Risk programming and sparse data: how to get more reliable results

J. Brian Hardaker[1], Gudbrand Lien[2,3], Marcel A.P.M. van Asseldonk[4], James W. Richardson[5] and Agnar Hegrenes[3]

[1] *School of Business, Economics and Public Policy, University of New England, Armidale, NSW 2351, Australia*
[2] *Lillehammer University College, Lillehammer, Norway*
,[3] *Norway and Norwegian Agricultural Economics Research Institute (NILF), Oslo, Norway*
[4] *Institute for Risk Management in Agriculture, Wageningen University, Wageningen, The Netherlands*
[5] *Department of Agricultural Economics, Texas A&M University, College Station, Texas USA*

Abstract— **Because relevant historical data for farms are inevitably sparse, most risk programming studies rely on few observations. We discuss how to use available information to derive an appropriate multivariate distribution function that can be sampled for a more complete representation of the possible risks in risk-based models. For the particular example of a Norwegian mixed livestock and crop farm, the solution is shown to be unstable with few states, although the cost of picking a sub-optimal plan declines with increases in number of states by Latin Hypercube sampling.**

Keywords— **Risk programming, states of nature, sparse data**

I. INTRODUCTION

Risk programming may be performed to support a decision by an individual farmer about what farm plan to follow next year [e.g. 1, 2], or it may be undertaken to evaluate a proposed innovation such as a new technology or a new policy instrument such as crop insurance [e.g., 3, 4].

The form of risk programming models ranges from quadratic (E,V) risk programming to direct maximization of expected utility using nonlinear programming.

A common feature of most risk programming studies is that the representation of the risk and dependency among per unit activity net revenues is based on 10 or fewer observations. The reason is that, in practice, the required historical data for a large number of years are not available for the farm being analysed, or, even when the records exist, the relevance of the older information is judged to be low.

Quadratic (E,V) risk programming applications based on sparse data, have been used to look at the reliability of estimated optimal farm plans [e.g., 6] and the confidence regions in the mean-standard deviation space [e.g., 7]. All these studies demonstrate that sparse data reduce the reliability of the risk programming results.

In this paper we use utility efficient programming (UEP) [8] to illustrate an approach aimed at improving reliability of results based on limited information in farm risk programming. The starting point is a sparse data set from which a multivariate probability function is specified by means of a multivariate kernel density estimation procedure. Then efficient sampling from that distribution is used to reduce the need to include large numbers of discrete states to get reliable solutions in a risk programming model. We demonstrate the approach using an example of a typical Norwegian mixed farm.

II. THE UTILITY EFFICIENT PROGRAMMING MODEL USED

The UEP model for the case farm was formulated as follows:

$$\max E[U] = pU(z,r),\ r \text{ varied,} \tag{1}$$

subject to:

$$Ax \le b \tag{2}$$

$$Cx + APx - LFx - Iz = f \tag{3}$$

$$x \ge 0 \tag{4}$$

where: $E[U]$ is expected utility, p is a vector of probabilities for states of nature, $U(z,r)$ is a vector of utilities of net income where the utility function is defined for a measure of risk aversion, r, z is a vector of net incomes for each state of nature S, A is a matrix of technical coefficients, x is a vector of activity levels, b is a vector of resource stocks, C is a matrix of gross margins, GMs, (without public payment schemes) for S states of nature, AP is a matrix of public payment schemes for S states of nature, LF is a matrix of fodder costs for livestock activities for S states of nature, I is an identity matrix, f is a vector of fixed costs.

A. Utility and certainty equivalent

Because we assume that the farmer is risk-averse, we are restricted to using a concave form of the utility function with $U'(z) > 0$, and $U''(z) < 0$. We used the negative exponential function:

$$E[U(z,r)] = \sum_{s=1}^{S} p_s \{1 - \exp(-r \times z_s)\} \qquad (5)$$

where r is a non-negative parameter representing the coefficient of absolute risk aversion with respect to net income and we assume that all states of nature are equi-probable so that $p_s = 1/S$.

This function exhibits constant absolute risk aversion (CARA), which is a reasonable approximation to the real but unknown utility function for wealth for variations in transitory (annual) income [5]. For simplicity, we assume that the farmer's relative risk aversion with respect to wealth $r_r(w) = 2$, implying moderate risk aversion. However, we do not measure utility and risk aversion in terms of wealth, but in terms of transitory income (i.e., a bad or good result in one year has little effect on wealth and hence on income levels in subsequent years). Since we use a negative exponential utility function in terms of transitory income, z, we need a relationship between $r_r(w)$ and r. Assuming asset integration we have [8]:

$$r = r_r(w)/w \qquad (6)$$

The level of the farmer's wealth (net assets), w, is assumed to be NOK (Norwegian kroner) 2 million, so a value of $r = 2/2\,000\,000 = 0.000\,001$ was used as the farmer's degree of absolute risk aversion.

We converted the expected utility of net income of any farm plan to an estimate of the certainty equivalent, CE, by taking the reverse of the utility function, i.e.:

$$CE(z,r) = -\ln\{1 - E[U(z,r)]\}/r \qquad (7)$$

The estimates of CEs are readily interpreted because, unlike utility values, they are expressed in money terms.

B. Activities and constraints

The case farm was chosen to reflect the conditions of a typical lowland farm in Eastern Norway. The main activities in the UEP model of the case farm can be classified into (1) crop activities, (2) livestock activities: dairy cows and sheep, (3) concentrate feed activities, (4) hire labour and rent land activities, and (5) public payment schemes as of the year 2005 [9].

The main constraints were: (1) owned and rented land, (2) land use and rotational limits, (3) marketing limit, (4) milk quota, (5) labour, both seasonal constraints and constraints on hired labour, and (6) limits on subsidies.

C. States of nature data

To represent the uncertainty in activity GMs (i.e. matrix C in the UEP model), we needed some information on per unit GMs over a set of possible states of nature, ideally spanning states that the future might bring. The way we tackled the task of providing such states of nature information is described below.

III. MAKING THE BEST USE OF THE DATA

If the historical data to be used to represent risk in returns for a risk programming study are sparse, it would be desirable to bring more information into the process of specifying the states of nature matrix. Inevitably, there must be much subjectivity in this process and there will be scope for disagreement on how best to proceed. For example, historical data may need to be updated for changes in technology and in the value of money, requiring some 'detrending'. Yet too much detrending may eliminate some of the 'noise' that represents uncertainty about the future.

There are often irregularities in sparse data due to sampling errors, and it may be useful to smooth out the irregularities by fitting distributions [10, 11]. Smoothing might be combined with the introduction of information or judgments about the upper and lower limits of each of the uncertain quantities of interest. For example, there may be plausible upper and lower bounds for crop yields. Sometimes it will be appropriate to add some assumptions about the appropriate functional form to describe the marginal distributions based on the characteristics of the random processes generating the data rather than simply on goodness of fit considerations.

Similarly, in the face of sparse data, it makes sense to consider whether the data set could be expanded by using data from other sources. For example, data from neighbouring farms may perhaps be used to supplement information from the farm being planned. If yields of crops are highly dependent on seasonal weather conditions, and if the relationship between weather data and yields can be effectively modelled, it might be possible to use a longer series of historical weather data to generate more 'observations' of yields, although the reliability of such an approach obviously is compromised if climatic change is occurring.

The assumptions made should imply an improvement in the modelling of the future risks to be faced, not the opposite.

A. Preliminary processing

By way of illustration, in obtaining data for the present paper we mainly used the method described by Hardaker et al. [5: 80-82] applied to similar data and a similar farm planning problem as described by Lien and Hardaker [4]. The data covered the years 1996 to 2005, which is a relatively long sequence in the derivation of a state of nature matrix for risk programming, but is a small statistical sample.

Historical data from Eastern Norway in the Norwegian Farm Accountancy Survey were used to estimate the historical variation in activity GMs per unit within farms between years. The consumer price index was used to bring the individual activities to 2005-money values. From the farm-level, historical unbalanced panel data we derived a de-trended activity GM per unit matrix, representative for one single farm.

Assuming that historical data are not fully relevant for the future, the derived data for the individual farm were combined with subjective judgements of an expert about the marginal distributions of the individual activity GMs. Each trend-adjusted marginal distribution was revised to match the subjectively assessed means and standard deviations. Thus, the reconstructed series has the subjectively elicited means and standard deviations while preserving the general shapes of the marginal distributions and the correlation and other stochastic dependencies embodied in the historical data.

IV. DATA SMOOTHING AND DISTRIBUTION FITTING

Because our data covered only 10 years, we decided that some smoothing of the marginal distributions for each activity was appropriate. There are several ways in which smoothing might be done; (i) by hand smoothing a cumulative density function (CDF) for each marginal distribution; (ii) by some curve fitting method applied to the CDFs; or, as in this study, by the multivariate kernel density estimation (MVKDE) procedure proposed by Richardson et al. [12].

The procedure is a smoothed multivariate distribution extension of the multivariate empirical distribution estimate procedure described by Richardson et al. [13]. A kernel density estimation function is used to smooth the limited sample data of variables in a system individually, and then the dependencies present in the sample are used to model the system using a copula to join the marginal distributions into a multivariate one. Given a small sample, the choice of copula is more or less confined to the normal copula, based on correlations calculated from the adjusted sample data (in our case) or derived subjectively [e.g., 14].

A. Simulating additional states of nature

Once a smoothed multivariate distribution is defined, stochastic simulation may be used to generate as many states of nature as required for input into a modelling analysis such as risk programming. The

MVKDE method samples of any size to be drawn to provide a matrix of discrete states of GM outcomes. With a large enough sample, appropriately drawn, the smoothed distribution will effectively be recreated by the sampled drawings.

We next considered ways to draw samples that are more representative of the smoothed distribution from which they come than are samples drawn purely at random. Hence, in addition to random Monte Carlo sampling, we also used Latin hypercube sampling, which is a modified form of stratified sampling that generates a distribution of plausible collections of parameter values from a given multidimensional distribution [15]. For both sampling methods, correlated vectors of per unit GMs were sampled from the derived multivariate distribution using the MVKDE procedure described above.

B. Computations to evaluate simulated states of nature

In order to compare the efficiency of the sampling methods we simulated states of nature matrices of various sizes using the two different sampling methods. Our purpose was to examine the stability of the solutions and to assess how many states were required, using the two sampling methods, to approach the presumed true optimum with reasonable certainty.

To keep the computing task of the efficiency of alternative sampling sizes and methods within bounds, the programming model was solved with eight different numbers of states of nature, i.e. 5, 10, 15, 20, 30, 50, 100, and 200, each with only eight replicates, all repeated for Monte Carlo and Latin hypercube sampling.

For each risk programming solution obtained, we evaluated the CE of net income that could be expected ex ante from implementing that solution. Each evaluation was done using stochastic simulation with 500 replicates, drawn from the same MVKDE smoothed distribution sampled with the Latin hypercube algorithm. Thus, the simulated CEs of net income reported in the results for each of the $8 \times 8 \times 2 = 128$ solutions obtained are generally different from those obtained from the UEP programming results because they are based on a large sample of possible GM realisations drawn from the assumed known distribution of possible states.

V. RESULTS

The ranges in activity levels across the eight replicates with Latin hypercube sampling are summarised in the columns labelled LH in Table 1. When these results are compared with the ranges in activity levels in columns labelled MC (Monte Carlo), the advantages of the more efficient sampling method are clear. With Latin hypercube sampling the ranges are mostly narrower compared with Monte Carlo sampling. The activity levels also become relatively more stable with smaller numbers of states with Latin hypercube sampling than with Monte Carlo sampling.

Table 1 Ranges in levels of activities in solution farm plans between replicates with an increase in number of states of nature[a]

	No. states:	5		50		200	
	Sampling:	MC[b]	LH[b]	MC	LH	MC	LH
Barley	ha^{-3}	233.3	162.5	0.0	0.0	0.0	0.0
Oats	ha^{-3}	233.3	188.7	115.0	10.2	187.0	7.3
Wheat	ha^{-3}	152.9	204.3	195.5	0.0	180.0	0.0
Potatoes	ha^{-3}	15.1	0.0	0.0	0.0	0.0	0.0
Oilseed	ha^{-3}	196.0	0.0	0.0	0.0	0.0	0.0
Carrots	ha^{-3}	29.3	15.3	12.9	3.6	6.9	2.6
Grass seed	ha^{-3}	0.0	0.0	0.0	0.0	0.0	0.0
Dairy cows	no.	13.2	7.9	10.6	1.6	6.4	1.1
Sheep	no.	75.0	75.0	61.3	16.4	75.0	0.0

[a] The ranges are the max. level minus the min. for each activity for that number of states.

[b] MC = Monte Carlo sampling: LH = Latin hypercube sampling.

Moreover, a comparison of the results for CEs of net income in Fig. 1 shows that plans based on Latin hypercube sampling converge towards the presumed optimum value of CE with fewer states of nature than for those plans based on Monte Carlo sampling.

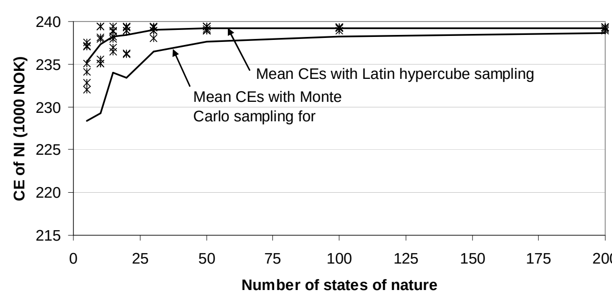

Fig. 1 Simulated results for CE of net income (NI), based on Latin hypercube sampling and the simulated mean CE based on Monte Carlo sampling for simulated numbers of states of nature between 5 and 200

The highest recorded CE of net income was NOK 239 402, which may be said to be the strict optimum if one accepts that the chosen smoothening procedure appropriately reflects the future risk. Fig. 1 shows that, at least for the model used here, it would be necessary to have a considerably larger number of states than are typically available from historical data to be reasonably sure that the solution is not appreciably suboptimal.

VI. DISCUSSION AND CONCLUSIONS

Although this analysis of the number of states of nature in risk programming models was based on one specific model with one particular data set, it is reasonable to argue that the findings will be broadly applicable. Our results imply that analysts undertaking risk programming studies need to be aware of bias from small samples. They need to give much more thought than seems to have been the case in the past to estimating multivariate probability functions that provide good descriptions of the risk to be faced in the planning period. It is likely that these descriptions will continue to be partly based on historical data, but there is a clear need to use other information and judgments to improve the relevance of the results. We have suggested some steps that might be taken in this direction and hope to see more discussion to improve the range of possible approaches.

Once an acceptable multivariate probability function of activity net revenues is obtained, efficient sampling from that distribution, for example by Latin hypercube sampling, as demonstrated in the paper, can reduce the need to include prodigiously large numbers of discrete states to get reliable solutions to risk programming models.

REFERENCES

1. Olson KD, Eidman VR (1992) A farmer's choice of weed control method and the impacts of policy and risk. Rev Agri Econ 14: 125-137.
2. Pannell DJ, Nordblom TJ (1998) Impacts of risk aversion on whole-farm management in Syria. Aust J Agri Res Econ 42: 227-247.
3. Dorward A (1999) A risk programming approach for analysing contractual choice in the presence of transaction costs. Eur Rev Agri Econ 26: 479-492.
4. Lien G, Hardaker JB (2001) Whole-farm planning under uncertainty: impacts of subsidy scheme and utility function on portfolio choice in Norwegian agriculture. Euro Rev Agri Econ 28: 17-36.
5. Hardaker JB, Huirne RBM, Anderson JR, Lien G, (2004) Coping with Risk in Agriculture, 2nd edn. Wallingford: CABI Publishing.
6. Lence SH, Hayes DJ (1995) Land allocation in the presence of estimation risk. J Agri Res Econ 20: 49-63.
7. Collender RN (1989) Estimation risk in farm planning under uncertainty. Amer J Agr Econ 71: 996-1002.
8. Patten LH, Hardaker JB, Pannell DJ (1988) Utility-efficient programming for whole-farm planning. Aust J Agri Econ 32: 88–97.
9. NILF (2005) Handbok for driftsplanlegging [Handbook in Farm Business Planning] 2005/2006. Oslo: Norwegian Agricultural Economics Research Institute.
10. Schlaifer R (1959) Probability and Statistics for Business Decisions. New York: McGraw-Hill.
11. Anderson JR, Dillon JL, Hardaker JB (1977) Agricultural Decision Analysis. Ames: Iowa State University Press.
12. Richardson JW, Lien G, Hardaker JB (2006) Simulating multivariate distributions with sparse data: a kernel density smoothing procedure, Poster paper contributed to the 26th International Conference of Agricultural Economics, Gold Cost, Australia, August 12-18, 2006.
13. Richardson JW, Klose SL, Gray AW (2000) An applied procedure for estimating and simulating multivariate empirical (MVE) probability distributions in farm-level risk assessment and policy analysis. J Agri Appl Econ 32: 299-315.
14. Fackler PL (1991) Modeling interdependence: an approach to simulation and elicitation. Amer J Agr Econ 73: 1091-1097.
15. McKay MD, Beckman RJ, Conover WJ (1979) A comparison of three methods for selecting values of input variables in the analysis of output from a computer code. Techometrics 21: 239-245.

The corresponding author:

- Author: Gudbrand Lien
- Institute: NILF
- Street: P.O. Box 8024 Dep.
- City: NO-0030 Oslo
- Country: Norway
- Email: gudbrand.lien@nilf.no