# Using the Spatial Statistics Approach to Analyze Yield Risk Pooling in the US

H. Holly Wang and Hao Zhang[1]

AAEA Annual Conference
July 28-31, 2002
Long Beach, CA

---

[1] The authors are assistant professors with the Department of Agricultural Economics and the Program in Statistics, Washington State University, respectively.

# ON THE POSSIBILITY OF PRIVATE CROP INSURANCE MARKET: A SPATIAL STATISTICS APPROACH

Abstract

Risk theory tells us if an insurer can effectively pool a large number of individuals to reduce the total risk, he then can provide the insurance by charging a premium close to the actuarially fair rate. There is, however, a common belief that the risk can be effectively pooled only when the random loss is independent, so that crop insurance markets cannot survive without government subsidy because crop yields are not independent among growers. In this paper, we take a a spatial statistics approach to examine the effectiveness of risk pooling for crop insurance under correlation. We develop a method for evaluating the effectiveness of risk pooling under correlation and apply the method to three major crops in the US: wheat, soybeans and corn. The empirical study shows that yields for the three crops present zero or negative correlation when two counties are far apart, which complies with a weaker condition than independence, finite-range positive dependency. The results show that effective risk pooling is possible and reveal a high possibility of a private crop insurance market in the US.

Crop insurance has been an important instrument for protecting farmers' income against low yield resulting from adverse weather and other natural disasters. Except for a few perils such as hail and fire, the multiple peril crop insurance (MPCI) has only been offered by the US government with a huge subsidy. MPCI was first introduced in 1938 on a trial basis, and extended in 1980 to most crops in the US. MPCI pays an indemnity to a farmer based on the difference between a pre-selected coverage yield level and the farmer's realized yield

level. For years, billions of dollars of financial deficits have been accumulated by the government to provide MPCI, and farmers' participation rate is still low. These problems have been attributed to moral hazard, adverse selection, and high administrative costs (Knight and Coble,1997; Skees, Black and Barnett,1997; Goodwin and Smith,1995). To deal with these problems, a Group Risk Plan (GRP) was introduced in 1994, which pays a farmer an indemnity only when the realized average yield of his county falls below the pre-selected coverage level. Evaluation of county yields instead of individual farm yields for indemnification greatly reduces the insurer's administrative costs. The disadvantage of GRP is that it does not protect farmers as effectively as MPCI when the farm yield and county yield are not highly correlated. In the late 1990s, revenue insurance programs were piloted that could protect farmers from both price and yield risks including those that were based on farm yields such as Income Protection, Crop Revenue Coverage and Revenue Assurance and those based on county average yields such as Group Revenue Insurance Programs.

To cope with the actuarial problems with federal crop insurance programs, three lines of study in the area of agricultural risk protection are currently advanced by economists. The first line is to investigate the theory of area-based insurance and the risk protection effectiveness of GRP (Miranda, 1991; Wang et al, 1998). The second line is to study revenue insurance which takes the advantage of usually negatively correlated price and yield risks and focuses on stabilizing the overall income (Hennessy, 1997; Skees et al, 1998). The third line is to study market instruments that are based on natural conditions, such as weather derivatives, which also deal with the moral hazard problem (Turvey,1999). The first two lines are focused on the current government programs, and the last line also fails to address the question of whether it is feasible for private insurers to provide agricultural insurance.

Although positive correlation among farm yields is the basis for GRP and weather derivatives, it is perhaps the major factor that has discouraged the consideration of private crop insurance. It is a common belief that effective risk pooling is built upon the independence

between risk exposure units, and that a private market for crop insurance is doomed to fail because of the systemic risk existing in crop yield (Miranda and Glauber, 1997). However, this belief has not been thoroughly studied. Therefore, it is important to investigate carefully whether private agricultural insurance and reinsurance markets can exist without or with a minimum government subsidy, what conditions are required, and whether these conditions are present in the current situation. The objectives of this research are (1) to explore necessary statistical conditions for effective risk pooling; (2) to investigate the pattern of US crop yields' correlation; and (3) to develop a method to evaluate the effectiveness of risk pooling under correlation and apply it to the US crops. Each of the objectives is pursued in one of the following three sections, and the last section consists of a summary and discussion.

**Statistical Foundations of Insurance**

The primary function of insurance is risk pooling. Mehr, Cammack and Rose (1985) offer the following definition, "*Insurance may be defined as a device for reducing risk by combining a sufficient number of exposure units to make their individual losses collectively predictable.*" In the insurance literature, the occurrence of an aggregate loss that is so large as to deplete the insurance fund is captured by the concept of ruin. It has been suggested that a possible objective criterion for the management of an insurance pool is to minimize the probability of ruin in a given time period or perhaps maximize returns subject to maintaining a specified probability of ruin (Bühlmann, 1970). These are the Safety-First criteria developed by Roy (1952) and Telser (1956). The aggregated premium surplus above the expected value of the aggregate loss required to maintain a particular probability of ruin is referred to as the buffer fund. In what follows, we will consider the statistical foundations of insurance based upon the framework provided by Cummins (1991), while incorporating the spatial characteristics of crop insurance.

Review of some basic concepts in premium rates is now in order. Premium rate setting generally begins with the concept of pure or net premiums (Hogg and Klugman, 1984; Borch,

1974; Goodwin and Smith,1995). The net premium is simply the expected indemnity per exposure unit. The gross premium, the amount paid by the insured per exposure unit in order to be eligible for coverage, is larger than the net premium by an amount referred to as the premium loading factor. We can examine the components of the gross premium by decomposing the gross premium into three parts, $P = P_N + A + L$, where $P$ is the gross premium, $P_N$ is the net premium, $A$ is the administrative cost load, $L$ is the buffer load, and $A + L$ is the premium loading factor.

Let $Y(\boldsymbol{s}, t)$ be the yield at the $t$-th year of the exposure unit with a two-dimensional spatial index $\boldsymbol{s}$ representing its location, and $c(\boldsymbol{s}, t)$ be the pre-specified coverage level. The exposure unit will experience a production loss in the amount of $X(\boldsymbol{s}, t) = \max\{0, c(\boldsymbol{s}, t) - Y(\boldsymbol{s}, t)\}$, and this amount is also the indemnity payment for this unit at year $t$ if we assume that price is normalized to be unitary. Once a model on the yield $Y(\boldsymbol{s}, t)$ is built, the distribution of $Y(\boldsymbol{s}, t)$ is then specified, from which we can obtain the distribution of $X(\boldsymbol{s}, t)$ and hence the mean loss $E(X(\boldsymbol{s}, t))$ for any given prespecified value $c(\boldsymbol{s}, t)$. The total loss of $N$ exposure units at year $t$ is therefore

$$S_N(t) = \sum_{i=1}^{N} X(\boldsymbol{s}_i, t), \tag{1}$$

where $i$ denotes the $i$-th exposure unit, and $\boldsymbol{s}_i$ is the spatial location of the $i$th unit. In the model given by equation (1), each individual loss is conceptualized as a random variable and the total loss experienced by the pool is random as well. The expected total loss of the pool at year $t$ is: $E(S_N(t)) = \sum_{i=1}^{N} E(X(\boldsymbol{s}_i, t))$. We need to obtain the variance of the total loss in order to know how much the actual total loss, $S_N(t)$, can differ from the expected total loss. The variance of the total loss at year $t$ is:

$$Var(S_N(t)) = \sum_{i=1}^{N} Var(X(\boldsymbol{s}_i, t)) + 2\sum_{i=1}^{N} \sum_{j=i+1}^{N} Cov(X(\boldsymbol{s}_i, t), X(\boldsymbol{s}_j, t)). \tag{2}$$

Under some regularity conditions, the following central limit theorem holds for a spatial

process

$$\frac{S_N(t) - E(S_N(t))}{\sqrt{Var(S_N(t))}} \sim N(0,\ 1). \tag{3}$$

See, for example, Bolthausen (1982) for stationary random fields and Guyon (1995, Theorem 3.3.1, page 112) for non-stationary random fields. Therefore,

$$P\left(\ |S_N(t) - E(S_N(t))| < 1.96\sqrt{Var(S_N(t))}\ \right) \approx 0.95. \tag{4}$$

In this case the buffer fund needs to be $E(S_N(t)) + 2\sqrt{Var(S_N(t))}$ in order to maintain a 0.025 probability of ruin for year $t$.

As mentioned previously, $E(S_N(t))$ can be computed once the yield $Y(\boldsymbol{s}, t)$ is modeled and all prespecified levels are known. $E(S_N(t))$ can also be calculated if the prespecified levels are unknown but have some known probability distribution. Calculation of $Var(S_N(t))$ is more cumbersome unless individual losses are independent. In addition, the losses are truncated and spatially dependent random variables. Even though some methods in spatial statistics may seem appropriate and applicable here, it is difficult to directly apply the methods because the losses are generally non-stationary unless we put some unrealistic assumptions of the prespecified levels $c(\boldsymbol{s}, t)$. In this paper, we will propose an approach to overcome the difficulty.

In order to gain some insight to the risk pooling theory, let us first consider the simplest case when the losses are independently and identically distributed.

### *An Example—The i.i.d. Case*

Even though the assumption of i.i.d. losses is unnecessarily strong and generally not realistic (Bühlmann, 1970), results are much simplified under these assumptions and can help us see some key issues in risk pooling. When the losses are i.i.d,

$$E(S_N(t)) = N\mu, \quad Var(S_N(t)) = N\sigma^2, \tag{5}$$

where $\mu$ and $\sigma^2$ are the mean and variance of loss from each exposure unit, respectively. The central limit theorem implies that, when $N$ is large, $S_N(t)$ is approximately normally

5

distributed and thus, for any $0 < \alpha < 1$,

$$P(|S_N(t) - N\mu| < z_{\alpha/2}\sqrt{N}\sigma) \approx 1 - \alpha,$$

where $z_{\alpha/2}$ is the positive value which the standard normal random variable exceeds with a probability $\alpha/2$. For $\alpha = 0.05$, the corresponding value of $z_{\alpha/2}$ is 1.96. Hence, to maintain a probability of ruin no more than 0.025, the liquid buffer fund needs to be approximately $1.96\sigma\sqrt{N}$ for $N$ individuals. The buffer load $L$ for each individual unit is $1.96\sigma/\sqrt{N}$, which decreases as the number of exposure units $N$ increases.

If we assume economies of scale in the administration function, then the administrative cost per exposure unit, $A$, will decline as the size of the insurance pool grows. Thus, for a sufficiently large insurance pool, the premium loading $A + L$ will not be large, i.e., the risk premium a risk averse insured must pay to obtain coverage will be small.

From this example we see that a critical ingredient for risk pooling is that the standard deviation of the average loss, i.e., $\sigma/\sqrt{N}$ in the i.i.d. case, diminishes as the size of the insurance pool increases. Under this condition, the buffer load also declines as the pool grows larger, thus ensuring (assuming economies of scale in administration) that the premium loading is relatively small. As we will see in the sequel, these properties may be retained when losses are not independent.

### Finite-Range Positive Dependency of Spatial Variables

In what follows we will explicitly relax the assumption of independence. Many statistical results that hold under the i.i.d assumption can be generalized into the dependence case. For example, under some mixing and regularity conditions, the central limit theorem (3) holds for the spatial process $X(\boldsymbol{s}, t)$ (See, e.g., Theorem 3.3.1, page 112 of Guyon, 1995): where the mixing conditions ensure that the dependency dies off sufficiently quickly as the lag distance increases for spatial process $X(\boldsymbol{s}, t)$.

In the context of crop insurance, although an established insurance company can use reserve funds from the earlier years' premium surplus to pay for a particular year's indemnity,

the discussion of effective risk pooling is still focused on the loss in any particular year. Therefore, it is particularly useful to investigate how yields or yield losses are spatially correlated for a given year. For example, how likely is it that the majority of the insured experience a loss in a particular year, in which case the insurer might be jeopardized? It seems that the spatial statistics approach has not been employed in the study of crop insurance, although spatial statistics has been successfully used in agriculture for other purposes (e.g., see Mercer and Hall 1911; Besag and Higdon, 1999, both for uniformity trials; Roberts, English, and Mahajanashetti, 2000 for precision agriculture; Zhang, 2001 for an application in plant pathology.)

Let us first review some terminologies and theories in spatial statistics. We refer readers to Cressie (1993, Chapter 2) for more details on spatial statistics. A spatial process, $X(\boldsymbol{s})$, is said to be second-order stationary if the mean is a constant and for any *vectors* $\boldsymbol{s}, \boldsymbol{h}$,

$$Cov(X(\boldsymbol{s}), X(\boldsymbol{s} + \boldsymbol{h})) = C(\boldsymbol{h}),$$

where $C(\cdot)$ is called the covariogram; and $\boldsymbol{h}$ is called the lag. Therefore, the covariogram measures the correlation structure of the spatial process. Here $X(\boldsymbol{s})$ refers to a general spatial process and should not be confused with $X(\boldsymbol{s}, t)$, the yield loss. Using data $X(\boldsymbol{s}_i), i = 1, \cdots, N$, the covariogram can be estimated by:

$$\hat{C}(\boldsymbol{h}) = (1/|N(\boldsymbol{h})|) \sum_{(i,j) \in N(\boldsymbol{h})} (X(\boldsymbol{s}_i) - \hat{\mu}_i)(X(\boldsymbol{s}_j) - \hat{\mu}_j) \tag{6}$$

where $N(\boldsymbol{h})$ is a set containing all pairs $(i, j)$ such that $\boldsymbol{h} = \boldsymbol{s}_j - \boldsymbol{s}_i$ for a particular level of $\boldsymbol{h}$; $|N(\boldsymbol{h})|$ is the number of pairs in the set $N(\boldsymbol{h})$; and $\hat{\mu}_i$ and $\hat{\mu}_j$ are the estimated means of $X(\boldsymbol{s}_i)$ and $X(\boldsymbol{s}_j)$ respectively.

$X(\boldsymbol{s})$ is said to be isotropic if its covariogram $C(\boldsymbol{h})$ depends on distance $\|\boldsymbol{h}\|$ only but not on the direction, i.e., the covariogram is non-directional. The covariogram is then a function of distance $h$ and, following a conventional notation, we use $C(h), h > 0$ to denote

the isotropic covariogram. Then $C(h)$ can be estimated by

$$\hat{C}(h) = (1/|N(h)|) \sum_{(i,j) \in N(h)} (X(\boldsymbol{s}_i) - \hat{\mu}_i)(X(\boldsymbol{s}_j) - \hat{\mu}_j) \tag{7}$$

where $N(h)$ is the set of all pairs $(\boldsymbol{s}_i, \boldsymbol{s}_j)$ such that the distance $\|\boldsymbol{s}_i - \boldsymbol{s}_j\|$ is $h$, $|N(h)|$ is the number of elements in $N(h)$. In reality, it is possible that $X(\boldsymbol{s}_i)$s have different means but still possess a stationary covariogram. In this case, we will first remove the mean and then estimate the covariogram. Note that in practice, only a few pairs are exactly the same distance apart, $N(h)$ is therefore a set that contains all pairs $(\boldsymbol{s}_i, \boldsymbol{s}_j)$ so that the distance $\|\boldsymbol{s}_i - \boldsymbol{s}_j\|$ is approximately $h$. We refer to Journel and Huijbregts (1978, p.194) and Cressie (1993, p. 70) for recommendation of the "tolerance" region.

Another measure of correlation for a spatial process is the correlogram that is defined as the covariogram divided by the variance:

$$\rho(\boldsymbol{h}) = \frac{C(\boldsymbol{h})}{\sigma^2}, \tag{8}$$

where $\sigma^2$ is the common variance of each $X(\boldsymbol{s}_i)$. In almost all interesting situations, the correlation between any two points becomes weaker when the lag distance increases. A process is said to be finite-range dependent, or m-dependent (Davis and Borgman, 1982) if any two points, $X(\boldsymbol{s}_i)$ and $X(\boldsymbol{s}_j)$ are independent when the lag distance $\|\boldsymbol{s}_i - \boldsymbol{s}_j\| > m$ for some real positive number $m$. The finite-range dependence implies the mixing conditions that ensure the central limit theorem in (3).

Here, we define a spatial process, $X(\boldsymbol{s})$, to be finite-range positive dependent (f.r.p.d.) if the correlation between any two points, $X(\boldsymbol{s}_i)$ and $X(\boldsymbol{s}_j)$ is positive when the lag distance $\|\boldsymbol{s}_i - \boldsymbol{s}_j\| < d$ for some d, and zero or negative otherwise. This f.r.p.d. assumption seems quite sensible in light of the fact that the correlation between the yield losses of two fields caused by adverse weather will eventually vanish when the two fields are very far away from each other. We can give sufficient conditions under which the central limit theorem (3) holds for f.r.p.d. processes. These conditions must ensure that dependency quickly dies off (so does

8

the correlation) when the lag distance increases. Discussion of these conditions is beyond the scope of this paper. Note in the definition of f.r.p.d, the spatial process needs not to be stationary.

For any location $\boldsymbol{s}_i$, denote by $O_i$ the set of all locations whose distances from $\boldsymbol{s}_i$ are at most $d$, i.e., $O_i = \{\boldsymbol{s}_j : 0 \leq \|\boldsymbol{s}_j - \boldsymbol{s}_i\| \leq d\}$. If the process $X(\boldsymbol{s})$ is f.r.p.d. and the variance of $X(\boldsymbol{s})$ is bounded by some constant $\sigma_m^2$, we can place the following bound on the variance

$$
\begin{aligned}
Var(\bar{X}_N) &= (1/N^2) \left[ \sum_{i=1}^{N} Var(X(\boldsymbol{s}_i)) + \sum_{i=1}^{N} \sum_{j=1, j\neq i}^{N} C(\boldsymbol{s}_j - \boldsymbol{s}_i) \right] && (9) \\
&\leq (1/N^2) \left[ N\sigma_m^2 + \sigma_m^2 \sum_{i=1}^{N} \sum_{j=1, j\neq i}^{N} \rho(X(\boldsymbol{s}_i), X(\boldsymbol{s}_j)) \right] \\
&\leq \sigma_m^2/N + (\sigma_m^2/N^2) \sum_{i=1}^{N} |O_i| && (10) \\
&\leq (\sigma_m^2/N) \left( 1 + \max_i(|O_i|) \right), && (11)
\end{aligned}
$$

where $|O_i|$ denotes the number of elements in $O_i$. Inequality (10) holds because each correlation coefficient is no greater than one, the sum over $j$ is no larger than the number of elements in set $O_i$ under f.r.p.d., and the inequality (11) holds because each set cardinality $|O_i|$ is no larger than the largest set size.

If the spatial process $X(\boldsymbol{s})$ is f.r.p.d. with a finite-range $d$, then for any site $\boldsymbol{s}_i$ the corresponding set $O_i$ has at most a finite number of elements. Indeed, if the exposure units (counties, farms or plots) have a minimum area, $a$, then $\max_i(|O_i|) \leq \pi d^2/a$, i.e., bounded by a constant that depends on $d$ but not on $N$ or $i$. We then have, in view of (10), $Var(\bar{X}_N) \to 0$, as $N$ goes to infinity, i.e., the variance of $\bar{X}_N$ diminishes as $N$ gets larger. If $X(\boldsymbol{s})$ represents the yield loss, this diminishment would mean that the buffer load per unit is very small for large $N$. Therefore, if a reasonably large number of units from a geographically large area participate in the insurance program, the premium charged to each participant can be very close to the actuarially fair level. As long as the farmer is risk averse, s/he should be willing to pay for such insurance.

We have just shown that $Var(\bar{X}_N)$ vanishes as $N \to \infty$ for a f.r.p.d. process. This is an asymptotic result that assures us that $Var(\bar{X}_N)$ can be arbitrarily small if the sample size is sufficiently large. In reality, only limited number of sample data are available. For a particular sample size, we may be able to evaluate $Var(\bar{X}_N)$ by directly applying (9) to get a better bound than (11).

If $X(\boldsymbol{s})$ represents the yield loss, the effectiveness of risk pooling depends much on how fast the variance of average loss diminishes. Several factors may affect this speed of diminishment: how fast the correlation of yields dies off when the lag distance increases, the specified coverage levels and range of dependence. Using some spatial statistics technique, we will show next how to calculate the variance of the average loss, as illustrated by the US crop yield data.

**The Pattern of US Crop Yield Dependence**

Because crop yields are correlated due to weather patterns, it is believed that private insurance companies can not effectively pool the risks without significant subsidies from the government. We intend to show here that such belief is not well grounded. We will explore the spatial dependence pattern of US crop yields for a given year to find out if the correlation between exposure units quickly dies off as two units become far apart, leaving enough room for risk pooling.

*Data and detrending*

County level crop yields from 1972 to 1997 provided by National Agricultural Statistics Service, United States Department of Agriculture, are used in this analysis. This is because farm level yield data are not available. Three major crops, corn, soybeans, and wheat are studied. There are 2,591 counties for corn, 2,000 counties for soybeans, and 2,641 counties for wheat in the data set (counties with less than five years of observations are dropped). The centroid latitude and longitude of each county are obtained from Bureau of the Census, United States Department of Commerce. These spherical references are then transformed

10

into plane coordinates using ArcInfo, so that they can be applied in the statistical software S-Plus (Mathsoft, Inc., Seattle) which has a spatial module. The temporal trend is removed first by location. For each location, yields are detrended by a log quadratic trend. Log quadratic trends are generally used for crop yields (Miranda and Glauber, 1997; Wang et al., 1998). Let $\xi(\boldsymbol{s}, t)$ be the detrended yield, where $\boldsymbol{s}$ is the county centroid and $t$ is the year. That is, $\xi(\boldsymbol{s}, t)$ is the deviation of the yield from its mean yield at year $t$. The yield is modeled as,

$$Y(\boldsymbol{s}, t) = \exp\{\alpha(\boldsymbol{s}) + \beta(\boldsymbol{s})t + \gamma(\boldsymbol{s})t^2\} + \xi(\boldsymbol{s}, t),$$

where $\alpha$, $\beta$ and $\gamma$ are location-specific parameters; and the mean of yield is $m(\boldsymbol{s}, t) = E(Y(\boldsymbol{s}, t)) = \exp(\alpha(\boldsymbol{s}) + \beta(\boldsymbol{s})t + \gamma(\boldsymbol{s})t^2)$, which is a log quadratic trend in the sense that $\ln(m(\boldsymbol{s}, t))$ is quadratic in time. The trend model is fitted for each location. It is very important for crop insurance to see how these $\xi(\boldsymbol{s}, t)$'s are correlated for a given year. We assume that the $\xi(\boldsymbol{s}, t)$ is second order stationary at each year with a variance, $\sigma_t^2$, across space.

### Correlograms

Based on the detrended yields, $\xi(\boldsymbol{s}, t)$, we calculated the covariogram and correlogram for each crop at each year from 1972 to 1997 using equations (6) and (8). Different directions are calculated and the results suggest it is reasonable to assume isotropy. Hence, reported in this paper are the isotropic correlogram, which is a function of the distance, calculated as follows:

$$\hat{\rho}_t(h) = \frac{1}{|N(h)|\hat{\sigma}_t^2} \sum_{(i,j) \in N(h)} (\xi(\boldsymbol{s}_i, t) - \bar{\xi}_t)(\xi(\boldsymbol{s}_j, t) - \bar{\xi}_t), \ h > 0,$$

where $N(h)$ is the set of all pairs $(\boldsymbol{s}_i, \boldsymbol{s}_j)$ such that the distance $\|\boldsymbol{s}_i - \boldsymbol{s}_j\|$ is $h$, $|N(h)|$ is the number of elements in $N(h)$, $\bar{\xi}_t$ is the sample mean of $\xi(\boldsymbol{s}_i, t)$ for year $t$, and $\hat{\sigma}_t^2$ is the estimated variance of $\xi(\boldsymbol{s}, t)$ for year $t$.

There are altogether 78 correlograms for the three crops and 26 years. For each crop, the correlograms of the 26 years show a consistent pattern. From these correlograms we

11

see that the positive dependency quickly dies off when the distance increases, i.e., the crop yields show a clear pattern of f.r.p.d. The distance for the positive dependency is at most $3 \times 10^6$ feet, or 570 miles, and in many instances, the range is much smaller than 570 miles. We only present the correlograms for year 1981, year 1991, and the average of 26 years for each of the three crops in Figure 1. Do the correlations die off sufficiently quickly to allow for effective risking pooling? We will propose a measure of effectiveness of risk pooling next.

## The Effectiveness of Risk Pooling

In this section, we introduce a measure of effectiveness for risk pooling, and then calculate the premium loading factor for each of the three US crops: wheat, soybeans and corn.

### *A measure of risk pooling effectiveness*

As we pointed out before, a critical ingredient for risk pooling is that the sample variance decreases as $N$ increases. One way to measure the effectiveness of risk pooling is to see how small the variance of $\bar{X}_t$ is compared with the the average variance of each individual $X(\boldsymbol{s}, t)$, i.e., $Var(\bar{X}_t)/(\sum_{i=1}^{N} Var(X(\boldsymbol{s}_i, t))/N)$. The ratio, denoted by $\phi_t$ for year $t$, will be called the coefficient of effectiveness and can be expressed in terms of the covariances or correlations:

$$
\begin{aligned}
\phi_t &= \frac{Var(\bar{X}_t)}{\sum_{i=1}^{N} Var(X(\boldsymbol{s}_i, t))/N} \\
&= \frac{1}{N \sum_{i=1}^{N} Var(X(\boldsymbol{s}_i, t))} \left( \sum_{i=1}^{N} Var(X(\boldsymbol{s}_i, t)) + \sum_{i=1}^{N} \sum_{j \neq i}^{N} Cov(X(\boldsymbol{s}_i, t), X(\boldsymbol{s}_j, t)) \right).
\end{aligned}
$$

Clearly, if the losses are independent, then $\phi_t = 1/N$. When the losses are positively correlated, $\phi_t$ will be bigger than $1/N$. In addition, if the $X(\boldsymbol{s}_i, t)$'s have the same variance, then

$$
\phi_t = \frac{1}{N^2} \left( N + \sum_{i=1}^{N} \sum_{j \neq i}^{N} Corr(X(\boldsymbol{s}_i, t), X(\boldsymbol{s}_j, t)) \right). \tag{12}
$$

The correlation coefficient $Corr(X(\boldsymbol{s}_i, t), X(\boldsymbol{s}_j, t))$ obviously depends on the prespecified levels as well as the joint distribution of yields $Y(\boldsymbol{s}_i, t)$ and $Y(\boldsymbol{s}_j, t)$. So far we have made no distributional assumptions on yields $Y(\boldsymbol{s}, t)$ other than stationarity. To get a feeling of how the losses $X(\boldsymbol{s}, t) = \max(0, c(\boldsymbol{s}, t) - Y(\boldsymbol{s}, t))$ are correlated, let us now assume yields

12

$Y(\boldsymbol{s}, t)$ are normal and that each of the prespecified levels $c(\boldsymbol{s}, t)$ is one standard deviation below the the mean of $Y(\boldsymbol{s}, t)$, i.e., $c(\boldsymbol{s}, t) = \mu(\boldsymbol{s}, t) - \sigma(\boldsymbol{s}, t)$. Then

$$X(\boldsymbol{s}, t) = c(\boldsymbol{s}, t) - \mu(\boldsymbol{s}, t) - \sigma(\boldsymbol{s}, t) \min \left( \frac{c(\boldsymbol{s}, t) - \mu(\boldsymbol{s}, t)}{\sigma(\boldsymbol{s}, t)}, \ \frac{Y(\boldsymbol{s}, t) - \mu(\boldsymbol{s}, t)}{\sigma(\boldsymbol{s}, t)} \right)$$
$$= c(\boldsymbol{s}, t) - \mu(\boldsymbol{s}, t) - \sigma(\boldsymbol{s}, t) \min \left( -1, \ Z(\boldsymbol{s}, t) \right)$$

where $Z(\boldsymbol{s}, t) = (Y(\boldsymbol{s}, t) - \mu(\boldsymbol{s}, t))/\sigma(\boldsymbol{s}, t)$ is a standard normal random variable. Hence, the correlation between the truncated losses can be expressed as the correlation between truncated standard normal random variables, i.e., $Corr(X(\boldsymbol{s}_i, t), \ X(\boldsymbol{s}_j, t)) = Corr(\min(-1, Z(\boldsymbol{s}_i, t)), \min(-1, Z(\boldsymbol{s}_j, t))$. This correlation is clearly a function of

$$Corr(Z(\boldsymbol{s}_i, t), \ Z(\boldsymbol{s}_j, t)) = Corr(Y(\boldsymbol{s}_i, t), \ Y(\boldsymbol{s}_j, t)) = \rho_t(\|\boldsymbol{s}_i - \boldsymbol{s}_j\|).$$

Let us write $g(c_1, c_2, \rho) = Corr(\min(c_1, Z_1), \ \min(c_2, Z_2))$ for two standard normal variables $Z_1$ and $Z_2$ that are also bivariate-normal with a correlation coefficient $\rho$. Then

$$g(c_1, c_2, \rho) = Corr(\min(c_1, Z_1) - c_1, \ \min(c_2, Z_2) - c_2)$$
$$= \frac{1}{2\pi(1 - \rho^2)^{1/2}} \int_{-\infty}^{c_1} \int_{-\infty}^{c_2} (x_1 - c_1)(x_2 - c_2) \exp(-\frac{x_1^2 - 2\rho x_1 x_2 + x_2^2}{2(1 - \rho^2)}) \, dx_1 dx_2$$
$$- \frac{1}{2\pi} \int_{-\infty}^{c_1} (x_1 - c_1) \exp(-x_1^2/2) dx_1 \int_{-\infty}^{c_2} (x_2 - c_2) \exp(-x_2^2/2) dx_2. \tag{13}$$

The integrals cannot be evaluated in closed form, but can be approximated by the Monte Carlo methods. We provide in Table 1 the correlation coefficients between the truncated normals corresponding to different $\rho$s where the standard normal variables are truncated at -0.75, -1 and -1.5. Our results were based on 500,000 simulations. We see that the truncated variables are significantly less correlated.

Intuitively, the truncated variables should be less correlated. Above the truncating values, the two truncated variables are constants and hence uncorrelated. Indeed, if the truncated values are far below the mean, the two truncated variables are constants with a probability close to one and hence are nearly uncorrelated. Note Table 1 was obtained for two standard

normals. If the two variables are not normal, we might expect the truncated variables are still less correlated. But the exact correlation coefficient of the truncated variables will depend the joint distribution of the two untruncated variables.

Recently, the Risk Management Agency of USDA allows farmers in the MPCI program to specify a coverage level $c(\boldsymbol{s}, t)$ up to 75% of the mean yield $\mu(\boldsymbol{s}, t)$, which is more than one standard deviation below the mean because the coefficient of variation for crops is usually less than $1/4$. Although farmers can select their own coverage levels, let us assume nonetheless that all farmers select their coverage levels to be one standard deviation below the expected yields, i.e., $c(\boldsymbol{s}, t) = \mu(\boldsymbol{s}, t) - \sigma(\boldsymbol{s}, t)$, and see how effective the risk pooling is. In this case, the losses have the same variance and (12) is applicable. If some farmers select coverage levels that are more than one standard deviations below the means, the losses only become less correlated according to what we found out from Table 1. We will continue to assume yields are normal. Since

$$Corr(X(\boldsymbol{s}_i, t),\ X(\boldsymbol{s}_j, t))$$
$$= Corr(\min c(-1, Z(\boldsymbol{s}_i, t)),\ \min(-1, Z(\boldsymbol{s}_j, t)))$$
$$= g(-1, -1, \rho_t(\|\boldsymbol{s}_i - \boldsymbol{s}_j\|)),$$

then by (12)

$$\phi_t = \frac{1}{N^2}\left(N + \sum_{i=1}^{N}\sum_{j\neq i}^{N} g(-1, -1, \rho_t(\|\boldsymbol{s}_i - \boldsymbol{s}_j\|))\right) \tag{14}$$

where $\rho_t(\cdot)$ is the covariogram of $Y(\boldsymbol{s}, t)$ and can be estimated from the data. Replacing $\rho(\|\boldsymbol{s}_i - \boldsymbol{s}_j\|)$ by the estimated covariogram $\hat{\rho}(\|\boldsymbol{s}_i - \boldsymbol{s}_j\|)$, we get an estimate for $\phi_t$.

Using the average correlogram of 26 years, we calculated the coefficient of effectiveness $\phi$ for US wheat, soybeans and corn by applying (14) with $\rho_t$ being replaced by the average correlogram. The coefficients of effectiveness are 0.00188, 0.00479 and 0.00320 for US wheat, soybeans and corn respectively. For each crop, we used all of the counties in our database. Hence $N = 2,641, 2,000$ and 2,591 for wheat, soybeans and corn. These coefficients mean, if

an insurer can pool all the growing areas in the US and each unit specifies the coverage level to be one standard deviation below its mean yield, over the 26 years the variance of average loss is only 0.188%, 0.479% and 0.32% of the variance of each individual loss. Or equivalently, to make the risk pooling for the correlated losses to be as effective as independent losses, the number of correlated units should be approximately 4.97 times, 9.58 times and 8.3 times as large as the number of uncorrelated units for wheat, soybeans and corn respectively, when all units select coverage levels to be one standard deviation below the mean yields. Choosing the coverage level to be one standard deviation below the mean would mean to insur against adverse events that occur about 15.9% of the time. If the units select lower coverage level, the losses will be less correlated and these factors will become smaller. These results show that, if a large number of counties buy the insurance, effective risk pooling is highly possible at least in the case that yields are normal. Moderate departure from normal distributions should not change the coefficients of effectiveness very much.

### *Premium loading*

Due to the weak correlation between the losses, the central limit theorem can be applied. Therefore, if the insurer's probability of ruin is set at $\alpha$, the premium loading for each insured at year $t$ is $L_t = z_\alpha (Var(\bar{X}_N))^{1/2} = z_\alpha \sqrt{\phi_t} \left( \sum_{i=1}^{N} Var(X(\boldsymbol{s}_i, t))/N \right)^{1/2}$, assuming the administrative cost load is zero. The variance of a loss is bounded by the variance of yield regardless of the distribution of yield, as shown in the Appendix. Hence $L_t \leq z_\alpha \sigma_t \sqrt{\phi_t}$.

The 26-year average variance, $\hat{\sigma}^2$, is estimated from the previous section as 52, 20, and 160 for wheat, soybeans and corn, respectively. Using $z_{0.025} = 1.96$ and the $\phi$s calculated in the previous subsection, we obtain the premium loading factor to be at most 0.6128, 0.6066 and 1.4024 bushels per acre for the three crops, respectively. In terms of dollars, if the prices are $3.5, $4.5 and $3 per bushel for wheat, soybeans and corn respectively, the corresponding premium loadings are $2.14, $2.72 and $4.20 per acre, which are quite moderate.

### Summary and Discussion

In this paper, we have reviewed the statistical foundations for crop insurance under dependence, and have taken a close look at the correlation structure between crop yields at the county level in the U.S. Using some techniques in spatial statistics, we have shown that the positive correlation between any two counties dies off quickly when the lag distance increases, a phenomenon of finite-range positive dependence. The yield losses, as truncated variables, are also f.r.p.d. and the correlation of losses dies off much faster. We provided a general method for investigating the effectiveness of risk pooling under dependence. Our method employs spatial statistics and is appropriate for crop insurance. The results indicate that risk pooling can be effective for the three crops (wheat, corn and soybeans) if the coverage levels are at least one standard deviation below the mean yields, hence suggest a high possibility of a viable private agricultural insurance market, if the insurance/reinsurance markets can cover most growing area in the US. These results may help alleviate the pressure on government to assume total responsibility of providing crop insurance.

Although in this work we have applied this method to county level yields, it is also applicable to individual farms if the farms' geographical locations and yields are known. With the availability of Global Position System, farm locations can be obtained easily. Insurance agencies should also have historical farm level yield records cumulated from farmers' years of enrollment in MPCI, and these farm yield data can be analyzed in the same way as the county yields. With the absence of individual farm data, it is impossible to examine the effectiveness of crop insurance at the individual farm level. However, we can show that if growers are scattered nearly uniformly in all the counties, the correlogram of individual farm yields is less than the correlogram of county average yields and hence dies off faster. How are the two coefficients of effectiveness, one based on individual farm data and the other on county data, compared with each other can only be examined under more assumptions, which is beyond the scope of this paper.

There are several open problems along this line for future research, one of which is the

temporal effect on risk pooling. The weather pattern is generally assumed to be independent and so are the yields of a farm over time. Pooling risks over time may further reduce the premium loading.

Note the buffer fund given as $E(S_N(t)) + 2(Var(S_N(t))^{1/2}$ is the upper bound. When the yield loss is finite-range dependent (or finite-range positive dependent with additional conditions to ensure mixing conditions), the actual total loss $S_N(t)$ is asymptotically normal. This implies that when $N$ is sufficiently large, $S_N(t)$ will be symmetric. Hence $S_N(t)$ has approximately a 50% chance to be below its mean $E(S_N(t))$. In the case that the actual total loss is below its mean, there will be a surplus in the buffer fund. Accumulation of such surpluses will reduce the risk of pooling, and its effect on risk pooling warrants further studies. There are other avenues that private insurers can do to reduce the pooled risk. For example, trading in international reinsurance market and diversification over crops may help risk pooling.

Due to the spatial dependence, more exposure units/farms are needed to equate pooling of dependent risks to independent risks. There might be a deterrent effect of the larger pool size on a private crop insurer seeking to enter the private crop insurance market. A study on the profitability of a private crop insurance will help find out how strong the deterrent effect may be.

## References

Besag, J. and D. Higdon. "Bayesian Analysis of Agricultural Field Experiments" (with discussion), Journal of the Royal Statistical Society, Ser. B, 61 (1999): 691-746.

Bolthausen, E. (1982) "On the C.L.T. for stationary mixing random fields." Annals of Probability, 10 (1982), 1047-1050.

Borch, K. The Mathematical Theory of Insurance: An Annotated Selection of Papers on Insurance published 1960-1972. Lexington, MA: D.C. Heath, 1974.

Bühlmann, H. Mathematical Methods in Risk Theory. New York: Springer-Verlag, 1970.

Cressie, Noel. Statistics for Spatial Data (Revised Edition), Wiley and Sons, Inc., New York, 1993.

Cummins, D.J. "Statistical and Financial Models of Insurance Pricing and the Insurance Firm." The Journal of Risk and Insurance, 58(1991): 261-302.

Davis, B. M. and L.E. Borgman. "A Note on the Asymptotic Distribution of the Sample Variogram." Journal of the International Association for Mathematical Geology, 14(1982):189-193.

Goodwin, B.K., and V.H. Smith. The Economics of Crop Insurance and Disaster Aid. Washington, D.C.: The AEI Press, 1995.

Guyon, X. Random Fields on a Network: Modeling, Statistics and Applications, Springer-Verlag, New York, 1995.

Hennessy, David A., Bruce A. Babcock, and Dermot J. Hayes. "Budgetary and Producer Welfare Effects of Revenue Insurance." American Journal of Agricultural Economics, 79(1997):1024-1034.

Hogg, R.V., and S.A. Klugman. Loss Distributions. New York: John Wiley, 1984.

Journel, A. G. and C. J. Huijbregts. Mining Geostatistics. Academic Press, London, 1985.

Knight, T.O., and K.H. Coble. "Survey of U.S. Multiple Peril Crop Insurance Literature Since 1980." Review of Agricultural Economics, 19(1997): 128-156.

Mehr, R.I., E. Cammack, and T. Rose. Principles of Insurance, 8th ed. Homewood, Illinois: Richard D. Irwin, Inc., 1985.

Mercer, W. B. and A.D. Hall "The Experimental Error of Field Trials", Journal of Agriculture Sciences (Cambridge), 4(1911):107-132.

Miranda, M.J. "Area-Yield Crop Insurance Reconsidered." American Journal of Agricultural Economics, 73 (1991): 233-242.

Miranda, M.J. and J.W. Glauber. "Systematic Risk, Reinsurance, and the Failure of Crop Insurance Markets." American Journal of Agricultural Economics, 79 (1997): 206-215.

Roberts, R. K., B. C. English, and S.B. Mahajanashetti. "Evaluating the Returns to Variable Rate Nitrogen Application ." Journal of Agricultural and Applied Economics, 32(April 2000): 133-143.

Roy, A. D. "Safety First and the Holding of Assets." Econometrica, 20(1952):431-449.

Skees, J.R., J. Harwood, A. Somwaru, and J. Perry. "The Potential for Revenue Insurance Policies in the South." Journal of Agricultural and Applied Economics, 30(July 1998):47-61.

Skees, J.R., J.R. Black, and B.J. Barnett. "Designing and Rating an Area Yield Crop Insurance Contract." American Journal of Agricultural Economics, 79(May 1997):430-438.

Telser, L. "Safety First and hedging." Review of Economic Studies, 23(1956):1-16.

Turvey C. G. "Weather Derivatives and Specific Event Risk." Selected Papers, 1999 AAEA Annual Meetings, Nashville, TN, August, 1999.

Wang, H. H., S. D. Hanson, R. J. Myers, and J. R. Black. "The Effects of Alternative Crop Insurance Designs on Farmer Participation and Welfare." American Journal of Agricultural Economics, 80 (November 1998):806-820.

Zhang, H. "On Estimation and Prediction for Spatial Generalized Linear Mixed Models." Biometrics, 58 (March 2002), 129-136
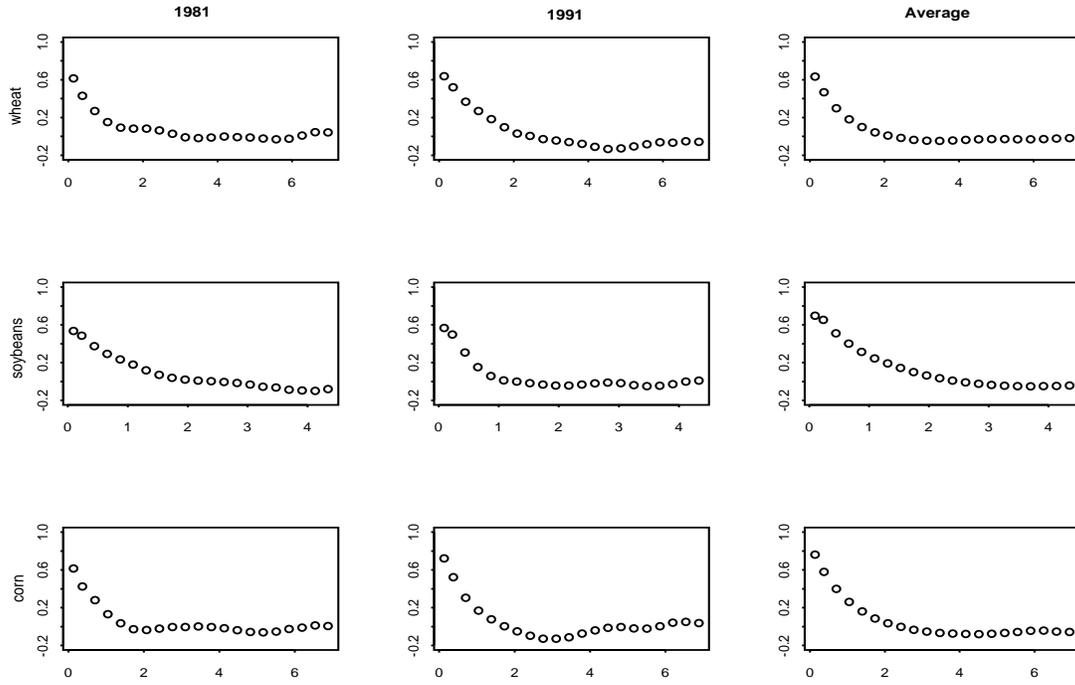
Figure 1: Correlograms of yields of wheat (top row), soybeans (middle row) and corn (bottom row); for 1981 (left), 1991 (middle) and the average of 26 years from 1972 to 1997.

| $\rho$ | -0.2 | -0.1 | 0.1 | 0.2 | 0.3 | 0.4 | 0.5 | 0.6 | 0.7 | 0.8 | 0.9 |
|---|---|---|---|---|---|---|---|---|---|---|---|
| $c_1 = c_2 = -0.75$ | -0.077 | -0.04 | 0.053 | 0.11 | 0.18 | 0.26 | 0.34 | 0.45 | 0.56 | 0.68 | 0.83 |
| $c_1 = c_2 = -1$ | -0.059 | -0.032 | 0.04 | 0.091 | 0.15 | 0.22 | 0.31 | 0.4 | 0.52 | 0.66 | 0.81 |
| $c_1 = c_2 = -1.5$ | -0.027 | -0.017 | 0.022 | 0.058 | 0.1 | 0.17 | 0.23 | 0.33 | 0.44 | 0.58 | 0.77 |

Table 1: The correlation coefficient of truncated standard normal random variables: $g(c_1, c_2, \rho)$ defined by (13)

# Appendix

Here we show the following theorem that implies that variance of loss is bounded by the variance of the yield.

**Theorem 1** *Let $Y$ be any random variable with mean $\mu$ and variance $\sigma^2$, and $X = \max(0, c-Y)$ for some constant $c < \mu$. Then $Var(X) \le \sigma^2$.*

**Proof.** First note that

$$X = c - \mu - \sigma \min\left(\frac{c - \mu}{\sigma}, \frac{Y - \mu}{\sigma}\right).$$

Then

$$Var(X) = \sigma^2 Var\left(\min\left(\frac{c - \mu}{\sigma}, \frac{Y - \mu}{\sigma}\right)\right) = \sigma^2 Var(\min(d, Z))$$

where $d = (c - \mu)/\sigma < 0$ and $Z = (Y - \mu)/\sigma$. Then $Var(Z) = 1$, and if $f_Z(z)$ denotes the density function of $Z$,

$$Var(\min(d,\ Z)) = Var(\min(d,\ Z) - d)$$

$$\le E[(\min(d,\ Z) - d)^2] = \int_{-\infty}^{d} (z - d)^2 f_Z(z)\, dz.$$

Since $z < d < 0$, then $z < z - d < 0$ and $z^2 > (z - d)^2$. The integral is bounded by

$$\int_{-\infty}^{d} z^2 f_Z(z)\, dz \le \int_{-\infty}^{\infty} z^2 f_Z(z)\, dz = Var(Z) = 1.$$

The proof is finished.