

Data Aggregation Issues for Crop Yield Risk Analysis

Margot Rudstrom¹, Michael Popp², Patrick Manning³
and Edward Gbur⁴

¹ Assistant professor, West Central Research and Outreach Center,
University of Minnesota, Morris, Minnesota.

² Associate professor, Department of Agricultural Economics and Agribusiness,
University of Arkansas, Fayetteville, Arkansas.

³ Research specialist, Department of Agricultural Economics and Agribusiness,
University of Arkansas, Fayetteville, Arkansas.

⁴ Professor, Agricultural Statistics Laboratory at the
University of Arkansas, Fayetteville, Arkansas.

With increased emphasis on risk management in agriculture and a lack of disaggregated or farm-level yield time series, decision makers are often faced with having to make adjustments to temporal yield risk measures obtained from readily available but aggregated yield data. This paper provides some empirical evidence on what type of aggregation bias to expect when measuring temporal yield risk using yield observations averaged across a region relative to yield risk estimated from quarter-section yield time series in wheat. This study highlights some of the challenges faced when estimating aggregation distortions in measuring yield risk defined by temporal variance, especially given the nature of the empirical data set used. Cluster analysis, visual examination of relative frequency distributions and mapping of yield risk clusters suggest that using a readily available, aggregate temporal yield risk measure has the tendency to underestimate yield risk observed at the quarter-section level and that clear, geographic yield risk boundaries do not exist in municipalities or across larger areas in this study. Further research on crops more risky than wheat appears promising.

Avec un plus grand intérêt sur la gestion du risk dans l'agriculture et un manque de données détaillées ou bien de collections de séries temporelles sur les rendements, les décideurs sont souvent tenus d'apporter des correctifs aux mesures du risk obtenues à partir des données de rendements qui sont disponibles. Cet article apporte une preuve empirique du type de biais lié à l'agrégation qui peut être présent dans le calcul du risk de rendement temporel obtenu à partir de rendements moyens de blé observés au niveau régional en comparaison du risk de rendement qui est estimé à partir de données basées sur des quart-de-sections.

Cette étude met en exergue quelques uns des obstacles qui se présentent dans l'estimation de distorsions liées à l'agrégation dans le calcul du risk de rendement défini par la variance temporelle, spécialement étant donné le caractère empirique des données utilisées. L'analyse de groupe, l'examen visuel de la distribution des fréquences relatives, et la cartographie de classes de risk de rendement suggèrent que l'utilisation de la mesure du risk de rendement basée sur des données disponibles de risk agrège temporel a tendance à sous-estimer le risk de rendement observé au niveau des quart-de-sections et qu'il n'y a pas de frontières de risk de rendement certaines, géographiques qui existent entre les municipalités ou bien à travers les zones plus larges examinées dans cette étude.

INTRODUCTION

With increasing emphasis on risk management in agriculture due to changes in agricultural policy and the globalization of markets (Harwood et al 1999), analysts attempting to capture production and price risk appropriate for decision makers at the farm level are often confronted with a lack of data suitable for yield risk analysis.¹ While county-level yield data are readily available, it is argued that such aggregated data do not accurately reflect variability conditions at the farm level and thus use of aggregated data may lead to erroneous research conclusions (Bechtel and Young 1999; Debrah and Hall 1989; Wang and Zhang 2002). Therefore, some research has presented results using sensitivity analysis on the second moment of data for such results to reflect differences between aggregate and farm-level data (Fulton, King and Fackler 1988; Popp, Dalsted and Skold 1997; Skees and Nutt in Mapp and Jeter 1988). It is this scarcity of data and the lack of a spatial data aggregation adjustment process that have prompted this research.

The objectives of this paper are:

- to provide further empirical evidence of the type of distortion that can be expected between farm-level wheat yield data and data that have been aggregated to some degree
- to suggest a method of identifying an appropriate aggregation level to use for risk analysis (i.e., if aggregate data are too distorting, how much further reporting detail is necessary to capture farm-level yield variability)
- to report on aggregation issues that are encountered when using cluster analysis for grouping yield data according to similar temporal variance.

The paper proceeds with a background to this research by summarizing some of the literature on this topic. The data source for the empirical analysis is then discussed in the context of the methods to be used. A statement of research hypotheses and a summary of findings follow.

LITERATURE REVIEW ON YIELD DATA AGGREGATION ISSUES

In efforts to address the difference between nonaggregated farm and aggregated regional yield data, three issues are typically highly relevant. First, the lack of consistent farm-level data from unbiased sources presents the most difficult barrier for analysis. Skees and Reed (1986) argue that farm-level yield data from crop insurance agencies may be biased as an adverse selection problem may exist with crop insurance participation. Yield data reporting efforts by governmental agencies are also restricted by transaction costs of reporting farm-level data and data privacy issues as well as survey respondent considerations (repeated surveying of the same individuals is discouraged to prevent respondent frustration). This typically leads to incomplete panel data so that a time series is available but not for the same field or farm for more than two or three years. Further, the size of the field has implications for yield variation (Marra and Schurle 1994; Eisgruber and Schuhman 1957). Second, detrending of yield data, to take account of technological change, for example, influences estimates of yield variation. In this context, there is continuing discussion in the literature about consistent guidelines on how to appropriately detrend yield data (Marra and Schurle 1994; Young 1980). Third, measurement and subsequent analysis of yield variability necessitates the use of an appropriate yield probability density function. The crop insurance analysis literature uses a variety of distribution assumptions ranging from normal (Botts and Boles 1960) to beta (Nelson 1990) to triangular (Mapp and Jeter 1988). Recent

research by Just and Weninger (1999) suggests the use of normal distributions for yield variability. They suggest further that detrending can introduce skewness and nonnormal kurtosis. Often the central limit theorem is used to support the use of the normality assumption (Wang and Zhang 2002).

BACKGROUND

Manitoba Crop Insurance Corporation (MCIC) is the provincial agency that manages crop insurance in Manitoba (Beattie 1994). It is a Crown corporation responsible for the development, administration and sale of revenue insurance, all risk crop insurance, additional hail insurance, livestock feed security insurance and honey insurance. Its mission is to provide yield protection to insured Manitoba farmers (and indirectly the agricultural supply industry) in years when crop loss occurs due to uncontrollable natural hazards and to deliver other income protection programs to Manitoba farmers as requested by the Minister of Agriculture. During the 1989–90 production year, more than 15,000 insurance contracts were sold representing approximately 80% of eligible producers.² In excess of \$138 million was paid out as insurance indemnities.

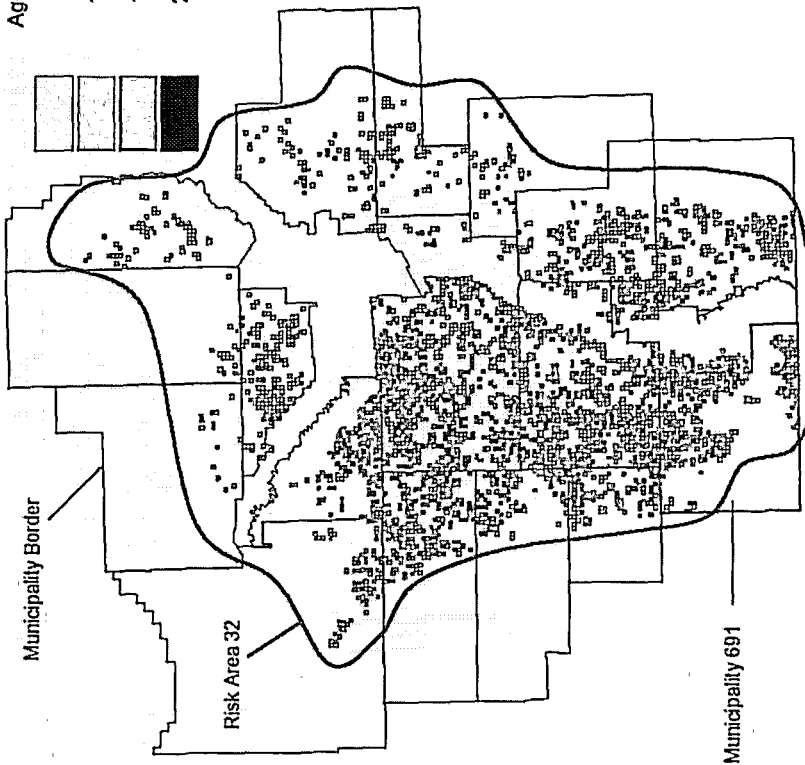
Manitoba is divided into a number of risk areas to reflect different crop growing conditions. Risk areas therefore represent relatively homogenous, contiguous areas in terms of crop yields with the underlying assumption that the variability of yields on the farm are not statistically different than the variability in the risk area. There are 16 large risk areas. Risk area 12 is the Red River Valley, with the heavier Osborne clay soils as the predominant soil type. Risk area 32 is a subarea of risk area 12 and comprises a number of municipalities (see risk area 32 and municipality borders in the map in Figure 1).

Within each municipality, land is divided into quarter-sections, which are identified with a legal descriptor. Yield information is available on a field level, with the field being located within a quarter-section. A quarter-section can have more than one field growing a particular crop in a particular year when the quarter-section is subdivided. This occurs relatively infrequently in the data set used.

For purposes of this paper, yield data aggregation occurs from quarter-section to municipality to risk area. To assess the appropriateness of aggregating yield data for measuring yield variability from quarter-section to municipality, the hypothesis that temporal variances of yield on each quarter-section within a municipality are equal is tested. If we fail to reject this hypothesis and higher-order moments are not of interest, then farm-level yield variances are expected to be similar across the municipality. Arriving at a representative yield variance estimate across an area with several quarter-sections still remains an issue, however. The following example illustrates this concern.

Municipality variance, calculated as the temporal variance of average annual yields observed across all quarter-sections in a municipality, can be different than the average of the quarter-section yield variance estimates even if individual quarter-section variance estimates are similar (Table 1). The three hypothetical scenarios illustrate situations where the average of quarter-section variances (the bold numbers shown in the bottom rows of each scenario titled Summary statistics) is either less than, equal to or greater than a more readily available measure of temporal yield risk using aggregated data (the bold numbers in the last column). Note that, in all situations, quarter-section variance estimates would likely be judged statistically similar and that quarter-section mean yields are relatively constant.

Cluster Yield Information		
Aggregate ^a	Range ^b	# of Q.S.
51.2	0.0 - 128.9	1,571
115.7	129.2 - 253.9	1,151
176.4	254.0 - 412.9	444
234.7	416.3 - 866.9	106



Relative Frequency of Quarter-Section Variance - Risk Area

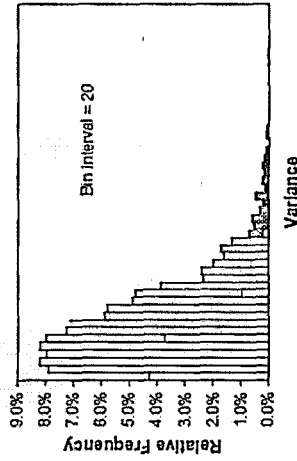


Figure 1. Geographically mapped cluster information for risk area 32

^aAggregate yield risk is the variance of annual average yields observed per cluster and is the same information presented in Table 4.

^bRange of individual quarter-section variance estimates observed within a cluster.

Table 1. Hypothetical examples of aggregated variance estimates with four quarter-section yield time series

Year	Quarter sections (Q.S.) in bu/acre ^a				Annual aggregate average yield
	1	2	3	4	
Scenario 1					
1	31	—	29	30	30.0
2	30	31	30	27	29.5
3	29	30	31	—	30.0
4	32	—	—	28	30.0
5	—	32	31	30	31.0
6	28	29	27	—	28.0
7	—	28	28	29	28.3
8	29	31	30	31	30.3
Temporal variance estimate	2.17	2.17	2.29	2.17	1.00
Mean yield	29.83	30.17	29.43	29.17	29.64
CV	4.93%	4.88%	5.14%	5.05%	3.38%
Summary statistics ^b			2.20		0.46
Scenario 2					
1	31	—	28	32	30.3
2	30	34	30	31	31.3
3	29	34	29	—	30.7
4	31	—	—	33	32.0
5	—	36	29	33	32.7
6	31	34	28	—	31.0
7	—	35	30	32	32.3
8	30	35	29	33	31.8
Temporal variance estimate	0.67	0.67	0.67	0.67	0.68
Mean yield	30.33	34.67	29.00	32.33	31.50
CV	2.69%	2.36%	2.82%	2.53%	2.61%
Summary statistics ^b			0.67		1.01
Scenario 3					
1	29	—	29	27	28.3
2	29	31	31	28	29.8
3	30	32	31	—	31.0
4	27	—	—	25	26.0
5	—	31	31	28	30.0
6	30	31	30	—	30.3
7	—	31	30	28	29.7
8	28	29	28	27	28.0
Temporal variance estimate	1.37	0.97	1.33	1.37	2.58
Mean yield	28.83	30.83	30.00	27.17	29.14
CV	4.05%	3.19%	3.85%	4.30%	5.52%
Summary statistics ^b			1.26		2.05

^aQuarter-section yield time series may not have observations each year due to crop rotations and other considerations.

^bSummary statistics are the average of temporal variance estimates across quarter-sections and the ratio of the temporal variance estimate of the annual aggregated average yields to the average of the temporal variance estimates across quarter-sections (see text discussion related to Table 1).

These scenarios illustrate that, even if quarter-section variance estimates are tested for equivalence and the hypothesis is not rejected (i.e. the average of the quarter section variance estimates is similar to each individual observation and therefore a good estimate of variance across the area), distortions from using the aggregate measure of municipality variance (the measure in the right-hand column) may still be introduced. These distortions are presented as the italicized ratio in the bottom right-hand corner of each scenario. Note that the ratio varies from 0.46 in the top scenario to 2.05 in the bottom scenario and thus use of the more readily available aggregate yield risk measure may either under- or overestimate yield risk for the area. Further, there does not appear to be a distinct relationship between the ratio and the initial data set given the unbalanced nature of the data.

Given the initial question of how to summarize a risk measure across an area, the above example illustrates that two measures may be chosen:

- using the average of individual quarter-section yield variance estimates as an estimate of quarter-section level variance across an area where individual quarter-section variances are statistically equal
- using the aggregate yield risk measure, that is readily available from reported average annual yield data across an area; further, the ratio of the two variance estimates captures the distortion from aggregation that results from using the more readily available measure.

Two questions arise from this discussion:

- What distortion ratio values are empirically observed? If the ratio is empirically close to one, the more readily available aggregate yield risk measure may be used without concern about introducing aggregation bias
- Is it worthwhile to arrange data into subsets of similar variance? That is, by how much do the empirical observations of the ratio change if the data are arranged into subgroups with similar variance?

Empirical evidence related to this last question might suggest:

- how researchers need to adjust aggregate variance estimates to reflect farm conditions across different levels of aggregation (i.e., municipality to farm vs. risk area to farm)
- how representative is research using aggregate yield risk measurements (i.e., a summary of when using aggregate yield risk over- or underestimates farm-level yield risk).

Finally, another question is whether clear geographical boundaries or patterns for similar yield risk observations emerge if yield data are grouped into subsets with similar risk characteristics? Should there be clear patterns, data reporting agencies (e.g., MCIC, Agriculture Canada, USDA, etc.) might be able to adjust their reporting of yield data to paint a clearer picture of areas with similar yield risk.

METHODOLOGY

To address the above questions, a Bartlett's test is conducted to test for equivalence of temporal yield variance across quarter-sections within each municipality. The Bartlett's test is used as a screening tool to determine whether data need to be grouped into subsets with similar yield risk characteristics. This test is performed at both the municipality and risk area level of aggregation.

The Bartlett's test statistic is $X_0^2 = 2.3026 \frac{q}{c}$ where:

$$q = (N - a) \log_{10} S_p^2 - \sum_{i=1}^a (n_i - 1) \log_{10} S_i^2$$

$$c = 1 + \frac{1}{3(a-1)} \left(\sum_{i=1}^a (n_i - 1)^{-1} - (N - a)^{-1} \right)$$

$$S_p^1 = \frac{\sum_{i=1}^a (n_i - 1) S_i^2}{N - a}$$

a = number of quarter-sections in the municipality or risk area

n_i = number of observations for a quarter-section i , $i = 1, \dots, a$

N = total number of observations in the municipality

S_i^2 = estimate of temporal yield variance at the quarter-section level.

The Bartlett's statistic is distributed χ^2 with $a - 1$ degrees of freedom. The null hypothesis for each municipality is $H_0: \sigma_1^2 = \sigma_2^2 = \dots = \sigma_a^2$. The same test is conducted for the risk area.

The Bartlett's test is sensitive to the assumption of normality (Montgomery 1991). For large samples normality generally is not an issue, as the central limit theorem applies. For small samples the degree of skewness and kurtosis relative to the normal distribution will affect the robustness of the Bartlett's test (Boos and Brownie 1989).

For those municipalities where temporal quarter-section variance estimates are not equal, quarter-sections are divided into subgroups of similar yield variation using cluster analysis. Cluster analysis allows objects to be placed into groups suggested by characteristics of the data. Objects in a cluster should be similar to each other in some sense — in this case, quarter-sections with similar temporal variances would be placed in the same group or cluster.

Using nonhierarchical clustering, the number of clusters can either be specified in advance or determined as part of the clustering procedure. The k -means method partitions the quarter-sections into k clusters (Johnson and Wichern 1998). An initial set of k quarter-sections are selected as seeds. Using these seeds, quarter-sections are assigned to a cluster whose centroid or mean of the characteristic in question is nearest. If a quarter-section is moved from one cluster to another, the centroids of the clusters receiving and losing the quarter-section are recalculated. The process is repeated until no more reassignments take place.

The results of the k -means procedure are sometimes sensitive to the seeds and thus it is important to evaluate the effects of different sets of starting seeds on the clustering results. The clustering procedure was therefore run using different seeds to identify whether results would be robust. Finally, determining the number of clusters (k) present in the data, the peak in the pseudo- F statistics was identified for values of $k = 1, \dots, 10$ for each municipality (Milligan and Cooper 1985).

K -means clustering was carried out using PROC FASTCLUS in SAS Version 8 (SAS Institute Inc., Cary, NC) with least squares as the distance criterion and a maximum of 15 iterations. Further, the analysis was performed on all municipalities and the risk area. It is not expected that the clustering is affected by spatial correlation of yields, as yield series that are spatially correlated would be placed in the same clusters if their temporal variation were sim-

ilar. However, given that nearby quarter-sections may not have data from exactly the same set of years, it isn't clear whether or not variances will be spatially correlated.

Once clusters have been identified, aggregation distortions in yield risk would be calculated similar to the statistics reported in Table 1. Further, the use of a geographical information system (GIS) allows plotting of yield risk clusters both at the municipality and risk area levels. Color-coding quarter-sections would then reveal patterns or geographic regions that have similar characteristics. Should such patterns emerge, this information might ultimately be used to potentially reduce aggregation distortions in reporting yield risk.

DATA MANAGEMENT

Field-level crop yield data were obtained from the Manitoba Crop Insurance Corporation (MCIC) for risk area 32 from 1980 to 1990. The crop analyzed is hard red spring wheat, as it is the predominant crop grown in risk area 32. Relative to other crops, the data are relatively complete and consistent.

The data are available by legal description to the detail of quarter-sections. For risk area 32, there were 21,987 individual field level yield observations (in bushels per acre) over the 11-year period once the following rules were applied to make the data set, both representative and manageable:

- To be included in the data set a quarter-section had to have wheat harvested at least in four of the 11 years to reflect the percentage of total crop acreage in wheat production over the period. With this rule, land that is typically not used for wheat production would not influence results.
- Field size per quarter-section allowed the gathering of several annual yield observations per quarter-section in 2,153 cases.

When a quarter-section had multiple fields for the same year, multiple annual observations were replaced with their simple average.

With these restrictions, the data set was used to obtain 3,272 quarter-section level temporal yield variance estimates for risk area 32. Yield mean and temporal variance were calculated for each quarter-section over the 11-year period. Neither the means nor variances were weighted for planted acreage.

Municipality mean yields are the simple average of all observations in the municipality. Aggregate temporal yield risk measures were calculated using the methods described in Table 1. The ratio of the two aggregate variance estimates, presented in the last column of Table 2, is the measure of distortion in variance estimates using aggregate data as discussed previously.

Shapefiles — a data format used with ArcView Version 3.2 GIS software (Environmental Systems Research Institute Inc., Redlands, CA) — of the municipalities of southern Manitoba at a level of detail of quarter-sections were obtained from Linnet (2001). Using these municipality files, a shapefile for the entire risk area was constructed.

RESULTS

Equivalence of Variance Tests

Bartlett's tests for equality of quarter-section temporal yield variances for each of the nineteen municipalities and risk area 32 are presented in Table 2. The null hypothesis of equal variances across all quarter-sections within the municipality was rejected for nine municipal-

Table 2. Summary statistics of risk area 32 municipalities, 1980-90

Municipality ID	Size ^a	N ^b	Size of sample ^c	Avg. # of years / quarter-section	Mean yield ^d	Bartlett's statistic	P-value ^e	Ratio ^f
231	837	986	163	6.05	34.0	126.53	0.982	0.72
281	1106	949	177	5.36	32.6	181.77	0.367	0.65
282	1368	524	101	5.19	34.7	118.00	0.106	0.69
361	1475	1315	226	5.82	36.4	283.92	0.005	0.74
451	1440	1226	222	5.52	35.4	235.08	0.246	0.69
461	1152	192	38	5.36	36.4	36.14	0.509	0.65
510	1738	4166	717	5.19	34.7	712.60	0.529	0.66
552	910	804	130	5.82	37.0	178.91	0.002	0.82
561	1678	3494	583	5.99	34.2	667.05	0.008	0.74
671	3857	491	77	6.38	36.6	110.92	0.006	0.72
691	1440	1891	305	6.20	36.1	352.06	0.030	0.70
692	691	148	26	5.69	36.3	24.01	0.519	0.53
721	1800	112	17	6.59	36.3	23.73	0.096	0.52
722	720	633	106	5.97	33.5	111.50	0.314	0.64
730	648	1000	162	6.17	32.1	191.19	0.052	0.37
741	2097	312	59	5.29	32.7	61.96	0.337	0.40
850	1728	458	80	5.73	33.6	104.84	0.028	0.20
881	883	420	69	6.09	35.0	99.26	0.008	0.48
991	1843	68	14	4.86	35.0	11.62	0.559	0.65
Risk area 32	27411	19189	3272	5.86	34.7	3943.89	<0.0005	0.61

^aTotal number of legal descriptors as a proxy of the actual size of the municipality. Some of the legal descriptors are river lots and represent areas that may be smaller than 160 acres.

^bN is the total number of annual yield observations in a municipality. Each quarter-section at the risk area level has at least four yield observations over the 11-year period.

^cTotal number of individual quarter-sections for which a least four annual yield observations were available. For some municipalities not all quarter-sections belonged to risk area 32 and thus the ratio of the size of the sample and the size of the municipality would be a biased estimate of sampling density (also see Figure 1).

^dThe average of all sample observations in a municipality or the risk area over quarter-sections and years.

^eEquality of variance is rejected at the 5% and 10% level of significance for seven and nine municipalities, respectively.

^fThe ratio of the temporal variance estimate of the annual aggregated average yields to the average of the temporal variance estimates across quarter-sections (see text discussion related to Table 1 as well as Tables 3 and 4 for values of risk measures).

ities at the 10% level. The ratio of the two aggregate variance estimates is always less than one when looking at individual municipalities or the risk area. Neither the size of the sample nor the value of the ratio appears to be related to the rejection of the null hypothesis. This finding suggests that cluster analysis needs to be performed to get to a better estimate of distortion within some municipalities and for the risk area.

Cluster Analysis

Cluster analysis was performed to arrange data into subsets of similar individual quarter-section variance. Pseudo- F statistics were used to determine the number of clusters for each municipality and the risk area where Bartlett's test lead to rejection at the 10% level. Cluster membership did not change substantially with the initial seeds chosen and thus the results were deemed robust.

Cluster results were also plotted in relative frequency distribution (RFD) charts to visually verify break points in quarter-section variance observations across a municipality or risk area. For the risk area (Figure 1), the chart shows four clusters that are not easily identifiable by looking at the graph without the cluster information presented in gray scale — i.e., the variance observations show a relatively continuous distribution except for the outliers on the right. For municipality 691 (Figure 2), the clusters are more readily apparent with peaks and troughs in the RFD. From visual analysis a fourth cluster may be justifiable for the right hand side outliers (for $k = 2$, pseudo- $F = 525.75$; $k = 3$, pseudo- $F = 679.59$; and $k = 4$, pseudo- $F = 658.59$).

Number and Size of Clusters

The results of the cluster analyses for municipalities where Bartlett's test was rejected are presented in Table 3. Clustering by temporal yield variance resulted in two to five clusters within a municipality or risk area. The number of clusters does not appear to be affected by size of the sample (e.g., note that the number of clusters does not change for municipalities 361 and 510, even though the number of individual quarter-section observations more than doubles). This finding may be particular to this data set or crop, however.

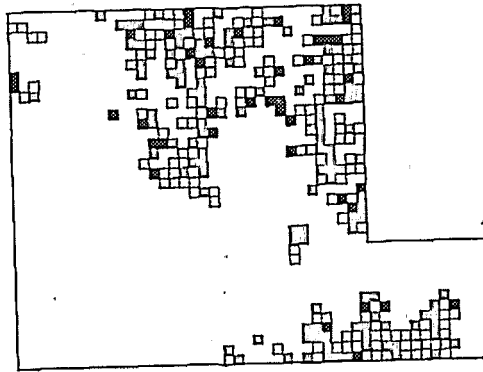
The number of individual quarter-section variance observations per cluster is also reported in Table 3. Note that the clusters are reported in order of increasing variance and that the number of individual quarter-section variance observations per cluster tends to be larger in the low-risk clusters than the higher-risk clusters across all municipalities and the risk area with the exception of municipality 881. High quarter-section variance observations are thus less frequent than low quarter-section variance observations for this crop.

Changes in Distortion




The ratio measuring aggregation distortion does not appear to show a consistent trend when changing from low-risk clusters to high-risk clusters. For five clusters the ratio declines with increasing risk, while no clear trend is apparent in municipalities 671, 721, 730 and 881. This observation is not surprising, as no relationship between the ratio and data characteristics is expected (recall discussion related to Table 1).

A casual observation from these findings is that for more risky observations (higher cluster numbers), aggregation tends to lead to larger underestimation of risk when using an aggregate measure relative to the less readily available average of individual quarter-sections. Perhaps more rigor toward determining aggregation bias would thus be justifiable for other crops exhibiting more yield variability.

To answer the question of what kind of differences a researcher could expect when using an aggregate yield risk measure relative to a measure obtained from data grouped into observations with similar risk, one can look to the ratios reported in Table 3. The minimum ratio of 0.27 is found in cluster 3 of municipality 850 and the maximum ratio of 1.51 is reported for cluster 1 in municipality 881. Using readily available data to measure yield risk thus typ-



Cluster Yield Information

	Aggregate ^a	Range ^b	# of Q.S.
	72.3	5.1 – 130.6	138
	117.5	132.6 – 270.4	126
	211.4	274.3 – 666.0	41

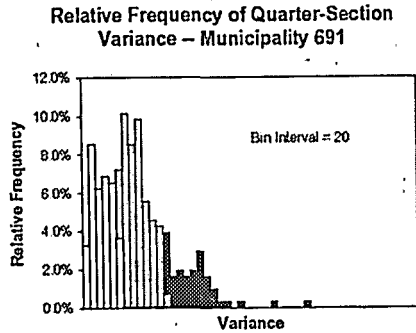


Figure 2. Geographically mapped clusters for municipality 691

^aAggregate yield risk is the variance of annual average yields observed per cluster and is the same information presented in Table 3.

^bRange of individual quarter-section variance estimates observed within a cluster.

ically under estimates risk at the quarter-section level (i.e., the ratio is less than 1 in 39 of the 46 ratios (85%) reported across municipalities and clusters in Table 3). Had clustering not been performed the range in distortions would have been from 0.20 in municipality 850 and 0.82 in municipality 552 (Table 2). The findings regarding the range in distortion are consistent with Debrah and Hall (1989), Eisgruber and Schuhman (1957), Carter and Dean (1960) and Freund (1960).

Bartlett’s Test

Given the limitations of the Bartlett’s test, *k*-means clustering was also carried out on all municipalities and the risk area (Table 4). Results were very similar to those reported above,

Table 3. Cluster analysis results for municipalities and risk area where Bartlett's test was rejected

Municipality ID	Item	Not clustered municipality	Clustered ^a				
			1	2	3	4	5
361	Aggregate yield risk (1)	102.8	66.9	18.7	134.3	223.7	
	Avg. of q.s. yield risk (2)	139.1	60.3	156.5	285.3	486.9	
	Ratio [(1)/(2)]	0.74	1.11	0.76	0.47	0.46	
	# of individual q.s. obs.	226	112	75	30	9	
552	Aggregate yield risk (1)	134.4	99.7	151.4	293.2		
	Avg. of q.s. yield risk (2)	163.1	77.4	196.7	390.4		
	Ratio [(1)/(2)]	0.82	1.29	0.77	0.75		
	# of individual q.s. obs.	130	61	54	15		
561	Aggregate yield risk (1)	113.9	78.5	113.2	182.5	274.3	
	Avg. of Q.s. yield risk (2)	154.5	62.2	168.2	280.6	488.3	
	Ratio [(1)/(2)]	0.74	1.26	0.67	0.65	0.56	
	# of individual q.s. obs.	583	241	219	105	18	
671	Aggregate yield risk (1)	97.2	47.3	85.1	97	162.3	271.4
	Avg. of q.s. yield risk (2)	134.1	45.1	102.2	160.7	237.1	371.5
	Ratio [(1)/(2)]	0.72	1.05	0.83	0.60	0.68	0.73
	# of individual q.s. obs.	77	23	20	16	15	3
691	Aggregate yield risk (1)	109.8	72.3	117.5	211.4		
	Avg. of q.s. yield risk (2)	157.3	72.8	185.6	354.8		
	Ratio [(1)/(2)]	0.7	0.99	0.63	0.60		
	# of individual q.s. obs.	305	138	126	41		
721	Aggregate yield risk (1)	38.7	24.3	133.0			
	Avg. of q.s. yield risk (2)	74.8	43.3	150.3			
	Ratio [(1)/(2)]	0.52	0.56	0.88			
	# of individual q.s. obs.	17	12	5			
730	Aggregate yield risk (1)	32.4	17.1	44.4	93.2	288.2	
	Avg. of q.s. yield risk (2)	86.8	43.2	98.5	187.4	358.9	
	Ratio [(1)/(2)]	0.37	0.40	0.45	0.50	0.80	
	# of individual q.s. obs.	162	82	60	15	5	
850	Aggregate yield risk (1)	18.4	16.7	37.8	63.9		
	Avg. of q.s. yield risk (2)	90.1	43.1	115.9	239		
	Ratio [(1)/(2)]	0.20	0.39	0.33	0.27		
	# of individual q.s. obs.	80	47	22	11		
881	Aggregate yield risk (1)	40.4	27.1	27.9	57.0	179.0	
	Avg. of q.s. yield risk (2)	83.7	18.0	62.6	108.7	200.1	
	Ratio [(1)/(2)]	0.48	1.51	0.45	0.52	0.89	
	# of individual q.s. obs.	69	19	16	26	8	
Risk area 32	Aggregate yield risk (1)	96.7	51.2	115.7	176.4	234.7	
	Avg. of q.s. yield risk (2)	157.6	69.3	185.5	315.5	502.5	
	Ratio [(1)/(2)]	0.61	0.74	0.62	0.56	0.47	
	# of individual q.s. obs.	3272	1571	1151	444	106	

^aData are reported for the municipality and then for the clusters in order of magnitude of variance — i.e., in municipality 361, the first cluster with 112 observations had the lowest individual quarter-section variances up to the last cluster with nine observations that had the highest individual quarter-section variances.

Table 4. Cluster analysis results for municipalities where Bartlett's test was not rejected

Municipality ID	Item	Clustered ^a										
		Not clustered					Clustered ^a					
		municipality	1	2	3	4	5	6	7	8	9	10
231	Aggregate yield risk (1)	164.5	142.0	237.0								
	Avg. of q.s. yield risk (2)	227.3	163.2	376.3								
	Ratio [(1)/(2)]	0.72	0.87	0.63								
	# of individual q.s. obs.	163	114	49								
281	Aggregate yield risk (1)	79.2	61.2	114.8								
	Avg. of q.s. yield risk (2)	122.7	75.5	219.6								
	Ratio [(1)/(2)]	0.65	0.81	0.52								
	# of individual q.s. obs.	177	119	58								
282	Aggregate yield risk (1)	149.5	99.6	121.2	218.9	406.8						
	Avg. of q.s. yield risk (2)	216.4	73.0	189.4	340.0	536.0						
	Ratio [(1)/(2)]	0.69	1.36	0.64	0.64	0.76						
	# of individual q.s. obs.	101	40	25	26	10						
451	Aggregate yield risk (1)	124.3	89.5	144.0	246.1							
	Avg. of q.s. yield risk (2)	181.1	87.1	239.7	410.8							
	Ratio [(1)/(2)]	0.69	1.03	0.60	0.60							
	# of individual q.s. obs.	222	110	90	22							
461	Aggregate yield risk (1)	83.7	41.2	121.5	296.0							
	Avg. of q.s. yield risk (2)	129.1	61.0	165.1	344.0							
	Ratio [(1)/(2)]	0.65	0.68	0.74	0.86							
	# of individual q.s. obs.	38	20	14	4							

Table 4. Continued

Municipality ID	Item	Clustered ^a										
		Not clustered municipality	1	2	3	4	5	6	7	8	9	10
510	Aggregate yield risk (1)	124.7	52.6	79.6	123.6	157.3	192.5	204.6	246.6	404.1		
	Avg. of q.s. yield risk (2)	188	42.1	109.9	176.6	239.3	316.9	399.2	500.8	615.9		
	Ratio [(1)/(2)]	0.66	1.25	0.72	0.70	0.66	0.61	0.51	0.49	0.66		
	# of individual q.s. obs.	717	122	174	153	132	73	36	19	8		
692	Aggregate yield risk (1)	61.6	47.8	61.2	62.7	96.0	121.5	239.6	482.7			
	Avg. of q.s. yield risk (2)	116.8	31.1	53.3	98.8	148.7	197.9	239.6	482.7			
	Ratio [(1)/(2)]	0.53	1.53	1.15	0.63	0.65	0.61	1.00	1.00			
	# of individual q.s. obs.	26	2	7	9	4	2	1	1			
722	Aggregate yield risk (1)	97.9	23.8	66.2	74.6	118.4	110.7	168.6	155.0	85.4	262.2	242.6
	Avg. of q.s. yield risk (2)	152.5	23.0	74.1	107.0	136.9	181.1	220.2	260.4	297.5	328.8	368.3
	Ratio [(1)/(2)]	0.64	1.03	0.89	0.70	0.86	0.61	0.77	0.60	0.29	0.80	0.66
	# of individual q.s. obs.	106	12	18	13	18	15	10	11	4	3	2
741	Aggregate yield risk (1)	31.5	17.7	36.6	42.7	87.2						
	Avg. of q.s. yield risk (2)	78.6	32.7	66.3	106.7	159.2						
	Ratio [(1)/(2)]	0.4	0.54	0.55	0.40	0.55						
	# of individual q.s. obs.	59	17	19	16	7						
991	Aggregate yield risk (1)	95.4	54.0	173.9								
	Avg. of q.s. yield risk (2)	147.3	78.4	239.2								
	Ratio [(1)/(2)]	0.65	0.69	0.73								
	# of individual q.s. obs.	14	8	6								

^aData are reported for the municipality and then for the clusters in order of magnitude of variance — i.e., in municipality 231, the first cluster with 114 observations had the lowest individual quarter-section variances up to the last cluster with 49 observations that had the highest individual quarter-section variances.

except that anywhere from two to ten clusters were found across the ten municipalities. This suggests that the Bartlett's test may be at best a weak screening tool on whether or not to perform cluster analysis. Further, clustering can lead to a relatively large set of subsets.

Geographic Variation

Mapping of clusters to identify risk patterns at the municipality or risk area level demonstrates that no clear geographical clustering exists. The maps in Figures 1 and 2 portray samples of maps generated using GIS. The implication of these findings from a data reporting agency's perspective is that smaller geographic risk areas based on similar yield variability for hard red spring wheat do not appear to be possible.

CONCLUSION

Risk areas are relatively large geographic areas and aggregating data from small units (quarter-sections) to large areas masks differences in yield variability across a risk area whether or not heterogeneous variance characteristics exist. When aggregating data, differences in variability across the units should be assessed in order to provide some guidelines to researchers on how representative their research recommendations are at the farm level when using aggregated data.

This research suggests that grouping data into subsets with similar variance using *k*-means cluster analysis, provides an opportunity to ascertain the level of distortion in yield risk measures when aggregating from individual quarter-sections to municipalities or larger areas. Distortions in using an aggregate measure relative to a less-biased measure, ranged from overestimating risk by just over 50% (a ratio of 1.53) and underestimating by reporting a yield risk measure of 27% of a more unbiased estimate (a ratio of 0.27). On average, underestimation of yield risk is more likely to occur at least in this empirical example.

Clustering provided a means to group data into like subsets and therefore allowed a more precise measurement of the type of aggregation distortion. There was no apparent direct relationship between the number of resulting clusters and the number of individual quarter section variance estimates analyzed. Further, mapping of clusters revealed no distinct geographical boundaries so that recommendations on improving data reporting for areas with similar yield risk cannot be made.

Recommendations for research from this paper are to conduct sensitivity analysis on variance to address difference in risk measured at the field relative to more aggregated yield information (i.e., municipality, risk area, etc.). Since this analysis focused on hard red spring wheat, which is not a very risky crop in terms of yield variability, it would be interesting to see if more risky crops exhibit similar patterns in terms of number of clusters, distribution of clusters and risk distortion.

NOTES

¹From here on, aggregation refers to moving from individual field observations (in this case, 160 acre quarter-section observations) or farm-level information to larger regions (i.e., counties, municipalities or risk areas).

²Because of the relatively high percentage of participation in the program adverse selection bias (Skees and Reed 1986) may not be much of a factor.

REFERENCES

- Beattie, J. C. 1994. Unlocking the corporate database: An agricultural application. In *Urban and Regional Information Systems Association Proceedings*, pp. 193-205.
- Bechtel, A. I. and D. L. Young. 1999. The importance of using farm-level risk estimates in CRP enrollment decisions. Selected Paper at the Western Agricultural Economics Association Annual Meetings.
- Boos, D. D. and C. Brownie. 1989. Bootstrap methods for testing homogeneity of variances. *Technometrics* 31: 69-82.
- Botts, R. R. and J. N. Boles. 1957. Use of normal-curve theory in crop insurance ratemaking. *Journal of Farm Economics* 39: 733-40.
- Carter, H. O. and G. W. Dean. 1960. Income, price and yield variability for principal California crops and cropping systems. *Hilgardia* 30 (6): 175-218.
- Debrah, S. and H. H. Hall. 1989. Data Aggregation and Farm Risk Analysis. *Agricultural Systems* 31: 239-45.
- Eisgruber, L. M. and L. S. Schuhman. 1963. The usefulness of aggregated data in the analysis of farm income variability and resource allocation. *Journal of Farm Economics* 45: 587-91.
- Freund, R. J. 1956. The introduction of risk into a programming model. *Econometrica* 24: 253-63.
- Fulton, J. R., R. P. King and P. L. Fackler. 1988. Combining farm and county data to construct farm-level yield distributions. Staff Paper P88-20. Minneapolis, MN: University of Minnesota, Department of Agricultural and Applied Economics.
- Harwood, J., R. Heifner, K. Coble, J. Perry and A. Somwaru. 1999. Managing risk in farming: concepts, research and analysis. Agricultural Economic Report No. 774. Washington, DC: USDA, ERS.
- Johnson, R. A. and D. W. Wichern. 1998. *Applied Multivariate Statistical Analysis*. Englewood Cliffs, NJ: Prentice-Hall Inc.
- Just, R. E. and Q. Weninger. 1999. Are crop yields normally distributed? *American Journal of Agricultural Economics* 81 (2): 287-304.
- Linnet. 2001. Quarter-section shapefiles for southern Manitoba. Winnipeg, MN.
- Mapp, H. P. and K. J. Jeter. 1988. Potential impact of participation in commodity programs and multiple peril crop insurance on a southwestern Oklahoma farm. Multiple peril crop insurance: A collection of empirical studies. Southern Cooperative Series Bulletin 334. Stillwater, OK: Oklahoma State University.
- Marra, M. C. and B. W. Schurle. 1994. Kansas wheat yield risk measures and aggregation: A meta-analysis approach. *Journal of Agricultural and Resource Economics* 19 (1): 69-77.
- Manitoba Crop Insurance Corporation. 1980-90. Field-level crop yield data, risk area 32. Portage la Prairie, Manitoba.
- Montgomery, D. C. 1991. *Design and Analysis of Experiments*. 3rd ed. New York, NY: Wiley.
- Milligan, G. W. and M. C. Cooper. 1985. An examination of procedures for determining the number of clusters in a data set. *Psychometrika* 50: 159-79.
- Nelson, C. H. 1990. The influence of distributional assumptions on the calculation of crop insurance premia. *North Central Journal of Agricultural Economics* 12: 71-78.
- Popp, M. P., N. L. Dalsted and M. D. Skold. 1997. Multiple peril crop insurance: An evaluation of crop insurance in northeastern Colorado. *Journal of the American Society of Farm Managers and Rural Appraisers* 19: 95-104.
- Skees, J. R. and M. R. Reed. 1986. Rate making for farm-level crop insurance: implications for adverse selection. *American Journal of Agricultural Economics* 68: 653-59.
- Wang, H. H. and H. Zhang. 2002. Model-based clustering for cross-sectional time series data. *Journal of Agricultural, Biological and Environmental Statistics* 7 (1): 107-27.
- Young, D. L. 1980. Evaluating procedures for computing objective risk from historical time series data. In *Risk Analysis in Agriculture: Research and Educational Development*. AE-4492. Urbana, IL: University of Illinois, Department of Agricultural Economics.

A vertical bar on the left side of the page, consisting of a series of yellow and orange rectangular segments. A small red diamond is located at the top of this bar.

COPYRIGHT INFORMATION

TITLE: Data Aggregation Issues for Crop Yield Risk Analysis
SOURCE: Can J Agric Econ 50 no2 JI 2002
WN: 0218204016006

The magazine publisher is the copyright holder of this article and it is reproduced with permission. Further reproduction of this article in violation of the copyright is prohibited.

Copyright 1982-2003 The H.W. Wilson Company. All rights reserved.