Creating Synthetic Agricultural Data Sets Using Copula Techniques :

I. Introduction and Background

How often does it happen that one sets out to develop computational techniques, intent on addressing the question of '*How to* ?', and subsequently realizes that the question. '*What* ?', - the basic concepts and definitions of the subject -, were not adequately addressed ? This has been our experience with the problem of generating multivariate random variables ; an experience that has been an intellectual adventure worth sharing.

The basic results on generating values from a univariate distribution are well known. If X is a real random variable with distribution function F, i.e. $F(x) = P(X \leq x)$, then $U = F(X)$ has the uniform distribution on the interval (0 , 1). This result implies that if U is uniform on (0, 1) , then $X = F^{-1} (U)$ has distribution function F. So, given one has the capability to generate uniform variates, one has a general method for generating continuous random variables with a specified distribution. (See Gentle (2003) for a discussion of standard techniques for generating values from a uniform distribution.) The idea of a copula can be viewed as generalizing these basic results to higher dimensions.

Definition.

Let $I = [0, 1]$ and let $I^m = [0, 1] \times \cdots \times [0, 1]$ in $\mathbb{R}^m$. A *copula* is a distribution function C on I^m with uniform marginals.

The seminal result of the subject is :

Theorem (Sklar 1959)

If F is a distribution function on $\mathbb{R}^m$ with one dimensional marginals $F_1 , \cdots , F_m$, then there is a copula C so that :

$$F (x_1 , \cdots , x_m) = C (F_1 (x_1) , \cdots , F_m (x_m)).$$

If F is continuous, then the copula C is unique, and is given by :

$$C(u_1 , \cdots u_m) = F (F_1^{-1}(u_1) , \cdots , F_m^{-1}(u_m))$$

Note that this result of Sklar is more than a means to the end of generating multivariate random variables : for continuous distribution functions with specified marginals, the copula C *characterizes* the distribution. In particular, the copula completely characterizes the dependence structure of the distribution. As the etymology of the word, 'copula' suggests, one may view a copula as the glue that determines how the marginals are joined together. This characterization is complete, but not necessarily convenient : one may hope that the dependence structure of the distribution might be adequately specified by a *finite* number of real values, corresponding to some measure of concordance. What 'adequate' means is obviously a matter of the intended application, and is not easy to rigorously define.

II. Copulas : Some Examples and Basic Results

Many, but not all, of the basic results and examples in two dimensions generalize to an arbitrary number of dimensions. For the sake of a concise exposition, the easy generalizations are left to the reader.

Examples.

1. The *product copula* $\Pi(u, v) = uv$.
2. $M(u, v) = \min(u, v)$.
3. $W(u, v) = \max(u + v - 1, 0)$.

The last two examples are extremal, in the following sense :

Theorem (*Frechet - Hoeffding bounds*)

If $C(u, v)$ is any copula on I^2, then for all (u, v) in I^2,

$$W(u, v) \le C(u, v) \le M(u, v)$$

There are various ways to derive new copulas from old ; for example, it is easy to see that a convex combination of copulas is a copula. One method of constructing entire families of copulas is based on the following result.

Theorem. Let ϕ be a function from $(0, 1]$ onto the nonnegative reals that is decreasing and convex, then $C(u, v) = \phi^{-1}(\phi(u) + \phi(v))$ is a copula.

Such a copula is called *Archimedean* ; the function ϕ is called the generator of the copula.

Examples.

4. Taking $\phi(t) = - \ln (t)$ produces the product copula.

5. $\phi(t) = \frac{1}{\alpha}(t^{-\alpha} - 1)$, $\alpha > 1$, gives the *Clayton copula* :

$$C(u,v) = (u^{-\alpha} + v^{-\alpha} - 1)^{-\frac{1}{\alpha}}$$

An extensive list of Archimedean copulas and their corresponding generators is given on pp. 94-97 of Nelson (1995).

An algorithm for generating random vectors from a continuous distribution with specified marginals and a specified copula C is conceptually straightforward.
(More about the breezy phrase 'conceptually straightforward' shortly.)
The bivariate case indicates the general idea :

Given a pair of random variables (X, Y) with specified marginal distributions G and H, respectively, and a specified copula C(u, v), one generates observation (x , y) via the following three step procedure :

(1) Generate S, T, independent random variables, uniform on (0 , 1) .

(2) The conditional distribution function of V given U = u is

$$C_u(v) = \frac{\partial C(u,v)}{\partial u} \quad \text{- (Easy exercise to verify this ...)}$$

Given a pair (s , t) from (1) , take $w = C_s^{-1}(t)$

(3) Take $x = G^{-1}(s)$, $y = H^{-1}(w)$.

Some measures of concordance can be given in terms of the copula.
Most notably :

Theorem : The Pearson correlation coefficient, ρ , for a pair (X , Y) of continuous random variables may be expressed in terms of the copula :

$$\rho = 12 \int_{I^2} C(u, v)\, du\, dv - 3$$

Theorem : The Kendall's tau, τ , for a pair (X , Y) of continuous random variables may be expressed in terms of the copula :

$$\tau = 4 \int_{I^2} C(u, v)\, d\,C(u, v) - 1$$

Results of this type are useful when one seeks to relate measures of association to the parameters defining the copula. To illustrate :

Example.

The function $C(u, v) = uv + \theta uv(1 - u)(1 - v)$ defines a copula for values of the parameter θ in the interval [-1, 1]. (This is the *Farley-Gumbel-Morgenstern* (FGM) copula.) Use the result quoted one finds $\rho = \theta/3$.

III. Measures of Association, Parameter Matching, and Choosing a Copula

Although the dependence structure of a distribution is completely characterized by the copula, one may not find this to be the most *convenient* characterization. One might prefer to think in terms of a finite set of parameters corresponding to some set of measures of association. (e.g. the correlation matrix). When creating a synthetic data set one plays a 'parameter matching game' in which one chooses the parameters defining the copula so that the measures of association for the synthetic data set match the corresponding values for the data set one wishes to copy. One hopes that matching these measures of association produces a synthetic copy adequate to the purposes for which it is intended.

What some experience, - and experimentation -, reveals is that while the generalization from the bivariate case to the case of an arbitrary number of dimensions may be conceptually straightforward, the mass of algebra required is formidable. To some extent these difficulties may be addressed with the appropriate use of modern software for symbolic manipulation, such as *Mathematica*. However, the issue goes beyond algebraic complexity : the parameter matching previously described may actually be impossible to achieve, when, for instance, the number of parameters defining the copula is smaller than the specified number of measures of association. As another example of the potential difficulties in the playing the parameter matching game, note that for FGM copula, defined above, $-1/3 \leq \rho \leq 1/3$.

The Gaussian copula, discussed in the next section, is readily implemented in any number of dimensions, and is naturally parameterized in terms of the correlation matrix of distribution. These features are the motivation for *starting* our investigation of methods of creating synthetic copies of agricultural data sets with the Gaussian copula. Before discussing the Gaussian copula in particular, it is worthwhile to discuss the general issue of the relationship between measures of association and the choice of a copula.

Thinking about the matter on an intuitive level, a pair of random variables may exhibit a variety of dependence structures. High values of one random variable may be associated with high values of another random variable (*right tail dependence*) ; or low values of one random variable may be associated with low values of another random variable (*left tail dependence*). For some copulas there is an inherent symmetry, in the sense that right and left tail dependence are equivalent : e.g., the Gaussian and Frank copulas. For other copulas the tail dependence is asymmetric ; for instance, the Clayton copula exhibits strong left tail dependence, but weak left tail dependence. Or perhaps the dependence structure is symmetric, but varies in strength ; e.g. tail dependence in which *extreme* values of one variable correspond with extreme values of the other variable. The Gaussian copula, for example, displays stronger tail dependence than the Frank copula. Ideally, one should choose a copula which is suited to capture the particular dependence structures one feels are most vital to preserve.

Similar remarks hold for a choice for a measure of association. This is too broad a subject to go into here. The reader is directed to the discussion based on the notion of *concordance* given in Nelsen (2002.) A particular measure of association, Kendall's tau, is discussed there at length, and also in Trivedi (2005). However, some properties of Kendall's tau are of sufficient interest to justify a ....

.... Digression.

Consider the set of distributions of continuous random variables on $\mathbb{R}^2$. An equivalence relation may be defined on this set by defining an equivalence class to be the set of all distributions having a specified pair of marginal distributions. (The *Frechet - Hoeffding class.*) It is easy to show that each equivalence class of distributions has the cardinality of $\mathbb{R}$. Hence one is led to ask whether there is a real-valued measure of association which indexes the set of copulas in some meaningful way. A partial answer in the affirmative is provided by Kendall's tau. Given the natural partial ordering of the set of bivariate copulas defined by $C_1 \leq C_2$ if $C_1(u, v) \leq C_2(u, v)$ for all (u, v), then $C_1 \leq C_2$ implies $\tau_1 \leq \tau_2$. (Trivedi (2005).) Moreover, $-1 \leq \tau \leq 1$, and all values in this interval are achieved ; in particular, $\tau(W) = -1$, $\tau(M) = 1$, and $\tau(\Pi) = 0$. Trivedi argues that a desirable feature of a family copulas is that the corresponding values of Kendall's tau cover [-1 , 1].

IV. The Gaussian Copula

Definition.

Let ϕ_Σ be the distribution function of a random variable which is N(0, Σ), where Σ is the correlation matrix. Let ϕ be the distribution function for a standard normal random variable. The *Gaussian copula* is :

$$C(u_1, \cdots u_m) = \phi_\Sigma(\phi^{-1}(u_1), \cdots, \phi^{-1}(u_m))$$

Note that the Gaussian copula is naturally 'parameterized' in terms of the correlation matrix. In the particular case m = 2, for which the family of Gaussian copulas are parameterized by the correlation ρ, the Gaussian copula, C_ρ , is the Frechet - Hoeffding lower bound, W(u, v) , when $\rho = -1$; the Frechet - Hoeffding upper bound, M(u, v), when $\rho = 1$, and the product copula when $\rho = 0$ Joe (1993) .

The Gaussian copula is undoubtedly the most popular choice in current applications. Note that extensive algebra is not necessary, and one does need to deal with the copula in explicit form, there is no dimensional restriction, and the correlation structure is directly set in the copula definition. This last point is a weakness as well as a strength -

A general distribution on $\mathbb{R}^m$ is *not* uniquely determined by its correlation matrix and the m one-dimensional marginals (See the figure that follows for a graphical illustration of this fact.) That having been said, faced with the problem of generating random variates for an arbitrary distribution, matching the marginals and correlation structure may often prove to be good enough. *Or perhaps not* ...A discussion of the dangers inherent in focusing exclusively on correlation as a measure of concordance, and an introduction to alternative, - possibly more suitable measures -, is presented in Ebmrechts, et al, (1999).

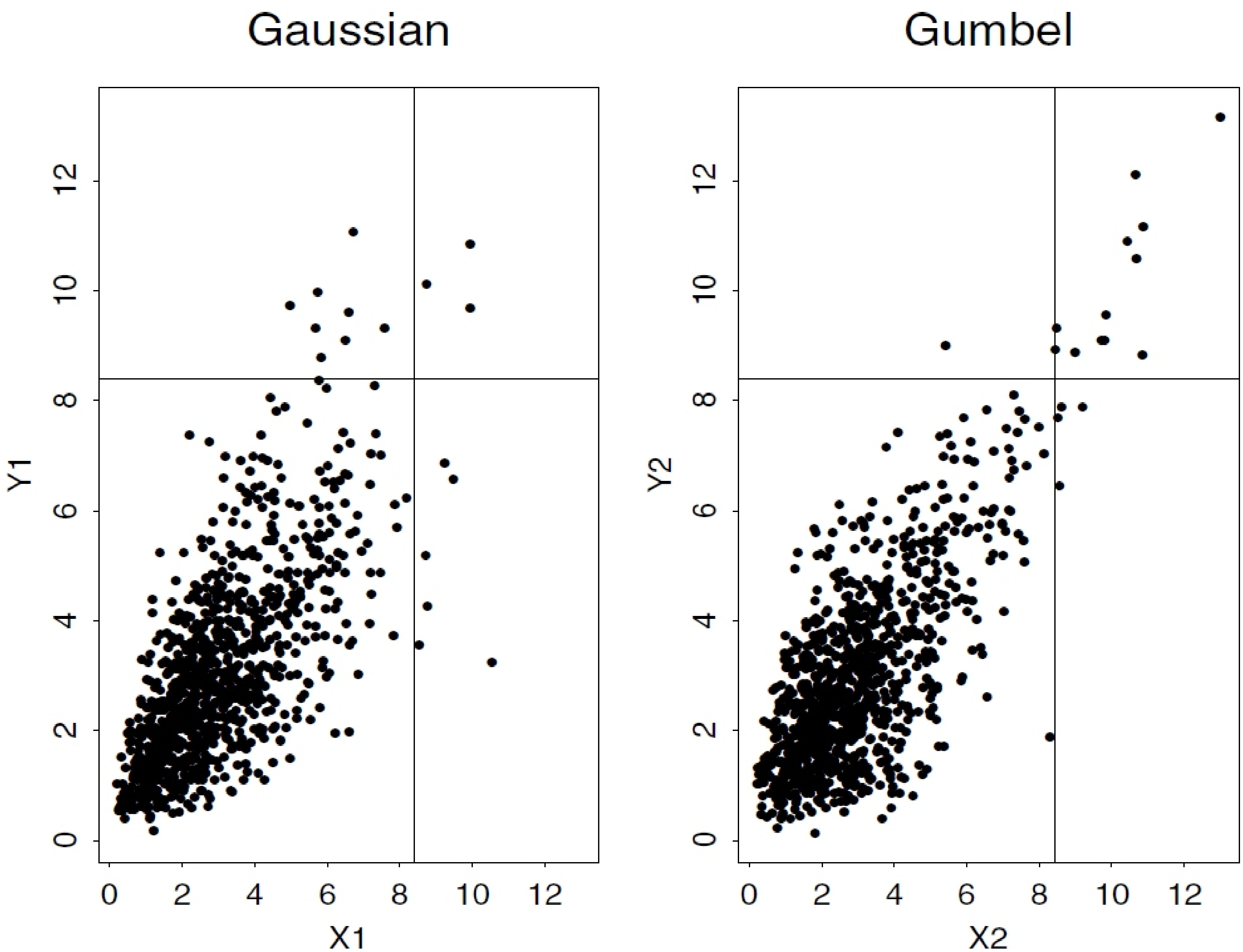


FIGURE 1. 1000 random variates from two distributions with identical Gamma(3,1) marginal distributions and identical correlation $\rho = 0.7$, but *different* dependence structures.

The procedure for generating random vectors using a Gaussian copula is easily implemented in SAS/IML, and is also easily described : given a specified correlation matrix Σ and a specified set of marginals , $F_1 , \cdots , F_m$, the algorithm for generating random variates using a Gaussian copula consists of three steps :

(1) Generate ($v_1 , \cdots , v_m$) according to a $N(0, \Sigma)$ distribution.

(2) Compute ($u_1 , \cdots , u_m$) , where $u_i = \phi^{-1}(v_i)$, $1 \leq i \leq m$.

(3) Compute ($x_1 , \cdots , x_m$) , where $x_i = F_i^{-1}(u_i)$, $1 \leq i \leq m$

(In the application to creating a synthetic copy of a data set, one should perhaps list a step (0) : Compute estimates of the marginal distribution and the correlation matrix.)

Whatever the shortcomings of the Gaussian copula, the ease with which it is implemented and its ubiquity in current applications, offer a compelling argument for starting a study of copula methods for creating synthetic data sets with an investigation of the Gaussian copula.

V. Modeling the Marginal Distributions

An essential part of any application of copula techniques to creating synthetic data sets is modeling the marginal distributions, hence we pause in the general flow of the discussion to consider this issue.

One should always be ambitious in ones goals, but still reasonable in those ambitions: one wishes to have a method of creating synthetic data sets which is *relatively* automatic and *relatively* general - A method that is computationally feasible only for generating a few dozen bivariate observations is clearly inadequate, but asking for a method to generate millions of records in five hundred variables is probably expecting too much.

In many instances the variables observed in agricultural data sets are non-negative, and the values tightly packed enough so that it is reasonable to consider the observations as realizations of a distribution with no atoms, except, possibly, a positive probability mass at zero. The present investigation will be limited to such cases.

Parametric modeling of the marginals can be time consuming, and requires expert knowledge and judgement beyond what is inherent in the data itself. A more automatic and more ‘data-driven’ approach to modeling the marginal distributions is based on simply interpolating the sample distribution of the data set one wishes to copy.

The interpolation method used in the test application described below used cubic splines to interpolate the sample distribution function for each variable.
A brief outline giving providing some background on cubic splines and its use in the present application follows :

The Interpolation Problem : Cubic Splines

Given a function F on [a, b] with values $y_0, y_1, \cdots, y_n$ at $a = x_0 < x_1 < \cdots < x_n = b$, respectively, an approximating function S for F on [a , b] so that $S(x_i) = y_i$, $0 \le i \le n$, is said to *interpolate* F.

In our application the value a corresponds to the minimum observed value of the variable in question ; the value b corresponds to the maximum observed value ; the function F corresponds to the sample cdf for the variable.

There are an infinity of different schemes for finding an interpolating function, and the interpolation problem represents an entire area of applied mathematics.
A good basic reference on numerical method is Kress (1998). Chapter 8 gives an introduction to the interpolation problem.

Definition

A *cubic spline* S(x) is an interpolating function so that on each interval $[x_j, x_{j+1}]$, $0 \le j \le n - 1$,

$$S(x) = S_j(x) = A_j x^3 + B_j x^2 + C_j x + D_j \ ;$$

subject to the conditions :

$$S(x_i) = y_i \quad , \ 0 \le j \le n$$

$$S(xj) = S(x_{j+1}) \ , \ 1 \le j \le n - 2$$

$$S'(xj) = S'(x_{j+1}) \ , \ 1 \le j \le n - 2$$

$$S''(xj) = S''(x_{j+1}) \ , \ 1 \le j \le n - 2$$

These conditions presented determine 4n - 2 linear equations in the 4n parameters A_j, B_j, C_j, D_j. Additional constraints in the form of end conditions give two more linear equations, thus giving a unique cubic spline in terms of the solution of 4n linear equations in the 4n parameters. For example, specifying values for the first derivative of S(x) at the endpoints, a and b , gives the *clamped spline* ; specifying that the second derivative vanish at the endpoints gives the *natural spline*.

The case for which each subinterval has fixed length $x_{j+1} - x_j = h = (b - a)/n$ the system of 4n defining equations in the 4n parameters is easy to set up. However, one must be careful in choosing the number of intervals to be large enough to achieve a good approximation, but not so large that some intervals contain no data points. The increased effort to set up the defining system of equations aside, it would probably be better to choose the intervals so that each corresponded to a given value of the sample cumulative distribution function. This might also greatly increase the execution time of corresponding code.

In the application to creating a synthetic copy of a data set using a Gaussian copula, the computation of an approximation to $F^{-1}(r)$, $0 < r < 1$, for a given marginal F, via $S^{-1}(r)$, comes down to :

(1) Finding the index j so that r is contained in $[y_j , y_{j+1}]$.

(2) Solving $S_j(x) = r$, for x in $[x_j , x_{j+1}]$,
e.g using Newton-Raphson iteration.

As noted before, there is some judgment to be exercised in choosing the number of intervals for each variable, but the required computations, as implemented using SAS/IML code, are generally executed quickly and with few problems.

VI. The Test Data Set

For our 'test' data set, we drew random samples from a subset of the Illinois farm operations reporting positive crop land acres on the 2007 US Census of Agriculture. Four variables were included in the analysis :

```
crop land acres. ( x1 )
soybean acres harvested ( x2 )
corn acres harvested ( x3 )
winter wheat acres harvested ( x4 )
```

There were 41,186 records with positive crop land acres (x1 > 0) ; of these, 32,391 were positive for at least one of x2, x3, and x4. From this latter set, a random sample of records was selected as the data set to be 'copied'. Of these 32,391 records, 29,416 were positive for x2 ; 26,846 were positive for x3, and 32,075 were positive for at least one of x2 and x3; 7,623 records were positive for x4.

There are some special constraints associated with this set of variables. In Illinois, soybeans (but not corn) are sometimes double cropped after winter wheat. Hence if $x4 = 0$, one must have $x2 + x3 \leq x1$. If $x4 + x2 + x3 \geq x1$ then one must have $x4 > 0$ and $x2 > 0$. Let y1, y2, y3, y4 denote the respective variable values for a record in the synthetic data set. If y4 was zero, and $y2 + y3 \geq y1$, then y1 was reset to $y2 + y3$.If y4 was positive, and $y2 + y3 \geq y1$, but y2 was zero, then y2 was reset to y4, and y1 was reset to $y2 + y3 + y4$. This editing of the synthetic data set, turned out to be largely a matter of 'tidying up': in a synthetic data set of 2000 records, usually only a few dozen failed to meet these constraints.

What were the most successful results, were arguably the least interesting results to come out the test. Briefly : for a sample of size N = 1600-2000, the SAS/IML code implementing the computations previously described ran quickly ; the marginal distributions were faithfully reproduced, as expected, and the sample correlations in the synthetic data set were not generally significantly different from the population correlations, - (although there seemed to be a slight downward bias for pairs for which one variable had a significant probability mass at zero (x4)).

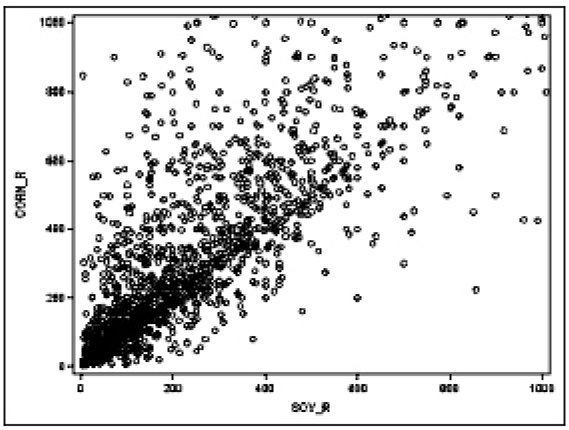
The less successful results, are more interesting. Simple graphical analyses revealed that some key features the distribution were not being captured. (A couple of bivariate plots to illustrate this follow. The second set of graphs gives a 'zoomed in' view of the real and synthetic data sets near the origin.)

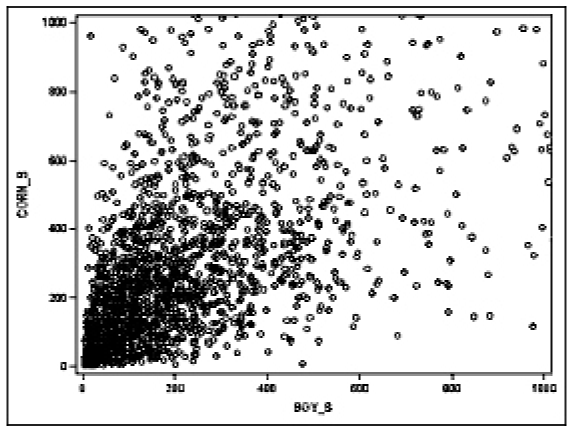
Even with these shortcomings, this methodology would certainly produce synthetic data sets adequate to many purposes. Moreover, the range of possibilities in terms of the choice of copula and measures of concordance , – these possibilities more and more frequently realized in software packages - , encourages one to further explore the application of copula based methodology to creating synthetic agricultural data sets.

Corn versus Soybean Acres
(both acreages < 1000)

Sample of REAL Data Points

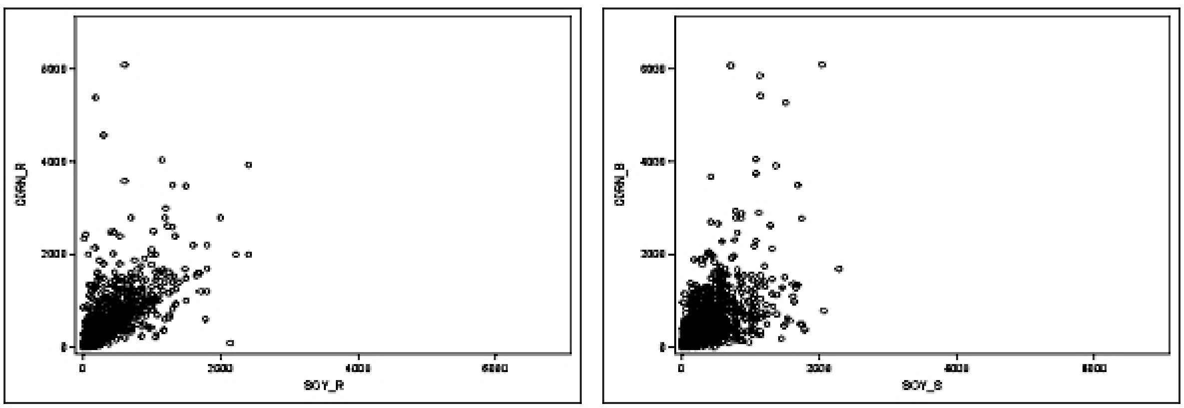
Sample of SIMULATED Data Points





Corn versus Soybean Acres
(both positive)

Sample of 2000 REAL Data Points **Sample of 2000 SIMULATED Data Points**



REFERENCES

Embrechts, P., McNeil, A., Straumann, D (1999) Correlation and Dependence in Risk Management: Properties and Pitfalls, RISK Magazine*,* May

Gentle, James E , (2003) Random Number Generation and Monte Carlo Methods, Springer

Joe, H (1993) Parametric Family of Multivariate Distributions with Given Margins, J. Multivariate Anal. 46 , 262-282

Kress, R(1998) Numerical Analysis, Springer

Nelsen R. (1995) Copulas, Characterization, Correlation and Counterexamples, Mathematics Magazine, Vol. 68, No. 3 , June, pp. 193 - 198

Nelsen R. (2002) Concordance and copulas : A survey , *Distributions with Given Marginals and Statistical Modeling* (C. M. Cuadras, J. Fortiana, and J. A. Rodríguez Lallena eds.), Kluwer Academic Publishers, pp. 169-178

Nelsen R. (2005) An Introduction to Copulas, Springer

Schweizer B. , Sklar A. (1983) Probabilistic Metric Spaces, North Holland

Sklar A, (1959), Fonctions de Reparition a n dimensions et leurs marges, Publ. Inst. Univ. Paris 8:229-231

Trivedi, P. , Zimmer, D. (2005) *Copula Modeling : An Introduction for Practitioners*, Foundations and Trends in Econometrics, Vol. 1 , No. 1, pp. 1 -111.