Journal of Agricultural and Resource Economics 39(2):253–270 ISSN 1068-5502

Exploiting the Relationship between Farm-Level Yields and County-Level Yields for Applied Analysis

Scott Gerlt, Wyatt Thompson, and Douglas J. Miller

County-level yield data are used in applied research and crop insurance policy in place of farm-level yield data, which are likely sparse, not broadly representative, and subject to selection bias. We exploit the fact that county-level yields are the aggregate of farm-level yields to derive bounds that can be reduced to direct relationships between county- and farm-level yields under certain conditions. Simulation experiments indicate that crop insurance premium estimates derived from this method have the potential for bias in certain conditions but are reasonably precise in other conditions, suggesting that these relationships are a new tool for applied analysts.

Key words: agricultural risk, crop insurance, crop yield distributions

Introduction

Farm-level yield data have become a primary factor in agricultural policy and crop insurance rating. However, researchers often have trouble analyzing these programs because existing farm-level yield data are sparse, might not be broadly representative, and might suffer from selection bias. Consequently, the basis for decision making is less firm, potentially leading to erroneous conclusions. National Agricultural Statistics Service (NASS) data represent many county-crop combinations and often have long time-series, particularly in key growing regions, providing researchers a foundation for empirical assessment of primary yield distributions. These data and the research literature based on them have been used as a basis for making decisions, including crop insurance and farm program analysis.

United States support for agricultural commodity producers has shifted to crop insurance, and this trend seems set to continue under the latest farm bill. Crop insurance subsidies amounted to $0.2 billion (in 2009 dollars) on $14.3 billion in crop value in 1991 and $5.4 billion in subsidies on $79.6 billion in crop value in 2009, whereas program payments to U.S. farms summed to about $12.3 billion in 2009 (and were as high as $24 billion in 2005) (White and Hoppe, 2012). The 2012 drought decreased the national average corn yield to just over 120 bushels per acre from predrought expectations of 166 bushels per acre (Office of the Chief Economist, 2012), highlighting the risk that crop insurance policies are intended to mitigate. The 2014 Farm Bill will continue to shift the focus of agricultural policy toward crop insurance and farm revenue protection programs, terminating direct payments altogether. Some of the proposed crop insurance programs provide options to farmers to choose coverage based on county- or farm-level yields.

Scott Gerlt is a research associate, Wyatt Thompson is an associate professor, and Douglas J. Miller is an adjunct associate professor in the Department of Agricultural and Applied Economics at the University of Missouri at Columbia.
Contributions by Gerlt and Thompson are supported in part by the Agriculture and Food Research Initiative Competitive Grant no. 2012-68002-19872 from the USDA National Institute of Food and Agriculture (NIFA). However, any findings and views expressed are the authors' own, and might not reflect those of USDA or NIFA.

Review coordinated by Christopher S. McIntosh.

Analysts trying to estimate the impacts of proposed farm bills must undertake ex ante assessment of crop insurance and similar policies. County-level yield data might be the starting point for such assessments and are appropriate for policies or policy options that operate at that level. However, analysts might also use county-level data for programs that operate at the individual level to make the problem tractable. Insurance options with a county-level yield trigger, like the U.S. Group Risk Plan, raise related questions about how county-level yield insurance relates to farmers' individual risks and, consequently, their participation rates. While RMA USDA Risk Management Agency (2010, 2011) collects and generally uses farm-level data, information supplied by farmers in the course of an insurance program might suffer from selection bias, and historical data for any particular crop on any particular farm might be too sparse, disallowing strong conclusions about the nature of farm-level crop-yield distributions.

Research has identified and addressed several potential challenges relating to accurate estimation of actuarially sound premiums based on county-level yield data. Miranda (1991) develops an innovative model that disaggregates farm-level yield variability into systemic and idiosyncratic components, elaborates propositions linking model parameters to crop insurance choices and outcomes, and simulates the model for farm-level data representing 102 soybean farms in Kentucky to calculate critical values of parameters and optimal premium rates. A key element of the Miranda model is a farm-specific parameter, beta, that governs the extent to which the systemic component apparent in the regional yield deviation from regional mean is reflected in any particular farm's deviation from its own mean.

The Miranda model has become a workhorse for crop insurance and farm-level policy studies. Researchers typically start with an assumption about the beta that defines systemic risk in farm-level yield variability—the key link from county-level data to farm-level data—and end with a simulation model to highlight the impact on program operation. Coble and Dismukes (2008) develop a representative farm using a beta equal to 1 and with yield variation calibrated to observations from farms participating in federal programs. They then test the implications of various commodity revenue programs. Cooper et al. (2012) use county data to augment crop insurance premium determination. Coble and Barnett (2008) allow the farm-specific beta to vary over an assumed normal distribution in their investigation of whether the addition of area revenue risk insurance program might account for the systemic component of farm risk and leave private insurers to handle idiosyncratic risk. Carriquiry, Babcock, and Hart (2008) estimate the beta parameter for nine crop reporting districts, then use a simulation model to calculate impacts of the estimated parameter on crop insurance programs. Coble, Heifner, and Zuniga (2000) start with a similar assumption regarding the link from county- to farm-level yield distribution, then go on to prove that the average farm yield variance in a county equals the county yield variance plus the average variance of farm-to-county yield differences. This framework is exploited to calculate optimal futures and put ratios under four alternative insurance programs. Barnett et al. (2005) compare regional and individual insurance programs with empirical application to farm-level yield data for tens of thousands of corn farms and thousands of sugar farms, but they are unable to estimate the farm-specific beta of Miranda's specification over data representing ten years.

While some researchers use county-level data to represent farm-level outcomes (e.g. Zulauf, Dicks, and Vitale, 2008), others have used a variety of ways to account for potential differences. Schnitkey, Sherrick, and Irwin (2003) rescale county data using farm records, and Goodwin (2009) added a shock to state yields to account for the extra variability at the farm level. Bulut, Collins, and Zacharias (2012) use a theoretical optimization framework to show that farmers prefer actuarially fair individual insurance to actuarially fair area insurance, but their choices will change if one or the other is provided free. Deng, Barnett, and Vedenov (2007) estimate a kernel function for farm-level yield distributions over selected RMA data supplemented by NASS county data and find that regional-level insurance might be preferred over individual insurance if the premium is substantially lower. Cooper (2010) assesses the effects of the Average Crop Revenue Election program, under

which payments are partly governed by regional triggers, on the distribution of farm revenue and finds increases in values at the low-end of this distribution.

Other researchers have investigated the distribution of crop yields. Goodwin (2001) shows that the correlation structure may be state contingent and vary with climate conditions. Just and Weninger (1999) stress certain problems, including the risk of understating yield variation if aggregated data are used to represent farm-level yield variation. Some researchers find the normal distribution adequately fit their data, some rule out the normal distribution, and some sidestep the problem by applying a nonparametric density function (Atwood, Shaik, and Watts, 2003; Claassen and Just, 2011; Goodwin and Ker, 1998; Harri et al., 2011; Just and Weninger, 1999; Ker and Coble, 2003; Ker and Goodwin, 2000; Koundouri and Kourogenis, 2011; Ramirez, Misra, and Field, 2003; Sherrick et al., 2004; Tack, Harri, and Coble, 2012). Claassen and Just (2011) focus specifically on the regional or systemic component of farm-level yields in RMA data, finding strong correlation within county and significantly higher farm-level yield variability.

Two general conclusions can be drawn from the yield-distribution literature. First, uncertainty remains about the best way to represent the farm-yield distribution, partly because county-level data often must be used in place of farm-level data. This suggests that no particular distribution should be considered definitively and universally applicable to crop yields. Second, some of these studies create a useful method of measure by analyzing the implications of various representations of data in terms of their implications for crop insurance programs.

A challenge that persists in the crop insurance, policy analysis, and crop yield literature is based on the difficulty of applying county-level yield data with farm-level yield distributions. While some studies estimate the link for specific commodity-county combinations, we are aware of no study that has so wide a scope that the results can safely be extrapolated to analyze or support implementation of federal programs that are available to the vast majority of crop farms. If researchers seek to inform policy makers by relying on quick assessments of proposals during the course of a debate, then narrowly focused studies or studies based on past programs rather than proposed programs might be of limited use.

We offer a new approach to this problem that builds from the definition of county data as the sum of its parts and some stylized facts about farms within the county in order to provide a tool for using county-level data for farm-level crop insurance program assessment. Whereas some studies have drawn on the fact that farm-level yields aggregate to county-level yields to supplement analysis (Barnett et al., 2005; Coble, Heifner, and Zuniga, 2000; Just and Weninger, 1999), we begin from this relationship to see what inferences about individual crop insurance can be drawn from county-level yield data. Our approach requires no assumption about the distribution, and the method can be applied regardless of the distribution of farm-level yield and even if distributions and interfarm correlation are not constant or vary under different weather conditions (Goodwin, 2001; Hennessy, 2009). There is some scientific knowledge about the relationships between individual observations and aggregate statistical measures of the distribution, but to the best of our knowledge no effort has been made to develop general guidelines that are less dependent on the underlying distribution or to apply these relationships to critically important questions relating to crop insurance.

A key conclusion to our work is that (to the extent that each individual farm represents an insignificant share of county production and yield variances are unrelated to interfarm yield correlations) the relationship between the county yield distribution and component farm distributions reduces to a tractable problem that is amenable to *ex ante* policy analysis or risk premium calculations.

Mathematical Representation (proofs available in appendices)

County yield is an aggregation of yields of farms located in the county, so county yield variance contains useful information about underlying farm variances. Consider a county, c, with n farms

with yields $\{x_1, x_2, \ldots, x_n\}$, each with a w_i fraction of acres in the crop for the county. By definition,

$$\sigma_c^2 = Var\left(\sum_{i=1}^{n} w_i x_i\right) = \sum_{i=1}^{n}\sum_{j=1}^{n} \rho_{i,j} w_i w_j \sigma_i \sigma_j, \tag{1}$$

where $Var(c) = \sigma_c^2$, $Var(x_i) = \sigma_i^2$, $Var(x_j) = \sigma_j^2$, $\rho_{i,j} = Corr(x_i, x_j)$, and $\sum_{i=1}^{n} w_i = 1$. Because the actual parameters are unobserved, let $\rho_{i,j}$, σ_i, and σ_j be treated as random variables such that $E[\rho_{i,j}] = \rho_c$ and $E[\sigma_i] = E[\sigma_j] = \sigma_f$. The expected interfarm correlation, ρ_c, is calculated as[1]

$$\rho_c = \frac{\sum_{i=1}^{n}\sum_{j\neq i}^{n} \rho_{i,j} w_i w_j}{\sum_{i=1}^{n}\sum_{j\neq i}^{n} w_i w_j} = \frac{\sum_{i=1}^{n}\sum_{j\neq i}^{n} \rho_{i,j} w_i w_j}{1 - \sum_{i=1}^{n} w_i^2}. \tag{2}$$

The expected farm standard deviation in the county, σ_f, can be expressed as

$$\sigma_f = \sum_{i=1}^{n} w_i \sigma_i. \tag{3}$$

These properties can be used to develop a relationship between county variance and farm yield standard deviations, namely

$$\begin{aligned} \sigma_c^2 \quad = \quad & (\sigma_f)^2 \rho_c - (\sigma_f)^2 \rho_c \sum_{j=1}^{n} w_j^2 + (\sigma_f)^2 \sum_{j=1}^{n} w_j^2 + \\ & \sum_{j=1}^{n} w_j \sigma_j Cov(\rho_{ij|j}, \boldsymbol{\sigma}) + \sigma_f \sum_{j=1}^{n} w_j Cov(\rho_{ij|j}, \boldsymbol{\sigma}), \end{aligned} \tag{4}$$

where $Cov(\rho_{ij|j}, \boldsymbol{\sigma})$ is the covariance between the jth row(column) of the correlation matrix and the vector of farm standard deviations, $\boldsymbol{\sigma}$. This relationship depends partly on shares of each farm's area in the county total and correlation among farms as well as the underlying standard deviations.[2] To our knowledge, this is the first time the relationship between aggregated random variables has been written in this form. The following propositions and lemmas are also the first known representations of the properties, unless otherwise noted.

(Proposition 1:)
$$\rho_c \in \left[-\frac{\sum_{i=1}^{n} w_i^2}{1 - \sum_{i=1}^{n} w_i^2}, 1\right]$$

This proposition establishes a range of feasible values for the average interfarm correlation within a county. As an aggregate of its parts, the county average correlation cannot be too negative, where shares of individual farms define exactly where that lower limit lies. On the other hand, it is theoretically possible that a county could have an interfarm correlation of 1 if all farm yields move alike. This result is independent of the distribution of farm yields.[3]

The lower limit of this proposition does not map directly to the Miranda method of yield decomposition into systemic and idiosyncratic components. The Miranda approach suggests that some fundamental factors drive all county yields alike, whereas our lower bound on average interfarm correlation is derived from farm-to-county aggregation and is simply an unavoidable lower bound irrespective of the fundamental factors. However, researchers might take from this proposition support to assume at least some value, depending on farm shares, for the beta parameter that otherwise represents systemic risk in applied work based on the Miranda model.

[1] For discrete random variables, the population mean is equal to the expectation.

[2] Due to Jensen's inequality, the square of the average farm standard deviation does not necessarily equal the average farm variance. The former will be larger if there is more than one farm; otherwise they are equal.

[3] If the distributions are known, Frechet-Hoeffding bounds can be used to bound the individual farm correlations and might imply an even tighter bound for the average (De Veaux, 1976).

Policy makers might make use of the fact that the lower bound on the average correlation among farm yields within a county depends on farm shares, as well as the lack of upper bound short of perfect correlation. While average effects might not be the sole goal of policies, placing bounds on average interfarm correlation within the county might already be a step forward. Moreover, a defined range of average interfarm correlation might be a useful tool for assessing regional insurance program options.

(**Lemma 1:**) $$\sigma_c^2 \in \left[0, (\sigma_f)^2\right]$$

County variance cannot exceed expected farm standard deviation squared. This finding is to our knowledge the first formal statement of this property, although it has been previously conjectured (Tack, Harri, and Coble, 2012). Furthermore, this lemma based on aggregation reinforces previous research, including empirical findings that suggest farm-level yield variation is greater than in county-level yield variation. Like any variance, county variance cannot be negative but can theoretically be 0. In practice, given that farms within a county have close spatial proximity, we would not expect all the conditions to be met such that the county has a variance of 0 (for example, only two farms in the county with identical share-weighted variances and perfect negative correlation). Therefore, it seems extremely unlikely that the theoretical lower bound would ever be observed.

(**Lemma 2:**) $$\sigma_c^2 \to (\sigma_f)^2\rho_c + E[\sigma_j Cov(\rho_{ij|j}, \sigma)] + \sigma_f E[Cov(\rho_{ij|j}, \sigma)] \text{ as } w_i \to 0 \; \forall \; i$$

The county variance approaches the product of the average farm standard deviation squared and the average interfarm correlation, plus a deviate based on the relationship and size of the standard deviations of the farm variances and correlations as individual farm shares decline. If the farm variances are homoskedastic or the interfarm correlations are constant, then the deviate approaches 0 if farm shares approach 0. Conversely, under very extreme assumptions about farm variances, the deviate has a theoretical maximum value of $(\sigma_f)^2$.

This proposition relates to the case that the weights approach 0, but there may be lessons for applied research nonetheless. The proposition is relevant provided there are not a handful of farms that account for the near entirety of production within the county, which is already a prerequisite of NASS reporting data from a county. The proposition does not require homogeneous farm size, but if the farms were homogeneous with regard to production, then $\sum_{j=1}^{n} w_j^2$ would equal $1/n$. In that case, with just twenty farms growing a crop in a county, the $\sum_{j=1}^{n} w_j^2$ term in the identity would equal just 0.05. Referring to the proof, the deviate would not drop out completely, as in the case that shares approach 0, but the importance of this term would be diminished and might be ignored for certain calculations.

(**Lemma 3:**) $$\sigma_c^2 \in \left((\sigma_f)^2\rho_c, (\sigma_f)^2\right] \text{ if } Cov(\rho_{ij|j}, \sigma) = 0 \; \forall \; i, j$$

This lemma starts from the case that the correlation among farm yields does not vary with the farm standard deviation. (That is to say, the correlation between yields on any two farms does not tend to be higher or lower as the yield standard deviations of those farms rises.) Absent such a relationship, county-level yield variance cannot be less than or equal to the product of average interfarm correlation and average farm standard deviation squared, and the county-level yield variance cannot be greater than the average farm standard deviation squared. As the individual shares of production decreases, the county-level variance decreases toward the lower bound, $(\sigma_f)^2\rho_c$, provided farm-to-farm yield correlation is not related to farm-level yield variance. In the case that there is one farm producing the crop in the county, then the county variance will equal the farm variance, as expected.

This proposition depends on an assumption that might not be very limiting in practice and has been used before in other fields. For example, Spearman (1910) and Brown (1910) use homoskedasticity, which is an instance of the assumption, with equation (1) to derive the

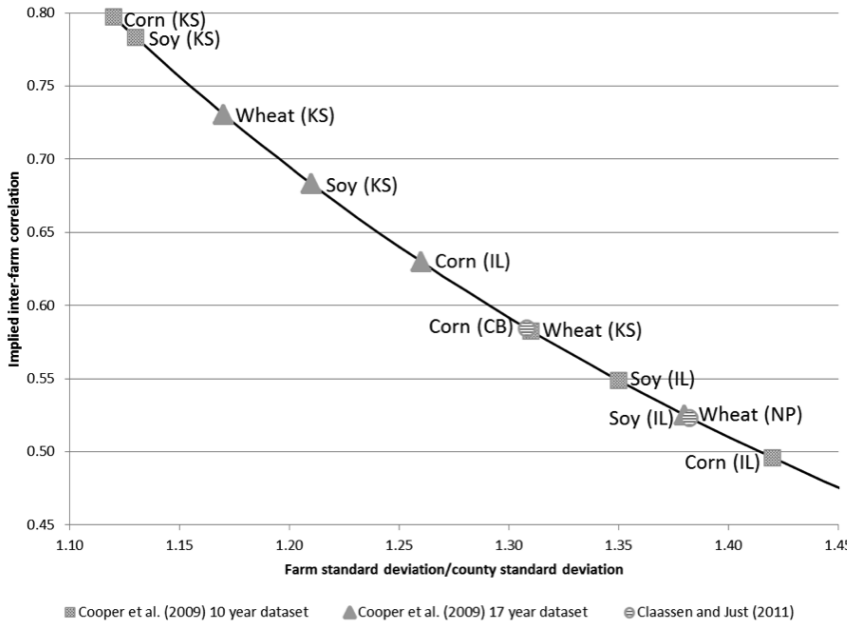

Figure 1. Empirical Studies of Farm to County Standard Deviation (X-Axis) and the Implied Average Interfarm Correlation Based on Proposition 2 (Y-Axis).

Notes: $\rho_c \approx \frac{1}{(\sigma_f/\sigma_c)^2}$ according to proposition 2.

Spearman-Brown prediction formula. Moreover, constant interfarm correlation, in addition to homoskedasticity, would be consistent with lemma 2 because it also corresponds to $Cov(\rho_{ij|j}, \boldsymbol{\sigma}) = 0$. Given that farms in a county are defined by their spatial proximity, they might be expected to have similar, if not identical, standard deviations and interfarm correlations.

(**Proposition 2:**) $\sigma_c^2 \to (\sigma_f)^2\rho_c$ where $\rho_c \in [0,1]$ if $Cov(\rho_{ij|j}, \sigma) = 0$ as $w_i \to 0 \; \forall \; i$

The county variance approaches the product of the average farm variance and the average interfarm correlation if the conditions of lemmas 2 and 3 hold. Given that (a) farm shares approach 0 and (b) the correlation of each farm's yield with other farms' yields does not relate to the level of the farm's yield variations, then the county-level yield variance relates simply to expected farm-level yield variance and the correlation among those yields.

Going farther, proposition 1 defines the boundary for correlation among farm yields, $\rho_c \in [0,1]$. Combined with the two conditions above, this bounds county-level yield variance between 0 and the average farm variance, or $\sigma_c^2 \in [0, (\sigma_f)^2]$, as shown based on different assumptions in lemma 1. As correlation among farm yields in the county rises, the county-level variance converges to the average farm variance squared, given the other conditions of this proposition.

Only a few studies have empirically estimated the ratio between average farm standard deviation and county standard deviation. Applying proposition 2 to these estimates implies interfarm correlation coefficients (figure 1). For example, Cooper et al. (2009) use data from two data sets of farm records for Illinois and Kansas and the NASS county data to calculate the average ratio for the county. These points are identified by state and crop, with the x-axis from the estimated value and the y-axis correlation based on proposition 2. Claassen and Just (2011) use RMA data for a similar comparison. For this study, the Corn Belt (CB) and Northern Plains (NP) are plotted. The corresponding correlations range from 0.8 to slightly less than 0.5, which conform to *a priori* expectations about strong correlation among farms within a county.

The conditions for proposition 2 do not appear to be problematic, as the range of correlations implied from previous studies proves quite reasonable. However, this is an empirical question for which the error can be measured for any particular case in which farm-level yield data are available. In a study with available data, researchers can test to see whether farm sizes are quite small and whether the correlation among farm yields is related to individual farm yield variances. To the extent that this relationship is weak among many farms in a county, proposition 2 can be validated.

However, farm-level yield data are not always available or easily accessible, nor do they always have a long enough time series for reliable statistical testing. The question we pose is whether using the results of these propositions, both at the boundary conditions and under alternative assumptions, can facilitate research or policy implementation that relies on county-level yield data in place of unavailable or insufficient farm-level yield data. A key question is the risk of error. Simulation experiments address this question in the next section.

Simulations

Simulations suggest that these relationships between county and farm yields have some value in the context of crop insurance and policy analysis and suggest more broad applicability. The simulations start with the farms in a hypothetical county that are characterized by their (1) number and share of production, (2) individual farm yield distributions that are characterized by mean and standard deviation that can be drawn randomly from distributions of their own, and (3) correlation between yields of each pair of farms in the county. The mathematical limits proven above are used to generate estimates of the farm-level yield based on observed county yields. When combined with a specific coverage level, which is characterized as a loss threshold that induces payments, the simulations generate the actuarially fair premium from the farm-level data and also the estimated premium based on county data and the mathematical limits. The simulations also include premiums calculated from unadjusted county data as point of comparison. These premiums give a measure of error that is relevant if the formulas derived mathematically are used in applied analysis and for crop insurance implementation. Another measure that is less relevant to applied work but goes more directly to the underlying approximation is the comparison of the ratio of county-level yield variance to farm-level yield variance, $\sigma_c^2/(\sigma_f)^2$, and correlation among farm variances, ρ_c. Under proposition 2, the ratio of variances should converge to the average correlation.

The simulations start by generating yields for n farms 500 times. The weights, means, standard deviations, and distributions of the farm yields are allowed to vary between scenarios. We impose a correlation matrix of our specification on the farm yields using a method that essentially reorders the results of independent marginal distributions to achieve a specified correlation matrix (Iman and Conover, 1982). This approach is distribution free, which allows the evaluation of different distributional assumptions. Although copulas have gained attention lately, imposing the correlation matrix is the only concern of this step. Copulas could also achieve this end, but are not necessary as long as the farms exhibit the imposed correlation structure.

The county-level yield for each iteration is simply the weighted sum of the farms, as is actual NASS county-level yield data used in applied research and policy design. The average farm and the county standard deviations are calculated from these synthetic data. Last of all, ρ_c is calculated from the correlation matrix of the 500 correlated yield draws.

While most of the parameters for the characterization of the hypothetical county are straightforward, the correlations are more complex. The correlation matrix must have unitary diagonals and be positive semidefinite. We rely on a method by Higham (2002) to find the correlation matrix nearest our specification. If the matrix of weights used in the algorithm is the identity matrix, the process is simply to compute the spectral decomposition of the correlation matrix, set the negative eigenvalues to 0, recompose the correlation matrix with the new eigenvalues, and set the diagonals to 1. This process is repeated until the correlation matrix converges.

Simulations with Full County Data

In this first set of simulations, we calculate crop insurance premiums for a corn yield protection plan based on all 500 draws of the farm data and the county in order to evaluate the convergence of the populations. The actuarially fair premium for farm i is calculated as

$$Prem(i) = \frac{1}{m}\sum_{d=1}^{m} max(0, \alpha E[y_i] - y_{i:d}) \times 5.68, \tag{5}$$

where α is the coverage level, $y_{i:d}$ is the dth draw of farm i, and \$5.68 is the yield protection price offered for corn in 2012 for much of the Corn Belt. The average premium for a county is derived by taking the weighted sum of the farms, $Prem(i)^* = \sum_{i=1}^{n} w_i Prem(i)$. The actuarially fair premium derived from inflating the county data is

$$Prem(c)^* = \frac{1}{m}\sum_{d=1}^{m} max\left(0, \alpha E[y_c] - \left(\frac{y_{c:d} - E[y_c]}{\sqrt{\rho_c}} + E[y_c]\right)\right) \times 5.68, \tag{6}$$

where ρ_c is calculated from correlated draws and $y_{c:d}$ is the weighted average of the farm yields from draw d. This specification inflates the county deviates consistent with proposition 2 so that the county-level yield standard deviation approximately equals the average farm-level yield standard deviation. Last of all, we include the noninflated premium from the county data, which is calculated as

$$Prem(c) = \frac{1}{m}\sum_{d=1}^{m} max(0, \alpha E[y_c] - y_{c:d}) \times 5.68. \tag{7}$$

Table 1 represents the simulation outcomes of nine scenarios. The first scenario assumes twenty identical farms with equal production shares and normally distributed farm yields with mean 175 and standard deviation 35 in every case. The distinguishing factor is cross-farm correlation: yields are randomly correlated between 0.5 and 0.9. The small number of farms with unrelated standard deviations and correlations corresponds to lemma 3. Consistent with the lemma, the ratio of the county variance to the farm variance is at least the average interfarm correlation. However, even with this small number of farms, $\sigma_c^2/(\sigma_f)^2$ and ρ_c are very close. The average farm premium for the first scenario is \$9.92 compared to \$10.01 estimated from the county data, implying that using county-level averages would result in an overestimate of the premium by just under 1%. Conversely, the unadjusted county premium is \$5.01, approximately half of the average farm premium. In this case, adjusting the county data provides a good estimate while not doing so would result in considerable error.

The next two scenarios build directly from the first. The second scenario in table 1 is identical to the first except for the increased number of farms within the county. This scenario corresponds more closely to proposition 2. In this scenario, the ratio of the county variance to the farm variance converges more closely to interfarm correlation relative to the previous scenario, supporting proposition 2. Once again, the premiums calculated precisely from the farm data and the premium estimated from the adjusted county data are only about 1% different, while the unadjusted county data premium is only about 50% of the average farm premium. The third scenario is identical to the second except the mean and standard deviation of each farm is drawn from uniform distributions. This scenario also corresponds to proposition 2 and adds greater farm-level heterogeneity, but results in the same values for the ratio of county variance to farm variance. However, the adjusted and actual premiums are 8% apart in this case, while the unadjusted county premium is less than half the actual premium.

Scenarios 4 through 6 are the analogues of the first three with farm yields drawn from beta distributions. The marginal farm beta distributions are bounded between 0 and 120% of the

Table 1. Simulation Parameters and Results, $\alpha = 0.75$

Farm Distribution	w_i	n	μ_i	σ_i	$\rho_{i,j}$	$\sigma_c^2/(\sigma_f)^2$	ρ_c	$Prem(i)^*$	$Prem(c)^*$	$Prem(c)$
1 Normal	$1/n$	20	175	35	Uniform [0.5,0.9]	0.725	0.710	9.92	10.01	5.01
2 Normal	$1/n$	500	175	35	Uniform [0.5,0.9]	0.701	0.700	9.54	9.66	4.63
3 Normal	$1/n$	500	Uniform [150,200]	Uniform [25,45]	Uniform [0.5,0.9]	0.701	0.700	10.31	9.49	4.54
4 Beta	$1/n$	20	175	35	Uniform [0.5,0.9]	0.693	0.676	19.87	19.02	10.88
5 Beta	$1/n$	500	175	35	Uniform [0.5,0.9]	0.668	0.667	19.87	18.54	10.36
6 Beta	$1/n$	500	Uniform [150,200]	Uniform [25,45]	Uniform [0.5,0.9]	0.662	0.663	20.66	18.52	10.26
7 Normal	McLean, IL	1,513	Uniform [150,200]	Uniform [25,45]	Uniform [0.5,0.9]	0.699	0.699	10.47	9.62	4.58
8 Normal	$1/n$	100	Uniform [150,200]	Uniform [25,45]	If neighbors, 0.9; If 1 separation, 0.7; Else, 0.5	0.531	0.526	9.83	8.97	2.02
9 Normal	$1/n$	100	If $i > 90$, uniform [160,180]; Else uniform [120,140]	If $i > 90$, uniform [40,50]; Else, uniform [25,35]	If $i\&j > 90$, 0.9; Else if $i\&j \leq 90$, 0.7; Else, 0.5	0.655	0.663	6.78	5.94	6.29

mean yield. The result is a left-tail skewness that some evidence suggests to be present in yield distributions. The average interfarm correlation for this set of scenarios is lower than in the first three as the simulations rely on Pearson's correlation coefficient, which can be sensitive to outliers. Even so, the ratio of county variance to farm standard deviation squared approximates the average interfarm correlation in the third set of scenarios, similar to the case of the first three scenarios that assumed normal distributions. As the propositions are independent of distribution, this result is consistent with preceding mathematical proofs. However, the premiums estimated from adjusted county data underestimate average farm premiums by 4–10%. Paralleling the first three scenarios, premiums from the unadjusted county data are about half of the actual average farm premiums.

Scenario 7 analyzes the case of heterogeneous farm size. We do this by using 2007 Census of Agriculture data for McLean County, Illinois. There are 1,513 farms within the county classified in twelve acreage categories that range from 1 to 9 acres up to 2,000 plus acres. All other assumptions are taken from scenario 3. Despite varying farm size, the results in this case are quite similar to the scenario with homogeneous farms. The ratio of variances converges to the average interfarm correlation. The premiums from the adjusted county data underestimate the average premium from farm data by 8%, as in scenario 3.

Scenario 8 is also a permutation of the third scenario, with the correlations altered. It is quite likely in reality that the correlation between yields for two farms is related to their proximity within the county. Therefore, we construct a hypothetical county with 100 farms arranged in a 10×10 grid. If two farms are immediately adjacent, their correlation is 0.9. If there is one farm separating them, their correlation is 0.7, otherwise it is 0.5. Each farm is homogenous with respect to size and shape. The results of the scenario indicate that conditional spatial correlation does not alter the results. The ratio of the variances is within 0.005 of the average interfarm correlation, and the difference between the average farm premium and the adjusted county premium is about 9%, or slightly greater than scenario 3. The unadjusted county premium is less than 25% of the actual average farm premium, which is the largest relative difference of the scenarios.

The last scenario in table 1, scenario 9, recasts scenario 3 to represent a county with uplands and bottomlands. Once again, a county with 100 farms is used. Ten farms are assumed to be the bottomlands, where yields have higher means but also a higher variances relative to yields of the

ninety upland farms. The bottomland farm yields have correlation coefficients of 0.9 with each other while the upland yields have correlation coefficients of 0.7 with each other. The correlation between yields in uplands and bottomlands is 0.5. This scenario corresponds to lemma 2 with negative covariances between the interfarm correlations and standard deviations. Therefore, the average interfarm correlation is larger than the ratio of the county variance to the farm variance. Even so, the two are still quite close. The absolute difference in premiums is less than 3%.

Simulations with Limited County Data

In each of the nine scenarios in table 1, the ratio of the county to farm variance and the average interfarm correlation are quite close, even under relaxed assumptions, indicating that the county variance and the average interfarm correlation can be used to provide a reasonable approximation of an unknown farm variance. However, the differences between the premium from the adjusted county data and the average of the farms do not necessarily converge, as shown in the table. Furthermore, the adjusted county premiums use all 500 observations from the farms. In reality, no time series provides that many observations of crop yield history. Limited time series particularly complicate premiums for low coverage levels as the calculations are based on only a few data points. In order to measure the error from these issues, we repeat the simulations but use a smaller sample size for the county and differing coverage levels.

This set of simulations begins by taking the 500 farm and county draws from the previous simulations and randomly selecting thirty county draws with replacement to represent a thirty-year time series of trend adjusted yields as might be used by an analyst. These county yield deviates from the observed mean are inflated by $1/\sqrt{\rho_c}$ to obtain a proxy dataset of the farm yields. From these adjusted yields, we use the nonparametric Kernel Density Estimator (KDE) to generate random yields. The KDE is formally defined as

$$f(x) = \frac{1}{nh}\sum_{i=1}^{n} K\left(\frac{x - x_i}{h}\right), \tag{8}$$

where $K(x) \in [0,1]$, $\int_{-\infty}^{\infty} K(x)dx = 1$, and h is a smoothing parameter. This nonparametric estimator allows us to fit the data without consideration to the proper parameterized distribution of the aggregation of the farm data. We use Silverman's rule of thumb (1986) to determine h:

$$A = min(\text{standard deviation, interquartile range}/1.34); \tag{9}$$

$$h = 0.9An^{1/5}. \tag{10}$$

Thirty deviates are generated from a Gaussian kernel around each of the thirty random adjusted county yields and the deviates are inflated to account for the variance bias in the second order kernel to obtain 900 nonparametrically distributed draws with the same mean and standard deviation as the small proxy dataset. The actuarially fair premium is calculated from equation (5). This process is repeated 500 times to generate a distribution of premiums.

Tables 2 and 3 contain the results of these simulations. The parameters of table 2 correspond to scenarios 2 and 5 in table 1, except the coverage level is allowed to vary. The results are expressed in terms of percentage deviation from the average of farm premiums. At the 50% coverage level, the simulated premium from the normal farm yield distributions was more than 10% below the actual farm premium. However, increasing the coverage level to 75% reduces the bias by almost half. The error nearly disappears at the 90% coverage level. Likewise, the difference between the two premiums was greatest at the 50% level for the beta distributions. This relative difference also decreases as the coverage level increased, but is always negative and greater in relative terms than when farm yields are normally distributed.

The relative errors in table 2 also decrease with an increase in coverage level. At the 50% level, the relative standard error is almost twice the actual farm premium. At this low coverage level, the

Table 2. Percent Deviation of Average County Based Premiums from Average Farm Premium, with Homoskedasticity and Constant Mean of Yields and Thirty Random County Observations

	Distribution	
Coverage Level	**Normal**	**Beta**
50%	11.2%	−22.1%
	[1.877]	[1.013]
75%	5.3%	−16.0%
	[0.561]	[0.451]
90%	0.7%	−7.4%
	[0.236]	[0.267]

Notes: Relative standard errors are in brackets. The average farm-based premium is used as the mean in the RSE calculation.

Table 3. Percent Deviation of Average County Based Premiums from Average Farm Premium, with Random Means and Standard Deviation of Yields and Thirty Random County Observations

	Distribution	
Coverage Level	**Normal**	**Beta**
50%	−52.8%	−40.8%
	[0.892]	[0.731]
75%	−7.1%	−16.3%
	[0.504]	[0.433]
90%	−1.3%	−5.1%
	[0.239]	[0.262]

Notes: Relative standard errors are in brackets. The average farm based premium is used as the mean in the RSE calculation. The random means are uniformly distributed between 150 and 200 and the standard deviations are randomly distributed between 25 and 45.

premiums are usually cheap and are determined by a small number of observations in the sample of thirty, making them subject to wide swings between iterations. At the 90% coverage level, the relative standard error decreases to approximately 24% of the actual farm premium.

Table 3 is similar to table 2 but allows farm means to be random between 150 and 200 and standard deviations to vary between 25 and 45. The accuracy of the second set of simulations is generally worse than the first set. However, the relative standard error generally decreases from the first set of simulations for the low coverage levels. Although these simulations are less accurate overall, they are more precise for low coverage levels. The limited sample, adjusted county premiums with the normal distribution are less than the "actual" farm distributions for all coverage levels with the additional randomness of parameters.

Several conclusions can be drawn from the second set of simulations. First, farm premium estimates based on county-level yields do not result in unbiased estimates. Second, the bias is inversely related to the coverage level. At high coverage levels, the issue largely disappears. Third, the standard error of the estimator decreases with the increase in coverage levels. Not only does the accuracy increase at higher coverage levels, but the precision also increases. Last of all, the bias and standard error of basing premiums on limited county-level yield as a proxy for unknown farm data appears to be sensitive to the distribution of farm yields. In general, we found that a normal distribution of farm yields gives rise to more accurate estimates than the beta distribution but with a lower degree of precision.

Conclusions

To the best of our knowledge, the fact that county-level yields are the aggregate of farm-level yields has not been completely exploited as a source of information. We use mathematical derivations and

simulation experiments to draw conclusions from this relationship between county- and farm-level yields.

Our approach provides applied researchers with a new tool to analyze policy design and implementation within realistic time and resource constraints. Our findings are relevant for applied researchers who assess the farm impacts, crop supply effects, and costs of crop insurance and other subsidy programs, both existing and proposed. County-level yield data are already used in crop insurance policy implementation in the form of transition yields, so better methods of linking these data to the underlying farm data can improve accuracy of premiums, potentially benefitting farmers or taxpayers.

Our approach has broader relevance as well. Other experiments that link county-level data to farm-level data—including yield-distribution assessments—can take advantage of our findings. Under certain assumptions that might be reasonable in some key producing regions, additional information about the relationship between county-level yield data and component farm-level yield data can be exploited for any attempt to use aggregated data in studies of farm-level yields. Second, we work with unconditional moments in this paper, but the method may be applied to state-contingent correlations and other conditional moments of the yield distributions such as lower partial moments (for a discussion of partial moments, see Antle, 2010). Finally, we use kernel density estimators in our application, but the method could be extended to other probability modeling tools linked to the underlying moments of the random variables. For example, some researchers have used copula-based methods to derive more flexible probability models, and the Cornish-Fisher expansion has a long history in the empirical literature on insurance and actuarial methods (see Bowers, 1967, for an early contribution).

The key conclusions from the mathematical proofs are important boundaries on county-level yields as they relate to farm-level yields. As an aggregate of its parts, the county average correlation cannot be too negative (such that shares of individual farms define exactly where that lower limit lies) but can theoretically be 1. Not only do farm-level yield variances exceed county-level variances in the few cases where previous researchers have found data to be available and assessed it, our calculations prove that average farm-level yield standard deviation must be at least to equal county-level standard deviation and is very likely to exceed it in any actual case. Moreover, we derive mathematical relationships between county-level and farm-level yields under plausible conditions that lend themselves to applied work. In a county with many farms and certain conditions regarding how the correlation among farm-level yield relates to farm variances, the county-level variance converges to the product of average farm standard deviation and average interfarm correlation. This relationship lends itself to using county-level yield data to estimate farm-level yield data with a minimum of additional information or assumptions.

Simulation results test the propositions and estimate the error in premiums if the mathematically derived formulas are applied using data representing a variety of hypothetical counties. Scenarios are differentiated by numbers of farms, yield variance and mean, distribution type, and more elaborate settings that account for the sorts of challenges that applied researchers expect when using real data. While the key propositions work well under relaxed assumptions, one key conclusion is that using county data may result in biased estimates of crop insurance premiums. However, another conclusion is that the bias tends to decrease as coverage rises, to the point that bias almost disappears at high coverage levels. The size of the error is also quite small in that case. The results are sensitive to assumed distribution type, suggesting an avenue for further research in this direction and potentially raising questions if county-level yield data are used in tests intended to shed light on farm-level yield distribution type.

We cannot draw broad conclusions about whether these methods should or should not be used for policy design and implementation. The advantage we see in an approach based on this method is that it can be applied quickly for all county-commodity combinations for which NASS provides data. This fact alone represents an advantage over methods that rely on data that are not widely available and calculations that take more time and are consequently narrow in relevance and potentially too

late to be useful for policy making. We also use simulations to demonstrate some risks of bias and error that can be quite small under certain conditions but substantial in other cases. Evaluating benefits and costs for this approach depends on the specific experiment.

[Received July 2013; final revision received April 2014].]

References

Antle, J. M. "Asymmetry, Partial Moments and Production Risk." *American Journal of Agricultural Economics* 92(2010):1294–1309.

Atwood, J., S. Shaik, and M. Watts. "Are Crop Yields Normally Distributed? A Reexamination." *American Journal of Agricultural Economics* 85(2003):888–901.

Barnett, B. J., J. R. Black, Y. Hu, and J. R. Skees. "Is Area Yield Insurance Competitive with Farm Yield Insurance?" *Journal of Agricultural and Resource Economics* 30(2005):285–301.

Bowers, N. L. "An Approximation to the Distribution of Annuity Costs." *Transactions of Society of Actuaries* 19(1967):295–309.

Brown, W. "Some Experimental Results in the Correlation of Mental Abilities1." *British Journal of Psychology* 3(1910):296–322.

Bulut, H., K. J. Collins, and T. P. Zacharias. "Optimal Coverage Level Choice with Individual and Area Insurance Plans." *American Journal of Agricultural Economics* 94(2012):1013–1023.

Carriquiry, M. A., B. A. Babcock, and C. E. Hart. "Using a Farmer's Beta for Improved Estimation of Expected Yields." *Journal of Agricultural and Resource Economics* 33(2008):52–68.

Claassen, R., and R. E. Just. "Heterogeneity and Distributional Form of Farm-Level Yields." *American Journal of Agricultural Economics* 93(2011):144–160.

Coble, K. H., and B. J. Barnett. "Implications of Integrated Commodity Programs and Crop Insurance." *Journal of Agricultural and Applied Economics* 40(2008):431–442.

Coble, K. H., and R. Dismukes. "Distributional and Risk Reduction Effects of Commodity Revenue Program Design." *Applied Economic Perspectives and Policy* 30(2008):543–553.

Coble, K. H., R. G. Heifner, and M. Zuniga. "Implications of Crop Yield and Revenue Insurance for Producer Hedging." *Journal of Agricultural and Resource Economics* 25(2000):432–452.

Cooper, J., C. Zulauf, M. Langemeier, and G. Schnitkey. "Implications of within County Yield Heterogeneity for Modeling Crop Insurance Premiums." *Agricultural Finance Review* 72(2012):134–155.

Cooper, J. C. "Average Crop Revenue Election: A Revenue-Based Alternative to Price-Based Commodity Payment Programs." *American Journal of Agricultural Economics* 92(2010):1214–1228.

Cooper, J. C., M. R. Langemeier, G. D. Schnitkey, and C. R. Zulauf. "Constructing Farm Level Yield Densities from Aggregated Data: Analysis and Comparison of Approaches." Paper presented at the annual meeting of the Agricultural and Applied Economics Association, Milwaukee, WI, July 26–28, 2009.

De Veaux, D. "Tight Upper and Lower Bounds for Correlation of Bivariate Distributions Arising in Air Pollution Modeling." SIMS 05, Stanford University, SIAM Institute for Mathematics and Society, Stanford, CA, 1976.

Deng, X., B. J. Barnett, and D. V. Vedenov. "Is There a Viable Market for Area-Based Crop Insurance?" *American Journal of Agricultural Economics* 89(2007):508–519.

Goodwin, B. K. "Problems with Market Insurance in Agriculture." *American Journal of Agricultural Economics* 83(2001):643–649.

———. "Payment Limitations and Acreage Decisions under Risk Aversion: A Simulation Approach." *American Journal of Agricultural Economics* 91(2009):19–41.

Goodwin, B. K., and A. P. Ker. "Nonparametric Estimation of Crop Yield Distributions: Implications for Rating Group-Risk Crop Insurance Contracts." *American Journal of Agricultural Economics* 80(1998):139–153.

Harri, A., K. H. Coble, A. P. Ker, and B. J. Goodwin. "Relaxing Heteroscedasticity Assumptions in Area-Yield Crop Insurance Rating." *American Journal of Agricultural Economics* 93(2011):703–713.

Hennessy, D. A. "Crop Yield Skewness and the Normal Distribution." *Journal of Agricultural and Resource Economics* 34(2009):34–52.

Higham, N. J. "Computing the Nearest Correlation Matrix–A Problem from Finance." *IMA Journal of Numerical Analysis* 22(2002):329–343.

Iman, R. L., and W. J. Conover. "A Distribution-Free Approach to Inducing Rank Correlation among Input Variables." *Communications in Statistics* B11(1982):311–334.

Just, R. E., and Q. Weninger. "Are Crop Yields Normally Distributed?" *American Journal of Agricultural Economics* 81(1999):287–304.

Ker, A. P., and K. Coble. "Modeling Conditional Yield Densities." *American Journal of Agricultural Economics* 85(2003):291–304.

Ker, A. P., and B. K. Goodwin. "Nonparametric Estimation of Crop Insurance Rates Revisited." *American Journal of Agricultural Economics* 82(2000):463–478.

Koundouri, P., and N. Kourogenis. "On the Distribution of Crop Yields: Does the Central Limit Theorem Apply?" *American Journal of Agricultural Economics* 93(2011):1341–1357.

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

Office of the Chief Economist. "World Agricultural Supply and Demand Estimates." WASDE, U.S. Department of Agriculture, Washington, DC, 2012. Available online at http://www.usda.gov/oce/commodity/wasde/.

Ramirez, O. A., S. Misra, and J. Field. "Crop-Yield Distributions Revisited." *American Journal of Agricultural Economics* 85(2003):108–120.

Schnitkey, G. D., B. J. Sherrick, and S. H. Irwin. "Evaluation of Risk Reductions Associated with Multi-Peril Crop Insurance Products." *Agricultural Finance Review* 63(2003):1–21.

Sherrick, B. J., F. C. Zanini, G. D. Schnitkey, and S. H. Irwin. "Crop Insurance Valuation under Alternative Yield Distributions." *American Journal of Agricultural Economics* 86(2004):406–419.

Silverman, B. W. *Density Estimation for Statistics and Data Analysis*. Monographs on Statistics and Applied Probability. London: Chapman and Hall, 1986.

Spearman, C. "Correlation Calculated from Faulty Data." *British Journal of Psychology* 3(1910):271–295.

Tack, J., A. Harri, and K. Coble. "More than Mean Effects: Modeling the Effect of Climate on the Higher Order Moments of Crop Yields." *American Journal of Agricultural Economics* 94(2012):1037–1054.

USDA Risk Management Agency. "A Risk Management Agency Fact Sheet of Agriculture: About the Risk Management Agency." Program Aid 1667-02, U.S. Department of Agriculture, Risk Management Agency, Washington, DC, 2010.

———. "Revised Premium Ratings for Corn and Soybeans." 2011. Available online at www.rma.usda.gov/help/faq/rerating.html.

White, T. K., and R. A. Hoppe. "Changing Farm Structure and the Distribution of Farm Payments and Federal Crop Insurance." Economic Information Bulletin 91, U.S. Department of Agriculture, Economic Research Service, Washington, DC, 2012. Available online at http://www.ers.usda.gov/publications/eib-economic-information-bulletin/eib91.aspx.

Zulauf, C. R., M. R. Dicks, and J. D. Vitale. "ACRE (Average Crop Revenue Election) Farm Program: Provisions, Policy Background, and Farm Decision Analysis." *Choices: The Magazine of Food, Farm & Resource Issues* 23(2008):29–35.

Appendix A: Proof of Equation (4)

Consider a county, c, with n farms $\{x_1, x_2, \ldots, x_n\}$, each with a w_i fraction of acres in the crop for the county. By definition,

$$\sigma_c^2 = Var\left(\sum_{i=1}^{n} w_i x_i\right) = \sum_{i=1}^{n}\sum_{j=1}^{n} \rho_{i,j} w_i w_j \sigma_i \sigma_j, \tag{A1}$$

where $Var(c) = \sigma_c^2$, $Var(x_i) = \sigma_i^2$, $Var(x_j) = \sigma_i^2$, $\rho_{i,j} = Corr(x_i, x_j)$, and $\sum_{i=1}^{n} w_i = 1$. Since the actual parameters are unobserved, let $\rho_{i,j}$, σ_i and σ_j be treated as random variables with $E[\rho_{i,j}] = \rho_c$ and $E[\sigma_i] = E[\sigma_j] = \sigma_f$. Since ρ_c is the expected interfarm correlation, it is calculated as

$$\rho_c = \frac{\sum_{i=1}^{n}\sum_{j\neq i}^{n} \rho_{i,j} w_i w_j}{\sum_{i=1}^{n}\sum_{j\neq i}^{n} w_i w_j} = \frac{\sum_{i=1}^{n}\sum_{j\neq i}^{n} \rho_{i,j} w_i w_j}{1 - \sum_{i=1}^{n} w_i^2}, \tag{A2}$$

while the average farm standard deviation, σ_f, is calculated as

$$\sigma_f = \sum_{i=1}^{n} w_i \sigma_i. \tag{A3}$$

Furthermore, let $\overline{\rho}_j$ be the weighted average of row (column) j in the correlation matrix, calculated as

$$\overline{\rho}_j = \sum_{i=1}^{n} w_i \rho_{ij}. \tag{A4}$$

Equation (A4) can be summed across the j's and $\sum_{j=1}^{n} w_j^2$ subtracted from both sides:

$$\sum_{j=1}^{n} w_j \overline{\rho}_j - \sum_{j=1}^{n} w_j^2 = \sum_{j=1}^{n}\sum_{i=1}^{n} w_j w_i \rho_{ij} - \sum_{j=1}^{n} w_j^2; \tag{A5a}$$

$$\frac{\sum_{j=1}^{n} w_j \overline{\rho}_j - \sum_{j=1}^{n} w_j^2}{1 - \sum_{j=1}^{n} w_j^2} = \frac{\sum_{j=1}^{n}\sum_{i=1}^{n} w_j w_i \rho_{ij} - \sum_{j=1}^{n} w_j^2}{1 - \sum_{j=1}^{n} w_j^2}. \tag{A5b}$$

Note that $\sum_{j=1}^{n}\sum_{i=1}^{n} w_j w_i \rho_{ij} - \sum_{j=1}^{n} w_j^2 = \sum_{i=1}^{n}\sum_{j\neq i}^{n} \rho_{i,j} w_i w_j$, which reduces the right hand side to

$$\frac{\sum_{j=1}^{n} w_j \overline{\rho}_j - \sum_{j=1}^{n} w_j^2}{1 - \sum_{j=1}^{n} w_j^2} = \rho_c. \tag{A5c}$$

This can be rearranged to isolate $\sum_{j=1}^{n} w_j \overline{\rho}_j$:

$$\sum_{j=1}^{n} w_j \overline{\rho}_j - \sum_{j=1}^{n} w_j^2 = \rho_c\left(1 - \sum_{j=1}^{n} w_j^2\right); \tag{A5d}$$

$$\sum_{j=1}^{n} w_j \overline{\rho}_j = \rho_c\left(1 - \sum_{j=1}^{n} w_j^2\right) + \sum_{j=1}^{n} w_j^2. \tag{A5e}$$

Moreover, consider the covariance between the correlations for farm j and the corresponding farm standard deviations where $\boldsymbol{\sigma} = \{\sigma_1, \sigma_2, \ldots, \sigma_n\}$:

$$Cov(\rho_{ij|j}, \sigma) = \sum_{i=1}^{n} w_i (\rho_{ij} - \overline{\rho}_j)(\sigma_i - \sigma_f); \tag{A6a}$$

$$= \sum_{i=1}^{n} w_i \rho_{ij} \sigma_i - \sigma_f \sum_{i=1}^{n} w_i \rho_{ij} - \overline{\rho} \sum_{i=1}^{n} w_i \sigma_i + \sigma_f \overline{\rho}_j \sum_{i=1}^{n} w_i. \tag{A6b}$$

Since $\sigma_f \sum_{i=1}^{n} w_i \rho_{ij} = \overline{\rho}_j \sum_{i=1}^{n} w_i \sigma_i = \sigma_f \overline{\rho}_j \sum_{i=1}^{n} w_i = \sigma_f \overline{\rho}_j$,

$$= \sum_{i=1}^{n} w_i \rho_{ij} \sigma_i - \sigma_f \sum_{i=1}^{n} w_i \rho_{ij}. \tag{A6c}$$

Since $Cov(\rho_{ij|j}, \sigma) = \sum_{i=1}^{n} w_i \rho_{ij} \sigma_i - \sigma_f \overline{\rho}_j$, the terms can be rearranged to conclude

$$\sum_{i=1}^{n} w_i \rho_{ij} \sigma_i = Cov(\rho_{ij|j}, \sigma) + \sigma_f \overline{\rho}_j. \tag{A6d}$$

The results of equation (A6c) can be used to derive an identity involving the covariance for the average farm correlations and the farm standard deviations:

$$Cov(\overline{\rho}_j, \sigma) = Cov(\sum_{i=1}^{n} w_i \rho_{ij}, \sigma); \tag{A7a}$$

$$= \sum_{j=1}^{n} w_j Cov(\rho_{ij|j}, \sigma); \tag{A7b}$$

$$= \sum_{j=1}^{n} w_j \sum_{i=1}^{n} w_i \rho_{ij} \sigma_i - \sigma_f \sum_{j=1}^{n} w_j \sum_{i=1}^{n} w_i \rho_{ij}; \tag{A7c}$$

$$= \sum_{i=1}^{n} w_i \sigma_i \sum_{j=1}^{n} w_j \rho_{ij} - \sigma_f \sum_{j=1}^{n} w_j \sum_{i=1}^{n} w_i \rho_{ij}; \tag{A7d}$$

$$= \sum_{i=1}^{n} w_i \sigma_i \overline{\rho}_j - \sigma_f \sum_{j=1}^{n} w_j \overline{\rho}_j; \tag{A7e}$$

$$= \sum_{i=1}^{n} w_i \sigma_i \overline{\rho}_j - \sigma_f (\rho_c (1 - \sum_{j=1}^{n} w_j^2) + \sum_{j=1}^{n} w_j^2); \tag{A7f}$$

$$= \sum_{i=1}^{n} w_i \sigma_i \overline{\rho}_j - \sigma_f \rho_c + \sigma_f \rho_c \sum_{j=1}^{n} w_j^2 - \sigma_f \sum_{j=1}^{n} w_j^2. \tag{A7g}$$

Since $Cov(\overline{\rho}_j, \sigma) = \sum_{i=1}^{n} w_i \sigma_i \overline{\rho}_j - \sigma_f \rho_c + \sigma_f \rho_c \sum_{j=1}^{n} w_j^2 - \sigma_f \sum_{j=1}^{n} w_j^2$, the terms can be rearranged to conclude:

$$\sum_{i=1}^{n} w_i \sigma_i \overline{\rho}_j = \sigma_f \rho_c - \sigma_f \rho_c \sum_{j=1}^{n} w_j^2 + \sigma_f \sum_{j=1}^{n} w_j^2 + \sum_{j=1}^{n} w_j Cov(\rho_{ij|j}, \sigma). \tag{A7h}$$

The results of equation (A6d) can be substituted into equation (1):

$$\sigma_c^2 = \sum_{j=1}^{n} w_j \sigma_j (\sum_{i=1}^{n} w_i \rho_{ij} \sigma_i); \tag{A8a}$$

$$= \sum_{j=1}^{n} w_j \sigma_j (Cov(\rho_{ij|j}, \sigma) + \sigma_f \overline{\rho}_j); \tag{A8b}$$

$$= \sum_{j=1}^{n} w_j \sigma_j Cov(\rho_{ij|j}, \sigma) + \sigma_f \sum_{j=1}^{n} w_j \sigma_j \overline{\rho}_j. \tag{A8c}$$

A similar substitution of equation (A7h) can be made into equation (A8c):

$$\sigma_c^2 = \sum_{j=1}^{n} w_j \sigma_j Cov(\rho_{ij|j}, \sigma) + \sigma_f(\sigma_f \rho_c - \sigma_f \rho_c \sum_{j=1}^{n} w_j^2 + \sigma_f \sum_{j=1}^{n} w_j^2 + \sum_{j=1}^{n} w_j Cov(\rho_{ij|j}, \sigma)) \tag{A9a}$$

$$\sigma_c^2 = \sum_{j=1}^{n} w_j \sigma_j Cov(\rho_{ij|j}, \sigma) + (\sigma_f)^2 \rho_c - (\sigma_f)^2 \rho_c \sum_{j=1}^{n} w_j^2 + (\sigma_f)^2 \sum_{j=1}^{n} w_j^2 + \sigma_f \sum_{j=1}^{n} w_j Cov(\rho_{ij|j}, \sigma). \tag{A9b}$$

Appendix B: Proof of Proposition 1

Every correlation matrix is a positive semidefinite matrix. By definition, for a positive semidefinite matrix ($\boldsymbol{M}$)

$$\boldsymbol{v'Mv} \geq 0, \tag{B1}$$

where $\boldsymbol{v}$ is a column vector. The following is a direct result of the preceding property:

$$\boldsymbol{w'Rw} = \sum_{i=1}^{n} \sum_{j \neq i}^{n} \rho_{i,j} w_i w_j + \sum_{i=1}^{n} w_i^2 \geq 0, \tag{B2}$$

where $\boldsymbol{w}$ is the column vector of yield weights and $\boldsymbol{R}$ is the correlation matrix of farm yields. Rearranging this equation,

$$\frac{\sum_{i=1}^{n} \sum_{j \neq i}^{n} \rho_{i,j} w_i w_j}{1 - \sum_{i=1}^{n} w_i^2} \geq -\frac{\sum_{i=1}^{n} w_i^2}{1 - \sum_{i=1}^{2} w_i^2}, \tag{B3}$$

which provides the lower bound because the left side is equal to the average interfarm correlation. If all the correlations equal 1, then the left side of the inequality is maximized. The average interfarm correlation is 1 in this case.

Appendix C: Proof of Lemma 1

First, note that $\frac{\partial \sigma_c^2}{\partial \rho_{i,j}} = w_i w_j \sigma_i \sigma_j \geq 0$ because $w_i \geq 0$, $w_j \geq 0$, $\sigma_i \geq 0$, and $\sigma_j \geq 0$. Therefore, the county variance defined in equation (1) is maximized when the cross-correlations, $\rho_{i,j}$, are maximized. The maximum value of correlation between any farms i and j is 1. If true for all farms, then equation (1) becomes

$$\sigma_c^2 = \sum_{i=1}^{n} \sum_{j=1}^{n} w_i w_j \sigma_i \sigma_j = (\sigma_f)^2. \tag{C1}$$

Conversely, given equation (1) and its derivative, the county variance with respect to the interfarm correlations is at its minimum when the interfarm correlations are at their minimum, -1. Due to the positive semidefinite characteristic of any correlation matrix, not all farms can be negatively correlated if there are more than two farms.

In the case in which there are only two farms with perfect negative correlation in the county, then

$$\sigma_c^2 = w_1^2 \sigma_1^2 + w_2^2 \sigma_2^2 - 2 w_1 w_2 \sigma_1 \sigma_2 = (w_1 \sigma_1 - w_2 \sigma_2)^2. \tag{C2}$$

In this two-farm case, the county variance must be at least 0 due to the squared term—as well as by the definition of variance. However, the county variance takes a value of 0 if $w_1 \sigma_1 = w_2 \sigma_2$, thus establishing that the theoretical lower bound of county variance is 0.

Appendix D: Proof of Lemma 2

It is worth noting that as $w_i \to 0$, $n \to \infty$. In other words, the lemma is for an increasing number of farms with weights uniformly going to 0. While w_i can equal $\frac{1}{n}$, it does not have to as long as no farms dominate production. If w_i did equal $\frac{1}{n}$, $w_i \to 0$ could simply be thought of as $\frac{1}{n}$ as $n \to \infty$. In this case, $\sum_{i=1}^{n} \frac{1}{n}$ would still have to sum to 1 while the individual weights of the farms would still go to 0.

Starting with equation (4), as $w_i \to 0$, $\sum_{i=1}^{n} w_i^2 \to 0$. Although σ_f and ρ_c are functions of w_i, they remain finite as $w_i \to 0$, just as $\sum_{i=1}^{n} w_i = 1$ as $w_i \to 0 \ \forall \ i$. Therefore, $-(\sigma_f)^2 \rho_c \sum_{j=1}^{n} w_j^2 + (\sigma_f)^2 \sum_{j=1}^{n} w_j^2 \to 0$ as $w_i \to 0$. $\sum_{j=1}^{n} w_j \sigma_j Cov(\rho_{ij|j}, \boldsymbol{\sigma}) + \sigma_f \sum_{j=1}^{n} w_j Cov(\rho_{ij|j}, \boldsymbol{\sigma})$ does not disappear as the number of farms must be increasing which offsets the effect of the decreasing weights. However, given that the distribution is discrete, these terms can be rewritten as expectations.

The deviate is said to have a maximum value of $(\sigma_f)^2$ in the text. Consider a correlation matrix $\boldsymbol{M}$ with $m_{i,j} = (-1)^{i+j}$ and a farm standard deviation of $\sigma_i = 2\sigma_f$ if i is even, $\sigma_i = 0$ otherwise. $\boldsymbol{M}$ is positive semidefinite, has all diagonals equal to 1 and contains values between or equal to -1 and 1 thereby satisfying the criteria of a correlation matrix. In this case, $\rho_c = 0$ while the left term of the deviate will reduce to $(\sigma_f)^2$ and the right term will reduce to 0. Even though it is unlikely to find a county with such a correlation matrix, this example does show that even if $\sigma_f^2 \rho_c = 0$, the county variance can equal the farm variance with asymptotics.

Appendix E: Proof of Lemma 3

This lemma follows from substituting the covariance into equation (4) and taking advantage of the boundary on the squared sum of weights, namely $\sum_{j=1}^{n} w_j^2 \in (0, 1]$. In this case, equation (4) reduces to

$$\sigma_c^2 = (\sigma_f)^2 \rho_c - (\sigma_f)^2 \rho_c \sum_{j=1}^{n} w_j^2 + (\sigma_f)^2 \sum_{j=1}^{n} w_j^2. \tag{E1}$$

Lemma 3 is consistent with lemma 1. County variance cannot be negative from this proposition, even though the average interfarm correlation can be negative. To see this, start with proposition 1 and take the minimum value possible for ρ_c, specifically $-\frac{\sum_{i=1}^{n} w_i^2}{1-\sum_{i=1}^{n} w_i^2}$. Substituting the lower bound into equation (E1) yields

$$\sigma_c^2 = (\sigma_f)^2 \left(-\frac{\sum_{i=1}^{n} w_i^2}{1 - \sum_{i=1}^{n} w_i^2} \right) - (\sigma_f)^2 \left(-\frac{\sum_{i=1}^{n} w_i^2}{1 - \sum_{i=1}^{n} w_i^2} \right) \sum_{j=1}^{n} w_j^2 + (\sigma_f)^2 \sum_{j=1}^{n} w_j^2 = 0. \tag{E2}$$

Appendix F: Proof of Proposition 2

If $Cov(\rho_{ij|j}, \sigma) = 0$ and the individual farm shares approach 0, then equation (4) reduces to

$$\sigma_c^2 = (\sigma_f)^2 \rho_c - (\sigma_f)^2 \rho_c \sum_{j=1}^{n} w_j^2 + (\sigma_f)^2 \sum_{j=1}^{n} w_j^2 \to (\sigma_f)^2 \rho_c, \text{ as } w_i \to 0 \ \forall \ i \tag{F1}$$