. Actually, I is first picked from the histogram, and then other parameters are estimated in this study. It is a difficult problem to test the null hypothesis that the sample is from an I-normal mixture versus the alternative that the sample is from an $(I-1)$-normal mixture or an $(I+1)$-normal mixture. We have adopted two tests, Akaike's information criterion (AIC) and a bootstrapping likelihood-ratio test. AIC was applied by Scolve, and by Bozdogan and Scolve to determine the proper number of clusters I. For the mixture model (3), AIC is used to choose the I that minimizes

$$AIC(I) = -2\mathrm{L}(\hat{\theta}) + 2N(I),$$

where L is the log likelihood, $\hat{\theta}$ is the MLE of the parameter θ, and $N(I)$ is the number of free parameters in the mixture model with I components. Because AIC methods tend to accept a higher number of clusters if the sample is small (Hurvich and Tsai), a bootstrapping likelihood-ratio test is also applied.

The likelihood-ratio test for the null hypothesis $I = I^*$ against the alternative hypothesis that $I = I^* + 1$ is not valid since the null parameter space is on the boundary of the parameter space rather than in its interior (as noted in Titterington, Smith, and Makov).

Rather than converging to a chi-squared distribution, the likelihood ratio diverges to infinity at a very slow rate when the sample size increases to infinity (Hartigan). McLachlan used a bootstrapping method to obtain the distribution of the likelihood-ratio test statistic for any finite sample size. Some theoretical justifications for the bootstrapping test were given by Feng and McCulloch.

In our particular test, the bootstrap procedure proceeds as follows: first find MLE $\hat{\theta}_0$ using $I = I^*$, and find MLE $\hat{\theta}_1$ using $I = I^* + 1$; then calculate the likelihood-ratio test statistics:

$$W = 2\left[\mathrm{L}(\hat{\theta}_1) - \mathrm{L}(\hat{\theta}_0)\right].$$

To bootstrap W, we generate bootstrap samples from the I^*-normal mixture with parameter $\hat{\theta}_0$, and from each of the bootstrap sample X^* construct the MLE $\hat{\theta}^*$ under the assumption that $I = I^* + 1$, and calculate as follows:

$$W(X^*) = 2\left[\mathrm{L}(\hat{\theta}^*, X^*) - \mathrm{L}(\hat{\theta}_0, X^*)\right].$$

From these quantities, the upper α-quantile of W or the p-value of the test can be found.

Classification

Three classification methods—Bayesian, minimum distance, and maximum probability density—are considered to classify a farm into one of the identified clusters. These methods have been used extensively in the past. McLachlan and Basford introduced the Bayesian method of classification in the first chapter of their book. Friedman and Rubin classified an object into a cluster by minimizing the distance between the object and the center of the cluster. The maximum probability density method follows the same classification strategy as the Bayesian method.

However, there is no ex ante criterion to evaluate these classification methods. Ex post measurements must be taken to justify the method case by case. In this study, a measure of misclassification is introduced in the following section which is calculated based on the clustering parameters and the parameters from each classification method. The lower the misclassification rates a method generates, the better that method is.

The first method we examine is the Bayesian classification that maximizes the posterior distribution. When an observed value x (average yield) is from the mixture distribution, the probability that it belongs to cluster i, according to the Bayesian formula, is:

$$P_i = \frac{p_i \varphi_i(x)}{\sum_{s=1}^{I} p_s \varphi_s(x)},$$

where $\varphi_i(x)$ is the probability density function of cluster i, i.e., $\varphi_i(x \mid \mu_i, \sigma_i)$ in our model. Therefore, a farm is classified to cluster i if $p_i \varphi_i(x) = \max_j \{p_j \varphi_j(x)\}$.

The second method is similar to the minimum distance classification. The standardized distance is used here. More specifically, the farm is classified to cluster i if

$$\frac{|x - \mu_i|}{\sigma_i} = \min_j \left\{\frac{|x - \mu_j|}{\sigma_j}\right\}.$$

The underlying assumption is that classified items should be close to the mean of their own cluster.

The third classification method is to maximize the probability density function directly, i.e., to classify x to cluster i if $\varphi_i(x) = \max_j \varphi_j(x)$.

Misclassification

As with any classification method, misclassification can occur. Here, the concept of misclassification is defined first. Assuming the parameter estimates are the true values, because of the randomness of the average farm yield, there is a chance that a farm belonging to cluster i is incorrectly classified into cluster j by a classification method. This is referred to as misclassification.

The reason for the misclassification is that all of the three classification methods provide fixed boundaries between each adjacent pair of clusters (such as shown later in figures 6–10), classifying farms whose average yields are smaller than a particular boundary into the left-hand cluster only and farms with larger average yields into the right-hand cluster. However, if the average yield observations come from the mixture distribution of several normal distributions (as is modeled), there is an overlapping between any two normal distributions, which means some farms whose average yield is smaller than the boundary may actually belong to the right-hand cluster. These farms will be misclassified.

The misclassification rate, the conditional probability that a farm is classified into cluster j given that it is from cluster i, is defined for each of the three methods (Bayesian, minimum distance, and maximum density) as α_{ij}, which can be directly calculated from the normal distributions:

(5)
$$\begin{aligned}\alpha_{i1} &= \Phi\left(\frac{\beta_1 - \mu_i}{s_i}\right),\\ \alpha_{ij} &= \Phi\left(\frac{\beta_j - \mu_i}{s_i}\right) - \Phi\left(\frac{\beta_{j-1} - \mu_i}{s_i}\right),\\ \alpha_{iI} &= 1 - \Phi\left(\frac{\beta_{I-1} - \mu_i}{s_i}\right), \qquad j = 2, \ldots, I-1,\end{aligned}$$

where $\Phi(\cdot)$ is the cumulative distribution function of the standard normal distribution, and $\beta_1, \ldots, \beta_{I-1}$ are the classification boundaries between each consecutive pair of the I clusters. Boundaries $\beta_1, \ldots, \beta_{I-1}$ are determined by $p_i \varphi_i(\beta_i | \mu_i, \sigma_i) = p_{i+1}\varphi_{i+1}(\beta_i | \mu_{i+1}, \sigma_{i+1})$ for the Bayesian method, $|\beta_i - \mu_i|/\sigma_i = |\beta_i - \mu_{i+1}|/\sigma_{i+1}$ for the minimum distance method, and $\varphi_i(\beta_i | \mu_i, \sigma_i) = \varphi_{i+1}(\beta_i | \mu_{i+1}, \sigma_{i+1})$ for the maximum density method.

There are always more than two misclassification rates. Because these rates are the conditional probabilities for one farm to be misclassified, an unconditional probability of misclassification can be calculated as:

$$P_m = \sum_{i=1}^{I} p_i \sum_{j=1, j \neq i}^{I} \alpha_{ij}.$$

This is the probability that any farm is misclassified, or the total misclassification rate, from a particular method. Minimizing this probability can be an objective criterion in selecting a classification method.

The Cases of Whitman County Wheat Farms and Iowa Corn Farms

Two cases are studied empirically using the model. Whitman is a county in eastern Washington State where dryland wheat production is a prominent industry, and it produces one of the highest yields in the world. Its area is several times larger than a typical county in the Midwest, and it has three distinct precipitation zones: a low precipitation zone receiving 9–14 inches annually, an intermediate precipitation zone receiving 15–18 inches, and a high precipitation zone receiving 19–24 inches. The cropping systems in the three zones are also different, with crops grown once every two years in the low precipitation zone with winter wheat/summer fallow as the primary rotation, twice every three years typically in the intermediate precipitation zone with winter wheat/spring barley/ summer fallow as the primary rotation, and annually with wheat rotated with peas or other crops in the high precipitation zone (USDA). The above conditions result in different wheat yield levels and risks across Whitman County. The county-based GRP is thus not an effective risk-management instrument for these farmers, and no Whitman farmer participated in GRP during 1997. This makes Whitman County a good candidate for sub-county ZGRP. In the case of corn farms in Iowa, the county area is relatively small and the natural conditions are similar across large areas. Therefore, a cross-county ZGRP might be suitable.

The Data and Exploratory Analysis

RMA has recorded dryland winter wheat yields for individual MPCI participants for a maximum of 10 production years during 1981–95. Although the RMA farm yield data come only from those farms who bought insurance in the past, the data cover the majority of farms and the representability will improve in the future as RMA is promoting the insurance programs among growers. The current farm participation rate is reported at 93.5% in Texas, and 82.5% in Nebraska (Coble et al.). A majority (64.5%) of the wheat acreage in Washington State was covered in 1999, though rates may have been lower in previous years.

The annual average yields over the period 1981–95 for 2,945 "farms"[3] are plotted in figure 1 for Whitman County wheat. No obvious trend is present.[4] Figure 2 shows the histogram of the temporal average yield ($\overline{X}_{ij}$) of all Whitman County farms. The distribution appears to have three modes, with the lowest one clearly differentiated from the two higher ones, which are close to each other.

The time trend for state average corn yield in Iowa is checked by plotting National Agricultural Statistics Service (NASS) data from 1982 through 1993 in figure 1. There is no strong trend, so that the temporal average can be calculated as the weighted average with the annual acreage as the weight.[5] The analysis first focuses on one local area

[3] One farm may be divided into multiple plots to purchase different policies. One observation in the data is not necessarily a farm; rather, it is one insurance unit. Regardless, the term "farm" will continue to be used here for convenience.

[4] A linear trend is estimated based on the average yield of Whitman County from 1981–95. The trend coefficient is 0.067, with a t-value of 0.14, which is insignificant at any reasonable critical level.

[5] A linear trend is estimated based on the average yield for the nine Iowa counties from 1982–93. The trend coefficient is 0.76, with a t-value of 0.35, which is insignificant at any reasonable critical level. However, in general, if a trend can be identified, the temporal average needs to be calculated based on the detrended yields.

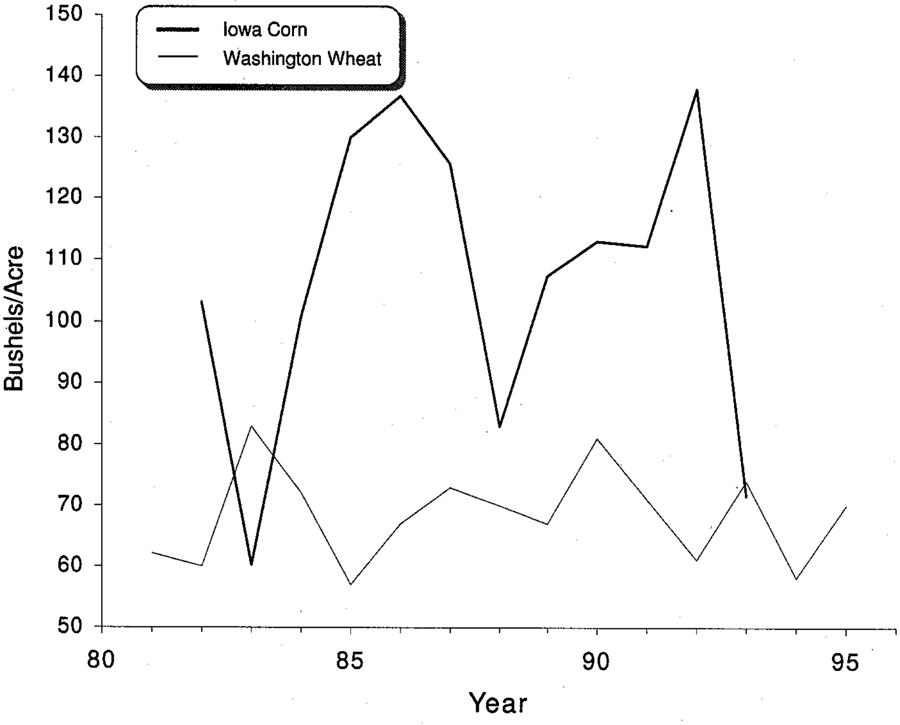

Figure 1. Whitman County wheat yield (1981–95) and Iowa state corn yield (1982–93)

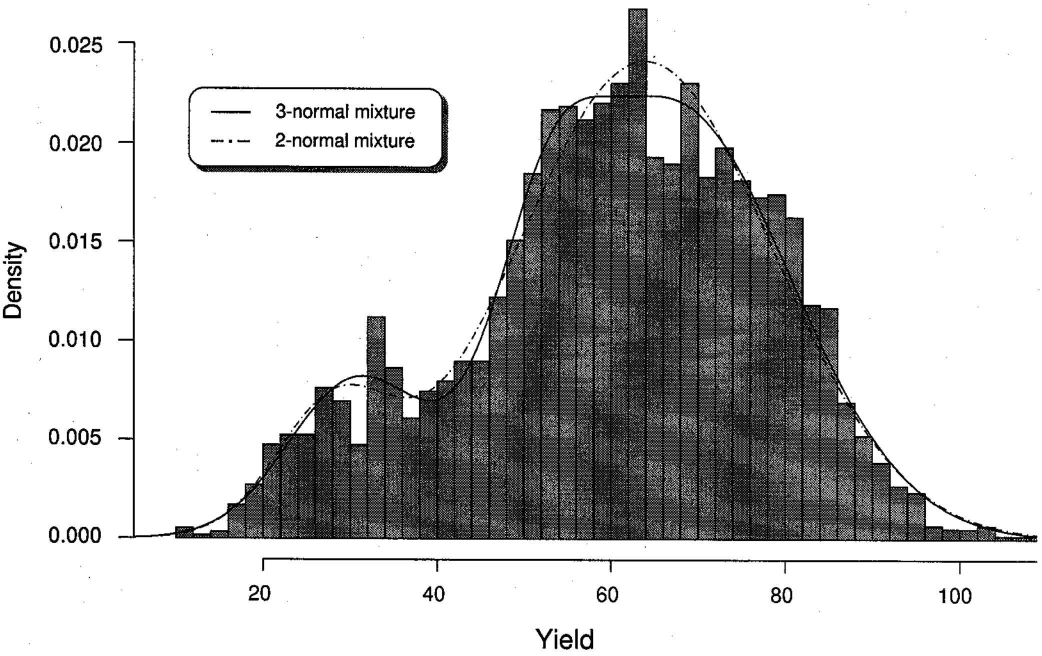

Figure 2. Histogram and estimated density curves for farm average wheat yield, Whitman County

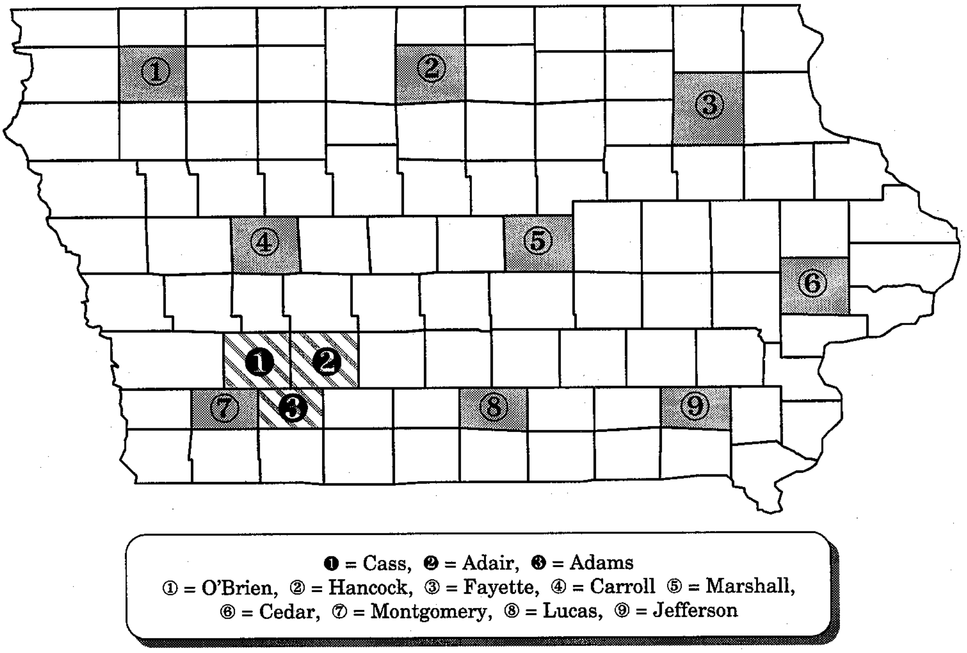
①
②
③
④
⑤
⑥
❶
❷
⑦
❸
⑧
⑨
❶ = Cass, ❷ = Adair, ❸ = Adams
① = O'Brien, ② = Hancock, ③ = Fayette, ④ = Carroll ⑤ = Marshall,
⑥ = Cedar, ⑦ = Montgomery, ⑧ = Lucas, ⑨ = Jefferson

Figure 3. Map of Iowa showing selected counties

in the Southwest Crop Reporting District, including the three Iowa counties of Adair, Adams, and Cass. This area has the highest county corn production record in recent years, but also is subject to high yield risk. FCIC's 10-year farm APH data from 1983–92, with 364 records in the data set, are used. The analysis is then extended to the whole state by choosing nine representative counties: Carroll, Cedar, Fayette, Hancock, Jefferson, Lucas, Marshall, Montgomery, and O'Brien. As seen by the map in figure 3, these counties are evenly distributed throughout the state. There are 1,335 records for the nine counties over the same 10 years. The histograms of the pooled yield distributions are shown in figures 4 and 5 for the three-county and nine-county Iowa cases, respectively.

Parameter Estimation

The approximated MLEs are presented in tables 1 and 2 for Washington wheat and Iowa corn, respectively. For Washington wheat (table 1), the three-component mixture model identifies 15% of the farms from the low precipitation zone, 11% from the intermediate zone, and the remaining 74% from the high zone. The mean yields of the three zones are 30.4, 52.6, and 67 bushels/acre, respectively, and the corresponding standard deviations are 7.6, 5.6, and 13.5 bushels/acre. The last two zones have closer mean yields. Even though the 2-normal mixture distribution also fits the data satisfactorily, it is difficult to interpret the components. It seems the 2-normal mixture combines two components of the 3-normal mixture into one. The AIC suggests choosing the 3-normal mixture. The null hypothesis for the 2-normal mixture versus the alternative 3-normal mixture is also rejected (at 4.4%) by the log-likelihood ratio test based on bootstrapping. This is

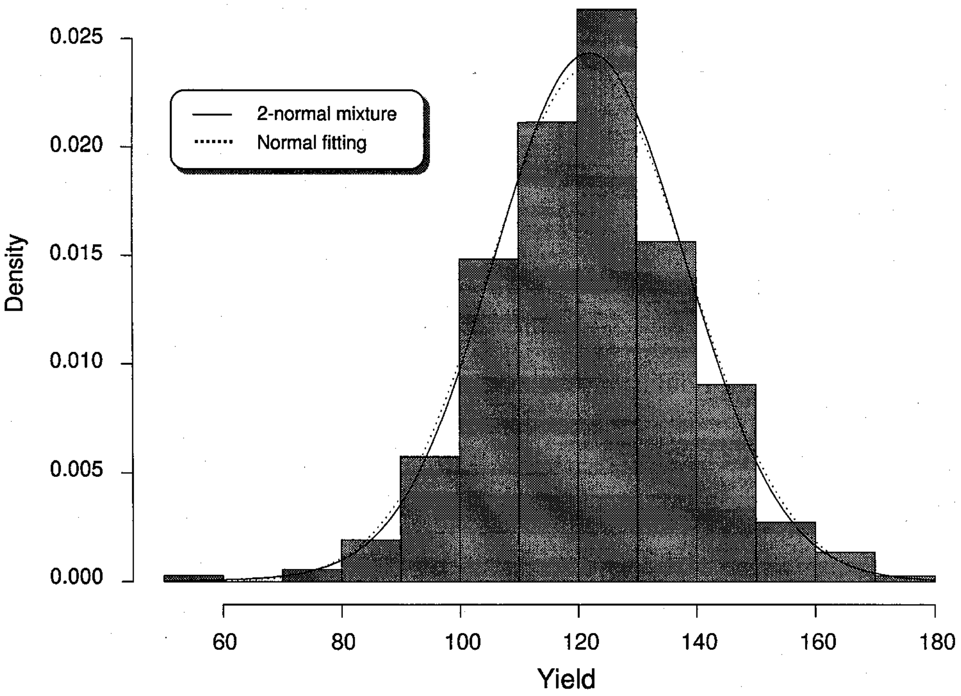

Figure 4. Histogram and estimated density curves for Iowa corn yield in Adair, Adams, and Cass counties

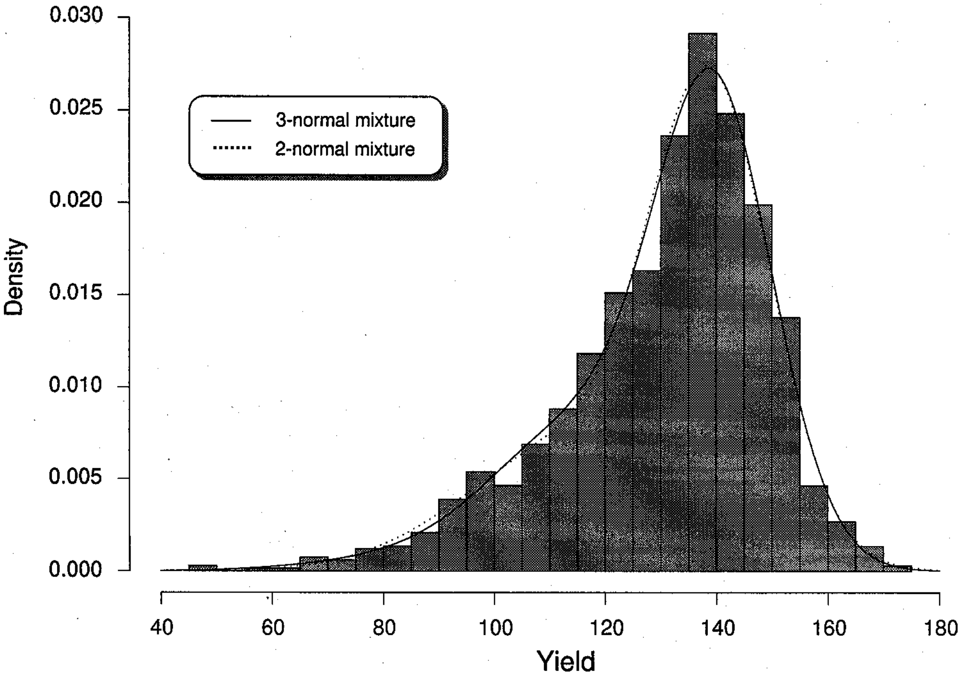

Figure 5. Histogram and estimated density curves for corn yield in nine Iowa counties

Table 1. Approximated Maximum-Likelihood Estimation: Whitman County (WA) Wheat Yield Mixture Model

Distribution	Proportion [p (%)]	Mean Yield [μ (bu./acre)]	Std. Dev. [σ (bu./acre)]	Log Likelihood	AIC
Mixture of 3 Normal Distributions:				-12,581	25,178
1	0.146	30.36	7.64		
	(0.019)	(1.025)	(0.551)		
2	0.107	52.56	5.63		
	(0.078)	(1.01)	(1.72)		
3	0.747	67.01	13.53		
	—	(2.10)	(0.91)		
Mixture of 2 Normal Distributions:				-12,586	25,182
1	0.10	28.09	6.61		
	(0.010)	(0.688)	(0.448)		
2	0.90	63.85	14.77		
	—	(0.394)	(0.301)		

p-value of log-likelihood ratio test based on bootstrapping = 0.044

Note: Values in parentheses are the estimated standard deviations of the estimators, given by the inverse of the information matrix.

consistent with the natural condition of having three zones in the county of Whitman. The fitted curves of 3- and 2-normal densities are plotted in figure 2, showing a better fit by the 3-normal curve.

For Iowa corn (table 2), the three adjacent counties of Adair, Adams, and Cass in the southwest region can be pooled into one normal distribution because the AIC favors one normal distribution to the 2-normal mixture model. In the 2-normal mixture model, only 3.4% of the farms belong to the low cluster. As a result, the bootstrapping method failed to generate a probability for this case. This result indicates that the three southwest Iowa counties can be pooled together, and only one cross-county ZGRP is needed. The average yield is 121.9 bushels/acre, with a standard deviation of 16.9 bushels/acre. The fitted curves of the single normal density and the 2-normal mixture density (plotted in figure 4) are nearly identical.

When studying the nine counties representing the entire state of Iowa, the model shows that a 2-normal mixture distribution fits the data fairly well. The AIC method selects the 2-normal instead of the 3-normal mixture model. The null hypothesis for the 2-normal mixture versus the alternative 3-normal mixture cannot be rejected at any reasonable level by the log-likelihood ratio test based on bootstrapping because the p-value is 92%. There are 36.6% of the farms in the low-yield cluster, with a mean of 116 bushels/acre and a standard deviation of 20.1 bushels/acre. The remaining farms (63.4%) are in the high-yield cluster, with a mean at 139.2 bushels/acre and a standard deviation of 10.7 bushels/acre. Based on these findings, the nine Iowa counties can be classified into only two zones, with one ZGRP corn program for each zone. Again, the fitted curves of the 2- and 3-normal mixture densities (as plotted in figure 5) are almost indistinguishable. There is no attempt to estimate a single normal distribution because the histogram is significantly skewed, suggesting it is not from a normal distribution.

Table 2. Approximated Maximum-Likelihood Estimation: Iowa Corn Yield Mixture Model

Distribution	Proportion [p (%)]	Mean Yield [μ (bu./acre)]	Std. Dev. [σ (bu./acre)]	Log Likelihood	AIC
— 3 Adjacent Southwest Counties (Adair, Adams, Cass) —					
Mixture of 2 Normal Distributions:				-1,543.5	3,097
1	0.034	117.92	31.29		
	—	(18.147)	(16.136)		
2	0.966	122.06	16.11		
	(0.075)	(0.946)	(0.969)		
Single Normal Distribution:				-1,545.03	3,094
1	1.000	121.91	16.87		
	—	(0.884)	(0.625)		
— 9 Counties (One in Each Crop District) —					
Mixture of 3 Normal Distributions:				-5,691.7	11,399.4
1	0.014	75.94	14.14		
	(0.045)	(42.483)	(15.294)		
2	0.362	117.17	17.56		
	(0.134)	(8.412)	(3.713)		
3	0.624	139.82	10.55		
	—	(1.261)	(0.962)		
Mixture of 2 Normal Distributions:				-5,693.8	11,397.6
1	0.366	116.01	20.13		
	(0.063)	(3.330)	(1.129)		
2	0.634	139.16	10.72		
	—	(0.711)	(0.652)		
p-value of log-likelihood ratio test based on bootstrapping = 0.92					

Note: Values in parentheses are the estimated standard deviations of the estimators, given by the inverse of the information matrix.

Classification

We have used the chosen 3-normal mixture distribution for classification of Whitman County wheat. The Bayesian classification is unsatisfactory for the wheat case, since no farm can be classified by this method into cluster 2. The reason is that, according to the clustering results from table 1 discussed above, the second and third clusters have means close to each other, and cluster 2 has only 11% of all farms while cluster 3 has 74%. The huge difference in the proportions causes more farms to be classified into cluster 3 by the Bayesian method.

Presented in table 3 are the classification results by each of the three methods. For the minimum distance classification, 17.4% of Whitman wheat farms are classified into cluster 1, 22.3% to cluster 2, and 60.3% to cluster 3. For the maximum probability density classification, 16.9% of Whitman wheat farms are classified into cluster 1, 31.1% to cluster 2, and 52% to cluster 3. The proportions of farms classified to the three clusters do not agree well among the alternative clustering methods or mixture model estimates.

Table 3. Classification Results by the Three Methods

		Percentage of Each Cluster by Classification Method		
Cluster	Estimated Percentage	Minimum Distance	Maximum Density	Bayesian
Whitman County (WA) Wheat:				
1	14.6	17.40	16.90	N/A
2	10.7	22.30	31.10	N/A
3	74.7	60.30	52.00	N/A
Iowa 9-County Corn:				
1	36.6	42.92	32.96	27.27
2	63.4	57.08	67.04	72.78

Table 4. Comparison of the Three Classification Methods for the Case of Nine Counties in Iowa

		Percentage of Low-Yield Cluster by Classification Method		
County	FIPS[a] County Code	Minimum Distance	Maximum Density	Bayesian
Carroll	19027	24.88	11.94	6.97
Cedar	19031	45.24	35.71	21.43
Fayette	19065	83.33	50.00	50.00
Hancock	19081	27.76	20.00	12.65
Jefferson	19101	87.97	82.59	77.84
Lucas	19117	100.00	98.77	97.53
Marshall	19127	37.60	22.48	13.57
Montgomery	19137	42.95	34.62	29.49
O'Brien	19141	25.00	13.83	12.77

[a] FIPS = Federal Information Processing Standards.

This is due to misclassification. As will be seen below, these classification results are actually what one should expect. One difficulty is that without knowledge of the farm identities, there is no way to confirm whether this classification agrees with the geographic precipitation zones in the county.

When we classify the farms from nine counties in Iowa into two clusters, we have additional information (although still not the farm locations) to check the classification. The classification results in table 3 show that 42.9%, 33%, and 27.3% of the farms are classified into cluster 1, the low-yield zone, by minimum distance, maximum density, and Bayesian methods, respectively, and the remainder fall under the high-yield zone. The classification percentage from the maximum density method is closest to the corresponding estimated parameter. Because we know the county identities of each farm, we can check the spatial distribution of the classification.

As shown in table 4 for the nine-county Iowa case, the three classification methods are consistent in that they all classify the majority of farms in three counties (Fayette, Jefferson, and Lucas) into the low-yield zone, followed by Montgomery, Cedar, and Marshall counties. This pattern shows that the yield is low in the southeast part of Iowa and increases toward the northwest. It is possible that the entire state can be grouped into two zones because (*a*) the three adjacent counties of Adair, Adams, and Cass are clustered into one zone; (*b*) the nine representative counties are evenly spread out through the state and are clustered into two zones; and (*c*) classification results from the nine counties show the low-yield zone lies to the southeast part of the state and the high-yield zone lies to the northwest part of the state. Data including all counties, which were not available in this research, will be needed in order to draw a firm conclusion about zones for the entire state.

Misclassification

We now briefly discuss the misclassification rates of the three methods. For simplicity, we assume the Whitman County wheat yield sample is from the mixture distribution with three normal components, with parameters as in table 3 (i.e., the estimates are the true parameter values). The Bayesian classification does not perform well due to the huge difference between p_2 and p_3. In fact, $p_3\varphi_3(x) > p_2\varphi_2(x)$ for all x. Therefore, no farm is classified into cluster 2. We therefore focus on misclassification rates of the other two methods.

From figure 6, which plots the three distance functions for Whitman County, we observe that the minimum distance method classifies x into cluster 1 if $x < 43.1$, to cluster 2 if x is between 43.1 and 56.8, and to cluster 3 if $x > 56.8$. Therefore, β_1 and β_2 are the classification boundaries of the three clusters, i.e., 43.1 and 56.8, respectively. The misclassification probabilities are listed in table 5. Cluster 1 is fairly well separated from clusters 2 and 3, but the conditional misclassification rate between clusters 2 and 3 is about 20% each way. The unconditional probability of misclassification is 0.2042, or 20.42%.

As seen from the plot in figure 7, the maximum probability density method classifies x into cluster 1 if $x < 42.6$, to cluster 2 if x is between 42.6 and 60.5, and to cluster 3 if $x > 60.5$. The misclassification probabilities (reported in table 5) are obtained from equation (5), with the β's replaced by 42.6 and 60.5. This method misclassifies 26.08% of all farms.

The misclassification for the Iowa case can be discussed in the same manner. The minimum distance method classifies all farms with an average yield lower than 131.1 into the low-yield zone, and the rest to the high-yield zone (figure 8). The break line for the two zones under the maximum density method is 126.0 (figure 9), and the break line imposed by the Bayesian method is 122.3 (figure 10). The misclassification rates for all three methods are reported in table 6. Table 6 data also show the unconditional probabilities of misclassification, indicating that 22.65% of all farms are misclassified by the minimum distance method, 18.3% by the maximum density method, and 26.49% by the Bayesian method.

While the maximum density classification method is outperformed by the minimum distance method for the Whitman County wheat case, it outperforms the other two methods for the Iowa corn case, according to the misclassification measure. In the Whitman County case, cluster 3 has a much larger number of farms than cluster 2; thus $\alpha_{32} = 0.2801$ of maximum density means many more farms are misclassified than under the corresponding minimum distance measure of $\alpha_{23} = 0.2253$, even though the numerical

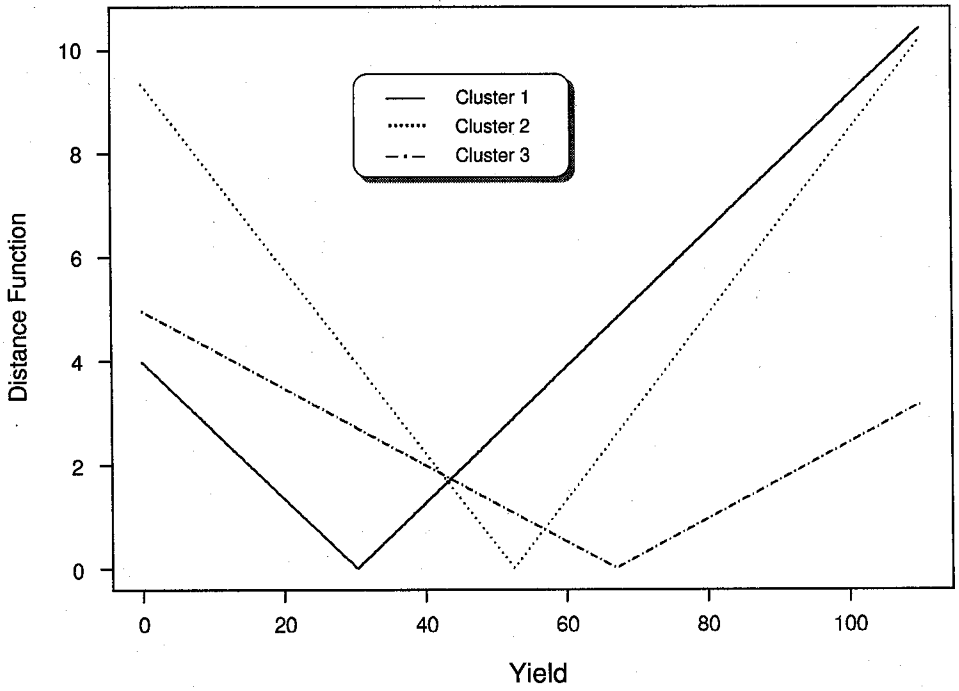

Figure 6. Plots of the distance functions for Whitman County wheat yield

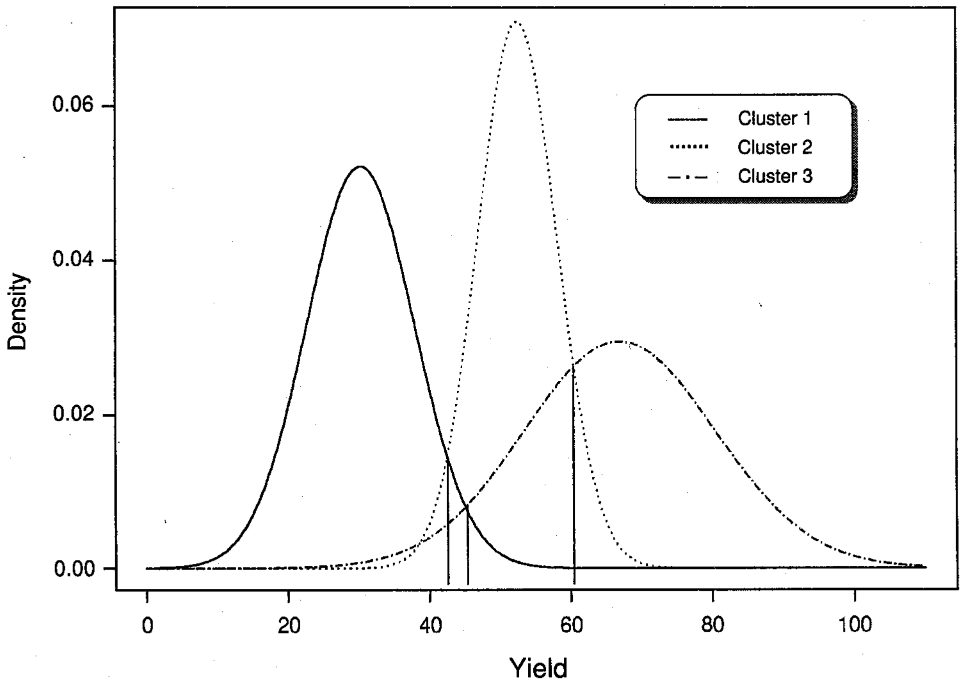

Figure 7. Probability densities for the three normal components for Whitman County wheat yield (maximum density method)

Table 5. Misclassification Rates for Whitman County Wheat Farms

	MINIMUM DISTANCE METHOD			MAXIMUM DENSITY METHOD		
	Farms belonging to:			Farms belonging to:		
Misclassified	Zone 1	Zone 2	Zone 3	Zone 1	Zone 2	Zone 3
Zone 1	—	0.0470	0.0388	—	0.0384	0.0356
Zone 2	0.0467	—	0.1865	0.0545	—	0.2801
Zone 3	0.0003	0.2253	—	0.0000	0.0787	—
Uncondit. Probability:		0.2042			0.2608	

Table 6. Misclassification Rates for Iowa Corn Farms

	MINIMUM DISTANCE METHOD		MAXIMUM DENSITY METHOD		BAYESIAN METHOD	
	Farms belonging to:		Farms belonging to:		Farms belonging to:	
Misclassified	Zone 1	Zone 2	Zone 1	Zone 2	Zone 1	Zone 2
Zone 1	—	0.2266	—	0.1098	—	0.0581
Zone 2	0.2264	—	0.3099	—	0.6230	—
Uncondit. Probability:	0.2265		0.1830		0.2649	

difference is not large. The classified percentages of two out of the three clusters are also closer to the corresponding estimates by the minimum distance method than those by the maximum density method. For Iowa corn data, the proportion of farms classified by the maximum density method in each cluster is closest to that of the model estimation among the three methods.

The classification performance may sometimes be improved by using the yields of each individual year instead of the average yields. However, because of the interdependence of yields, missing yield values, and nonnormal yield distribution, improvement of the classification by using the yearly yields is questionable. Conversely, misclassification is not critical in the implementation of ZGRP, because insurers have the farms' identifications and do not need to rely on the statistical model for determining to which cluster a farm belongs once the clusters have been identified.

Summary and Conclusions

In this analysis, a zone-based Group Risk Plan (ZGRP) is proposed as a substitute for the current county-based GRP. ZGRP retains the advantages of the current GRP programs in eliminating the moral hazard and adverse selection problems associated with Actual Production History (APH), while it can reduce the disadvantages of the county-based GRP by improving its risk-reducing effectiveness for farmers in a heterogeneous county. The zone can be either sub-county or cross-county, depending on the natural conditions in the farming environment.

There is no economic justification for using a county as the area base of the GRP except for the convenience of data availability. The cost of adopting ZGRP involves setting up

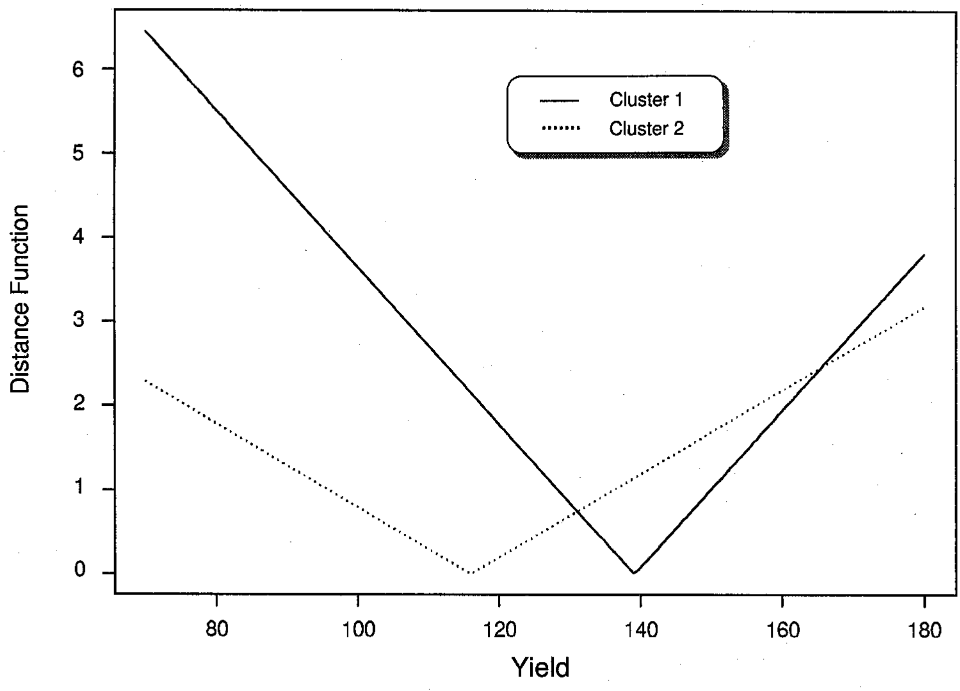

Figure 8. Plots of the distance functions for Iowa corn yield

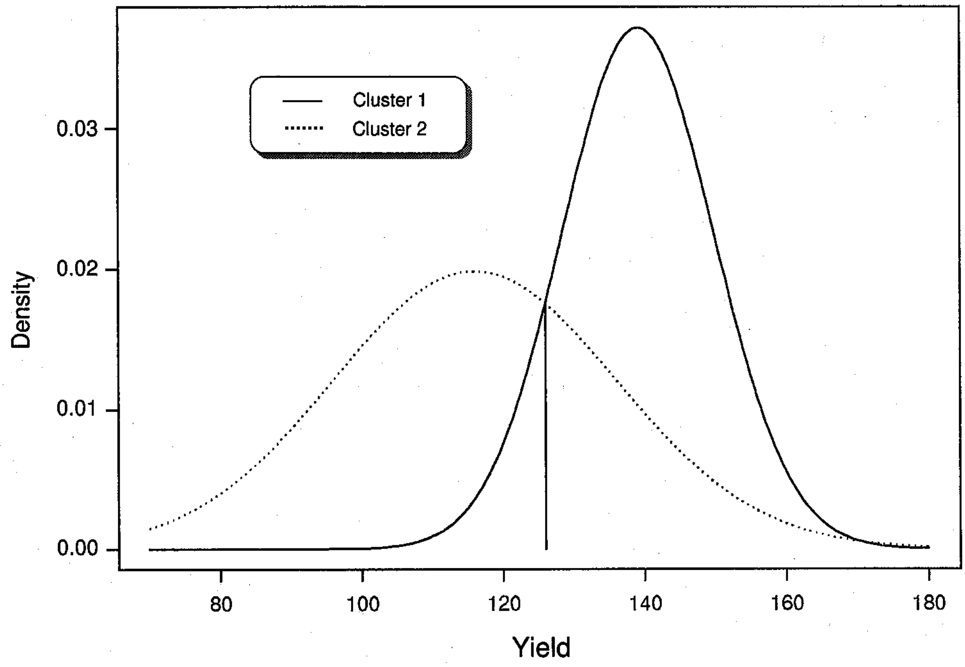

Figure 9. Probability densities for the two normal components for Iowa corn yield (maximum density method)

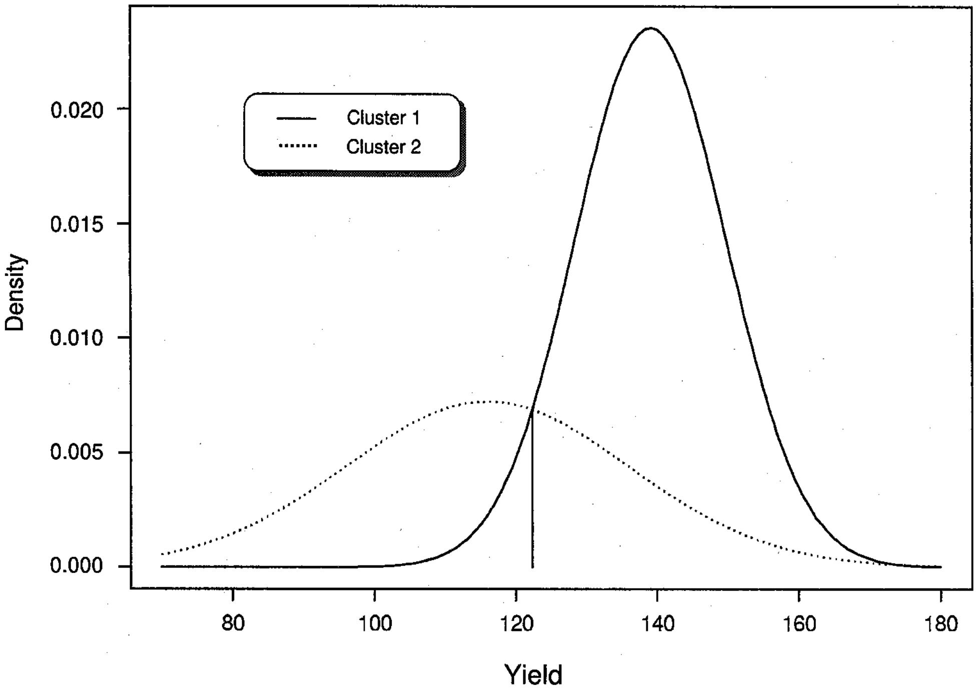

Figure 10. Probability densities for the two normal components for Iowa corn yield (Bayesian method)

a statistical system, similar to the existing one for county average yield, to estimate the zone average yield, once the zones have been identified. There is no need to record individual farm yields, because (*a*) a sampling system can be used to estimate the zone average yield each year based on a few sample farm yields in each zone, and (*b*) the farms are classified into zones based on their location and no reclassification is needed. Therefore, the costs are much lower than those associated with identifying the yield of each individual farm every year, as required under the APH program. Specifically, for adoption of ZGRP, there may be only a couple of zones broken up from each of certain counties, other counties are pooled into zones, and the remainder of the counties are retained as a unit. The costs would not be very substantial when considering the benefits derived from ZGRP's improved risk-management effectiveness.

A statistical approach was developed to cluster and classify each subject into an appropriate category via the mixed-effects model when data are correlated across both time and location. Averages over time are used for clustering and classification. It is noted that these averages have approximately a mixture distribution. Parameters can be estimated by maximizing the approximated likelihood function.

This model was applied to wheat yields in Whitman County of Washington State to identify clusters of farms and classify farms into different clusters for the purpose of design, evaluation, and establishment of a sub-county zone-based crop insurance instrument. Three clusters were identified, which directly correspond to Whitman County's three precipitation zones with individually distinct farming practices. The model was also applied to corn yields in Iowa, indicating cross-county zone-based crop insurance may also be suitable. In fact, the results suggest it is very likely the counties in southeast Iowa can be pooled into a low-yield zone and those in northwest Iowa into a high-yield zone.

The major contribution of this study is the identification of a method for clustering farms based only on their yield records, without requiring detailed analysis of the agronomic and socioeconomic factors which affect yield distributions. Although the classification methods are not ideal, especially when the clusters are close to each other, this will not present a serious problem in actual implementation of ZGRP. Once the geographic zones are identified, the insurers will be able to classify each insuring farm according to their knowledge of the farm location.[6]

One weakness of the current numerical analysis is the use of RMA recorded farm yields for clustering. RMA records only include insurance participants' yields, not all farms' yields—and the RMA participation rate was no more than 50% during the years for Whitman wheat data. Consequently, identified clusters may be biased here. However, this is just a numerical example, and the RMA record is expanded each year due to more and more farms participating in the federal crop insurance programs. Using more recent RMA yield data will reduce this problem.

[*Received October 1999; final revision received August 2000.*]

References

Bozdogan, H., and S. L. Scolve. "Multi-Sample Cluster Analysis Using Akaike's Information Criterion." *Annals Institute of Statis. Mathematics* 36(1984):163–80.

Coble, K. H., T. O. Knight, G. F. Patrick, and A. E. Baquet. "Crop Producer Risk Management Survey: A Preliminary Summary of Selected Data." Information Rep. No. 99-001, Dept. of Agr. Econ., Mississippi State University, 1999.

Everitt, B. S., and D. J. Hand. *Finite Mixture Distributions.* London: Chapman and Hall, 1981.

Feng, Z. Z., and C. E. McCulloch. "Using Bootstrap Likelihood Ratios in Finite Mixture Models." *J. Royal Statis. Soc., Series B*, 58(1996):609–17.

Friedman, H. P., and J. Rubin. "On Some Invariant Criteria for Grouping Data." *J. Amer. Statis. Assoc.* 62(1967):1159–78.

Hartigan, J. A. "A Failure of Likelihood Asymptotic for Normal Mixtures." In *Proceedings of the Berkeley Conference in Honor of Jerzy Neyman and Jack Kiefer*, Vol. II, eds., L. LeCam and R. A. Olshen, pp. 807–10. Belmont CA: Wadsworth and Brooks, 1985.

Hennessy, D. A., B. A. Babcock, and D. J. Hayes. "Budgetary and Producer Welfare Effects of Revenue Insurance." *Amer. J. Agr. Econ.* 79(1997):1024–34.

Hurvich, C. M., and C.-L. Tsai. "A Cross-Validatory AIC for Hard Wavelet Thresholding in Spatially Adaptive Function Estimation." *Biometrika* 85(1998):701–10.

McLachlan, G. J. "On Bootstrapping the Likelihood Ratio Test Statistic for the Number of Components in Normal Mixture." *Appl. Statis.* 36(1987):318–24.

McLachlan, G. J., and K. E. Basford. *Maximum Models: Inferences and Applications to Clustering.* New York: Marcel Dekker, Inc., 1987.

Miranda, M. J. "Area-Yield Crop Insurance Reconsidered." *Amer. J. Agr. Econ.* 73(1991):233–42.

[6] Although this analysis does not provide a specific method to identify the geographic zones corresponding to yield clusters, which will need intuitive knowledge of a particular county (such as the break lines of the precipitation zones in Whitman County), the clustering method can, at least, identify the number of zones needed in each county, and the classification methods can provide the approximate location of each of these zones, if farm location is known.

Nelson, C. H. "The Influence of Distributional Assumption on the Calculation of Crop Insurance Premia." *N. Cent. J. Agr. Econ.* 12(1991):71–78.

Scolve, S. L. "Application of the Conditional Population-Mixture Model to Image Segmentation." *IEEE Transactions on Pattern Analysis and Machine Intelligence*, PAMI-5 (1983):428–33.

Skees, J. R., J. R. Black, and B. J. Barnett. "Designing and Rating an Area Yield Crop Insurance Contract." *Amer. J. Agr. Econ.* 79(1997):430–38.

Titterington, D. M., A. F. M. Smith, and U. E. Makov. *Statistical Analysis of Finite Mixture Distributions.* London: John Wiley and Sons, Ltd., 1985.

U.S. Department of Agriculture. "Palouse Cooperative River Basin Study." Cooperative study by Soil Conservation Service, Forest Service, and Economics, Statistics, and Cooperatives Service of Whitman County, WA. Washington DC: Government Printing Office, 1978.

U.S. General Accounting Office. "Crop Insurance Additional Actions Could Further Improve Program's Financial Condition." Report to the Ranking Minority Member, Committee on Agriculture, Nutrition, and Forestry, U.S. Senate. Pub. No. GAO/RCED-95-269, Washington DC, September 1995.

Wang, H. H., S. D. Hanson, R. J. Myers, and J. R. Black. "The Effects of Yield Insurance Designs on Farmer Participation and Welfare." *Amer. J. Agr. Econ.* 80(1998):806–20.

Wang, H. H., and H. Zhang. "Model-Based Clustering for Cross-Sectional and Time-Series Data." Work. pap., Dept. of Agr. Econ., Washington State University, Pullman, 1999.

Williams, J. R., G. L. Carriker, G. A. Barnaby, and J. K. Harper. "Crop Insurance and Disaster Assistance Designs for Wheat and Grain Sorghum." *Amer. J. Agr. Econ.* 75(1993):435–47.