Dependence Structure in Agricultural Index Insurance Design
and Product Development

Dissertation

Presented in Partial Fulfillment of the Requirements for the Degree Doctor of Philosophy in the Graduate School of the Ohio State University

By

Pu Liu, M.A.S.

Graduate Program in Agricultural, Environmental and Development Economics

The Ohio State University

2010

Dissertation Committee:

Mario J. Miranda, Advisor

Elena Irwin

Stan Thompson
Abstract

Index insurance refers to a broad class of insurance products that indemnify the insured based not on verifiable losses, but rather a distinct variable that ideally is highly correlated with losses, a so-called index. Index insurance products based on rainfall, area-yields, regional livestock mortality, remotely-sensed vegetation indices, and other variables have been promoted as cost-effective risk management tools for agricultural producers in developing countries where traditional insurance is likely to fail due to high transactions costs. Index insurance is generally free of moral hazard, is less susceptible to adverse selection, and is less expensive to administer than conventional insurance. Index insurance, however, has been criticized on the grounds that, in practice, available indices are not sufficiently correlated with losses to provide effective protection against common farm or household risks.

In this dissertation, various issues are addressed pertaining to statistical methods used in the development, design, and economic assessment of index insurance products. Of special interest is whether statistical methods commonly used by actuaries and economists to rate and analyze index insurance products can adequately capture the potentially complex dependence structures that might exist among indices and losses, particularly at the tails of the distributions. I propose a
novel approach to statistical analysis of index insurance products based on copulas, and the spatial autoregressive model with variant spatial autoregressive parameters. Bivariate copulas are especially well-suited for capturing complex dependencies that exist among bivariate random variables, particularly at the tails, but have been rarely little used to analyze agricultural index insurance. The spatial autoregressive model with variant spatial autoregressive parameters is suited for characterizing tail dependence among multivariate random variables by introducing the effect of distances.

Three distinct applications are proposed. In Chapter 2 copulas and their properties are introduced. In Chapter 3, the degree of bivariate tail dependence that may exist among common indices will be assessed empirically, using bivariate copulas and Iowa county rainfall as a case study. In Chapter 4, how copulas can be used to design optimal index insurance products will be demonstrated, using Henan Province, China as a case study. In Chapter 5, I discuss how to investigate tail dependence among multivariate indices using spatial autoregressive error model.
Dedication

Dedicated to my parents, Xuezhi Liu and Mumei Zheng,

and my husband, Le Li.
Acknowledgements

I would like to express my sincere gratitude to my advisor, Professor Mario J. Miranda. Throughout my four years study at the Department of Agricultural, Environmental and Development Economics, The Ohio State University, Professor Miranda has given me constant guidance, invaluable advice and warm encouragement. It is his tremendous support that makes this dissertation possible. He has devoted great support to my whole Ph.D. study and career development. His admirable personality and professional ability are of immense benefit to my study and my life in the future.

I also want to thank Professor Elena Irwin, Professor Abdoul Sam and Professor Stan Thompson. They offered me a great deal of time and efforts to not only this dissertation, but also to my career development. I do appreciate their valuable suggestions and comments on this dissertation and my career development.

I would like to express my sincere thanks to Professor Claudio Gonzalez-Vega, Professor Tim Haab, Professor Matthew Roberts and Professor Cam Thraem. Thank you so much for your time and help.

My special thank go to my parents and husband. None of my achievements would have been possible without your support. Your trustful and unlimited love
encouraged me in every step of my life and will always be on my side.
Vita

2003…………………………………B.A. Finance, Nankai University

2006…………………………………M.A. Finance, Nankai University

2010…………………………………M.A.S. Statistics, The Ohio State University

Fields of Study

Major Field: Agricultural, Environmental, and Development Economics
Table of Contents

Abstract ........................................................................................................................................... ii

Dedication ........................................................................................................................................ iv

Acknowledgements ........................................................................................................................... v

Vita ................................................................................................................................................... vii

List of Tables .................................................................................................................................. xi

List of Figures .................................................................................................................................. xiii

CHAPTER 1: INTRODUCTION ................................................................................................. 1

CHAPTER 2: COPULAS ............................................................................................................. 7

  2.1 Definition of Copulas ............................................................................................................. 7
  2.2 Parametric Families of Copulas .......................................................................................... 9
  2.3 Dependence ............................................................................................................................ 11
  2.4 Tail Dependence ................................................................................................................... 15
  2.5 Empirical Estimation and Selection of Copulas ................................................................. 19
    2.5.1 Estimation of Copulas .................................................................................................... 19
    2.5.2 Selection of Copulas ..................................................................................................... 21

CHAPTER 3: TAIL DEPENDENCE AMONG WEATHER INDICES: THE CASE

viii
5.3.2 Estimation Results ................................................................. 70

5.4 Panel Data Model........................................................................ 73

5.4.1 The Model ............................................................................. 73

5.4.2 A Monte Carlo Simulation ..................................................... 74

5.4.3 The Case of Iowa County-Level Rainfalls............................... 79

5.5 Conclusion .............................................................................. 80

References ..................................................................................... 61
List of Tables

Table 2.1  Archimedean Copula Generator Functions........................................10
Table 2.2  Extreme-Value Copula Dependence Functions..................................11
Table 2.3  Kendall’s Tau and Spearman’s Rho for Selected Archimedean Copulas as a Function of the Generator Function Parameter $\theta$.................................13
Table 2.4  Lower and Upper Asymptotic Tail Dependence for Archimedean Copulas.............................................................................................................16
Table 3.1  Number of Taus with Higher Values in Drought and Flood Years........30
Table 3.2  Number of Adjacent Taus with Higher Values in Drought and Flood Years..................................................................................................................33
Table 3.3  Percentage of Copulas’ Rankings in Adjacent Counties......................36
Table 3.4  Summary of Estimates of Parameters for Three Archimedean Copulas and the Average Tail Dependence of Adjacent Counties.................................39
Table 4.1  Agricultural Insurance Sold in China in 2007.......................................43
Table 4.2  Willingness to Pay and Contract Parameters.....................................55
Table 5.1  The Number of Years in Which Statistics are Greater than the Critical Value...........................................................................................................67
Table 5.2  Estimates of Parameters in Equation (5).............................................71
<table>
<thead>
<tr>
<th>Table 5.3</th>
<th>RMSEs of the Estimates of $\beta_0$</th>
<th>77</th>
</tr>
</thead>
<tbody>
<tr>
<td>Table 5.4</td>
<td>RMSEs of the Estimates of $\beta_1$</td>
<td>79</td>
</tr>
<tr>
<td>Table 5.5</td>
<td>The Estimation Results</td>
<td>80</td>
</tr>
</tbody>
</table>
List of Figures

Figure 2.1  Probability Density Contour Plots of (X,Y) with Standard Normal Marginals and the Archimedean Copulas…………………………………...17

Figure 4.1  Wheat Yields and Trend in Henan……………………………………..49

Figure 4.2  Histogram of Detrended Yield in Henan, 1951-2006…………………..49

Figure 4.3  Histogram of March Rainfall in Henan 1951-2006…………………..50

Figure 4.4  Relationship between Rainfall and Indemnity………………………...53

Figure 4.5  Clayton Copula…………………………………………………………..57

Figure 4.6  Frank Copula……………………………………………………………..57

Figure 4.7  Gumbel Copula…………………………………………………………...58

Figure 5.1  Plots of the Autoregressive Parameter $\rho$ with Respect to $x$...........72
Chapter 1: Introduction

Researchers and practitioners in the field of development finance have in recent years exhibited growing interest in the use of index insurance contracts to manage the risks faced by poor agricultural producers (e.g., Miranda and Vedenov 2001; Barnett and Mahul 2007; Bryla and Syroka 2007). Unlike conventional insurance, which indemnifies the insured based on verifiable losses, index insurance indemnifies the insured based on the observed value of a specified “index”. Ideally, an index is a random variable that is objectively observable, reliably measurable, and highly correlated with the losses of the insured, and which additionally cannot be influenced by the actions of the insurer or the insured. Indices that have been employed or proposed in the design of index insurance contracts for agricultural risk management include area-yields, rainfall, temperature, satellite-measured vegetation indices, and regional livestock mortality rates (Skees, Hartell and Hao 2006; Khalil et al. 2007).

Index insurance avoids many of the problems that have plagued conventional agricultural insurance, which indemnifies based on verifiable farm-level losses. Because the insured cannot significantly influence the value of the index, and thus the indemnity paid by the contract, index insurance is
essentially free of moral hazard. Because an index insurance contract's indemnity schedule and premium rates are typically based on publicly available information, not privately held information, index insurance is largely free of adverse selection problems. And because index insurance does not require individually-tailored terms of indemnification or separate verification of individual loss claims, index insurance is less expensive to administer. These features of index insurance can substantially reduce its cost relative to conventional insurance, making index insurance more affordable, particularly to agricultural producers of limited financial means (Skees 2008).

Index insurance, however, suffers from the drawback that it does not cover all losses that may be experienced by an agricultural producer (e.g., Miranda 1991; Doherty and Richter 2002). In particular, since the indemnity provided by index insurance is based on an index, rather than verifiable losses, it is possible for the insured to suffer a significant loss without the insurance contract providing an indemnity. The potential benefits of index insurance ultimately depend on the statistical relation between the indemnities provided by the index insurance contract and the losses suffered by the insured.

There are many statistical and actuarial assessments that must be made in the design and development of index insurance products for the developing countries. For example, the expectation of the indemnities to be paid by an index insurance contract must be estimated as the first step in setting the premium. The standard
error of the expected indemnity estimate must also be computed to allow the insurer to appropriately load the premium for parametric uncertainty. The probability distribution of indemnity payments for a portfolio of index insurance contracts must be derived in order to compute the maximum probable loss for an insurer's book of business. And how well index insurance indemnities match the losses of the insured must be reliably estimated to assess the potential demand for an index insurance product.

The precision with which these actuarial computations can be performed ultimately depend on the joint distributions of indices and losses and the statistical methods used to model them. Actuaries have developed a broad range of statistical techniques to model conventional loss claim distributions based primarily on univariate statistical models that assume independence across claims. Univariate statistical methods, however, are of limited applicability to index insurance because, unlike conventional insurance, index insurance indemnities are based, not on verifiable losses, but rather a distinct random variable, the index. Also, sets of indices, such as rainfalls at different locations within a defined geographical area, usually exhibit correlation due to systemic weather effects, making the assumption of independence untenable for index insurance portfolio analysis and undesirable for efficient premium rates computation.

Economists have performed a substantial number statistical studies that have raised doubts about the value of index insurance as a risk management tool
The primary criticism of index insurance is that it is possible for the insured to suffer a loss yet not receive an indemnity. The potential severity of this problem, often referred to as “basis risk”, is typically measured empirically using the Pearson linear correlation coefficient, a statistical measure of the degree of linear dependence that exists between a pair of random variables: the lower the linear correlation, the greater the basis risk.

Linear correlation coefficients between losses and indices, however, are unreliable measures of basis risk for two reasons. First, if an index is strongly related to losses, but in a nonlinear fashion, then basis risk can easily be addressed by the use of an appropriate nonlinear indemnity schedule. Linear correlation coefficients, however, are by design unable to detect nonlinear relationships. Second, index insurance contracts provide indemnities only for extreme losses associated with extreme values of the index. The relationship between extreme values of losses and an index could be strong, yet be missed by an empirical linear correlation coefficient because it weighs all observations equally.

Actuarial and statistical assessments of index insurance products call for the use of flexible multivariate statistical methods that can faithfully capture the potentially nonlinear distributional dependencies that exists among indices and between specific indices and insurable losses, particularly in the extremes of the distributions. Copulas, which provide a theoretical framework for capturing
complex dependencies among random variables, are well-suited for this task.

Financial analysts began to take a strong interest in copulas as a result of the financial crisis of 2007-9 (e.g., Ang and Chen 2002; Embrechts, McNeil, and Straumann 1999; Hotta, Lucas, and Palaro 2008; Jouanin et al. 2001; Fantazzini 2008; Stefanova 2007; and Embrechts, Lindskog, and McNeil 2001). As a result of the crisis, financial analysts began to ask whether stock returns are more highly correlated during financial crises than in normal times, thus rendering stock portfolios riskier than predicted by conventional asset pricing models. The questions being addressed by financial analysts are analogous to those that must be addressed in index insurance design: in both cases, one is concerned with the degree of dependence exhibited by two or more random variables at the extremes of their distribution. Copulas provide a formal framework for addressing questions such as these. However, copulas remain rarely little used in index insurance applications.

In this dissertation, I address various issues pertaining to statistical methods used in the development, design, and economic assessment of index insurance products. Of special interest is whether statistical methods commonly used by economists can adequately capture the potentially complex dependence structures that might exist among indices and losses, particularly at the tails of the distributions. In the following section, Chapter 2, I review the basic features of copulas and notions of dependence that are relevant to index insurance analysis. In Chapter 3, I describe the research aimed at detecting higher tail dependence among common
weather indices, using Iowa county rainfall as a case study. In Chapter 4, I demonstrate how copulas can be used to design optimal index insurance products, using Henan Province, China as a case study. In Chapter 5, I discuss how to investigate tail dependence among multivariate indices using spatial autoregressive error model. A Monte Carlo simulation is conducted and the Iowa county-level June rainfall is used as case study.
Chapter 2: Copulas

2.1 Definition of Copulas

A copula is a function that describes how univariate marginal distributions are combined together to form a multivariate distribution (Nelsen 2006; Embrechts, Lindskog, and McNeil 2001; Trivedi and Zimmer 2007; Yan 2007). Formally, a two-dimensional copula $C(u,v)$ can be written as a function $C : [0,1]^2 \to [0,1]$ such that (Nelsen 2006)

1. $C(u,0) = C(0,v) = 0, \forall u, v \in [0,1]$
2. $C(u,1) = u$ and $C(1,v) = v, \forall u, v \in [0,1]$
3. $C(u_2,v_2) - C(u_2,v_1) - C(u_1,v_2) + C(u_1,v_1) \geq 0, \forall u_1 \leq u_2, v_1 \leq v_2$.

Equivalently, an $n$-dimensional copula is a joint cumulative distribution function of $n$ interdependent random variables, each of which, on the margin, is uniformly distributed on the unit interval.

The role that copulas play in capturing the interdependency among jointly distributed random variables is explained by Sklar's Theorem (Nelson 2006). Sklar's Theorem states that any continuous $n$-dimensional cumulative distribution function $F : \mathbb{R}^n \to [0,1]$ can be uniquely written
\[ F(x_1, x_2, \ldots, x_n) = C(F_1(x_1), F_2(x_2), \ldots, F_n(x_n)), \] (1)

where \( C \) is an \( n \)-dimensional copula and \( F_i \) is the \( i^{th} \) marginal cumulative distribution function associated with \( F \). Conversely, if \( C \) is an \( n \)-dimensional copula and \( F_i : \mathbb{R} \to [0,1] \) is a univariate cumulative distribution function, then \( F \) as defined above is a cumulative distribution function on \( \mathbb{R}^n \) with marginal cumulative distributions \( F_i \). The joint probability density function associated with a differentiable cumulative distribution function \( F \) can be recovered from its copula decomposition through the relation

\[ f(x_1, x_2, \ldots, x_n) = c(F_1(x_1), F_2(x_2), \ldots, F_n(x_n)) \prod_{i=1}^n f_i(x_i), \] (2)

where \( c \) is the joint probability density function associated with \( C \) and \( f_i \) is the univariate probability density function associated with \( F_i \).

The general properties of copulas, stated here for two jointly distributed random variables \( X_1 \) and \( X_2 \), include: (1) \( X_1 \) and \( X_2 \) are independent if and only if \( C(u,v) = uv \); (2) the copula of \( X_1 \) and \( X_2 \) is invariant under strictly increasing transformations of \( X_1 \) and \( X_2 \); (3) for any copula \( C \) and for all \( u \) and \( v \), there is a upper bound and a lower bound for copulas called Fréchet-Hoeffding bounds,

\[ \max(u + v - 1,0) \leq C(u,v) \leq \min(u,v) . \] (3)

Copulas can be useful in index insurance analysis because they provide a general, flexible framework for modeling the joint distributions of indices and losses, which is the essential first step required when estimating premium rates and assessing basis risk. When using copulas to model indices and losses, one is free to specify the
forms of marginal distributions independently of one another and independently of the form of the copula function. For example, in building a bivariate model of an index and losses, it is possible to posit that one of the two random variables is log-normally distributed, the other is beta distributed, and the dependency between the two is captured by a Student-t copula. This flexibility is particularly useful in index insurance design given that there is often no reason to suppose that the index and loss distributions belong to the same distributional family. The flexibility also offers the modeler the freedom to search among different copula functions to find the one that best explains the observed dependency between index and losses.

2.2 Parametric Families of Copulas

A number of parametric families of copulas are commonly used in statistical analysis of dependence. The two most frequently used parametric copula families are elliptical copulas, which include the Gaussian and Student-t copulas. The two-dimensional Gaussian copula distribution is (Freez and Valdez 1998):

$$C(u, v) = \Phi_{\rho}(\Phi^{-1}(u), \Phi^{-1}(v)),$$ for any $u, v \in [0,1]$, (4)

where, $\Phi^{-1}$ is the inverse of standard normal cumulative distribution function, and $\Phi_{\rho}$ represents the standard bivariate normal distribution with correlation $\rho$.

Two-dimensional Student-t copula is defined analogously to the Gaussian copula by using a multivariate extension of the $t$ distribution with parameter $\alpha$:
Another widely studied parametric family of copulas is the Archimedean copulas. An Archimedean copula is not derived directly from Sklar’s Theorem, but is constructed through a generator $\phi$:

$$C(u, v) = \phi^{-1}[\phi(u) + \phi(v)],$$

where $\phi:[0,1] \rightarrow [0, \infty)$ is a continuous, strictly decreasing, convex function with $\phi(1) = 0$ (Nelsen 2006). Three widely used one-parameter Archimedean copulas are shown in Table 2.1.

<table>
<thead>
<tr>
<th>Family</th>
<th>Parameter</th>
<th>$\phi_{\theta}(t)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Clayton</td>
<td>$\theta \geq -1, \theta \neq 0$</td>
<td>$\frac{1}{\theta}(t^{-\theta} - 1)$</td>
</tr>
<tr>
<td>Frank</td>
<td>$\theta \neq 0$</td>
<td>$-\ln \frac{e^\theta - 1}{e^\theta - 1}$</td>
</tr>
<tr>
<td>Gumbel</td>
<td>$\theta \geq 1$</td>
<td>$(-\ln t)^\theta$</td>
</tr>
</tbody>
</table>

Table 2.1 Archimedean Copula Generator Functions

Another parametric family of copulas that have been used in financial analysis and may prove applicable to index insurance design is the “extreme-value” copulas. Based on extreme value theory, the two-dimensional extreme-value copula function has the form (Nelsen, 2006):
\[ C(u, v) = \exp[\log(uv) \cdot A(\frac{\log(v)}{\log(uv)})], \quad (7) \]

where \( A \) is a mapping of the interval \([0,1] \rightarrow [0.5,1]\) and must satisfy certain conditions: (1) \( A(0)=A(1)=1 \), (2) \( \max(t,1-t) \leq A(t) \leq 1 \) for all \( t \in [0,1] \), and (3) \( A \) is convex. \( A \) is often referred to as the “Pickands dependence function”. The most widely used extreme value copulas are Marshall-Olkin copula and Gumbel-Hougaard copula. The parameters and dependence functions of the two copulas are shown in Table 2.2.

<table>
<thead>
<tr>
<th>Family</th>
<th>Parameters</th>
<th>( A(t) )</th>
</tr>
</thead>
<tbody>
<tr>
<td>Marshall-Olkin</td>
<td>0 \leq \alpha, \beta \leq 1</td>
<td>1 - \min(\beta t, \alpha(1-t))</td>
</tr>
<tr>
<td>Gumbel-Hougaard</td>
<td>\theta &gt; 1</td>
<td>((t^\theta + (1-t)^\theta)^{1/\theta})</td>
</tr>
</tbody>
</table>

Table 2.2 Extreme-Value Copula Dependence Functions

### 2.3 Dependence

Copulas allow an analyst to isolate, study, and model dependence among random variables independently of their marginal distributions. But what does one mean by “dependence”? How does one measure it? And how does one best capture it for the purposes of index insurance product design and market development?
The most widely used measure of dependence between two random variables is the Pearson linear correlation coefficient. It is obtained by dividing the covariance of the two variables by the product of their standard deviations. Specifically, the Pearson correlation coefficient of two variables $X$ and $Y$ is defined by

$$
\rho_{X,Y} = \frac{\text{cov}(X,Y)}{\sigma_X \sigma_Y},
$$

where $\text{cov}(X,Y)$ is the covariance of $X$ and $Y$, and $\sigma_X$ and $\sigma_Y$ are the standard deviations of $X$ and $Y$, respectively.

A problem with the linear correlation coefficient is that it attempts to summarize in a single number the global linear dependence between two random variables. However, one cannot expect the linear correlation coefficient to adequately summarize complex dependencies. In index insurance analysis, one is generally concerned only with the tails of the distributions. Moreover, one is also primarily concerned with whether the dependency is monotonic, without regard to whether it is linear. The Pearson correlation coefficient, which measures linear correlation globally, is thus inadequate for index insurance applications.

More informative measures of association for the purposes of index insurance analysis are the Kendall's tau and Spearman's rho measures of association. Kendall’s tau measures a form of dependence known as concordance (Kruskal 1958; Hollander and Wolfe 1973; Lehmann 1975). Informally, a pair of random variables is concordance if “large” values of one tend to be associated with “large” values of the other and “small” values of one with “small” values of the other. For two-dimensional copulas $C$, Kendall’s tau is given by (Nelsen 2006):
and Spearman’s rho by:

$$\rho_c = 12 \int uv dC(u, v) - 3 = 12E(uv) - 3;$$  \hfill (9)

where $E$ is the expectation operator. With an Archimedean copula $C$ with generator function $\varphi$, the Kendall’s tau is:

$$\tau_c = 1 + 4 \int_0^1 \frac{\varphi(t)}{\varphi'(t)} dt. \quad \hfill (10)$$

Nelsen (2006) gives the Kendall’s tau and Spearman’s Rho for selected Archimedean copulas, shown in Table 2.3. Both Kendall’s tau and Spearman’s rho are invariant under strictly increasing transformations of the random variables. This property is referred to as “scale-invariance”.

<table>
<thead>
<tr>
<th>Family</th>
<th>( \tau )</th>
<th>( \rho )</th>
</tr>
</thead>
<tbody>
<tr>
<td>Clayton</td>
<td>( \frac{\theta}{\theta + 2} )</td>
<td>Complicated closed form</td>
</tr>
<tr>
<td>Frank</td>
<td>( 1 - \frac{4}{\theta}(1 - D_1(\theta)) )</td>
<td>( 1 - \frac{12}{\theta}(D_1(\theta) - D_2(\theta)) )</td>
</tr>
<tr>
<td>Gumbel</td>
<td>( \frac{\theta - 1}{\theta} )</td>
<td>No closed form</td>
</tr>
</tbody>
</table>

* \( D_k(\theta) = \frac{k}{\theta} \int_0^\theta t^{k-1} dt \)

Table 2.3 Kendall’s Tau and Spearman’s Rho for Selected Archimedean Copulas as a Function of the Generator Function Parameter \( \theta \).

\(^{1}\) There is no closed-form expression for Spearman’s rho with Archimedean copulas.
Given data, Kendall’s tau and Spearman’s rho can easily be estimated using their sample counterparts. The sample Kendall’s $\tau$ is expressed as:

$$\tau = \frac{n_c - n_d}{0.5n(n-1)}, \quad (11)$$

where $n_c$ is the number of concordant pairs, and $n_d$ is the number of discordant pairs in the data set with $n$ pairs. According to Nelsen (2006), concordance of a pair of random variables means if “large” values of one tend to be associated with “large” values of the other one, and “small” values of one with “small” values of the other one. To be more specific, let $(x_i, y_i)$ and $(x_j, y_j)$ be two observations from paired variables $X$ and $Y$. $(x_i, y_i)$ and $(x_j, y_j)$ are concordant if $x_i < x_j$ and $y_i < y_j$, or if $x_i > x_j$ and $y_i > y_j$. Similarly, $(x_i, y_i)$ and $(x_j, y_j)$ are discordant if $x_i < x_j$ and $y_i > y_j$, or if $x_i > x_j$ and $y_i < y_j$.

The sample Spearman’s rho is also based on concordance and discordance (Kruskal 1958; Lehmann 1966):

$$\rho = 1 - \frac{6\sum d_i^2}{n(n^2-1)}, \quad (12)$$

where $d_i$ is the difference between the ranks of corresponding values $X_i$ and $Y_i$, and $n$ is the number of observations in the data set.

Because index insurance indemnity schedules can be nonlinear, the ability to write an effective insurance contract for a particular loss distribution depends ultimately on whether the index and losses exhibit strong monotonic dependence, not strict linear dependence. As such, Kendall's tau and Spearman's rho provide
measures of association that are superior to Pearson's linear correlation coefficient for
assessing the viability of index insurance contracts. However, both Kendall's tau and
Spearman's rho remain global measures of association that are unable to capture
variations in the degrees of dependence in the tails of the distribution.

2.4 Tail Dependence

Efforts to explain asymmetries in dependence at extremes have lead to the
introduction of the notion of tail dependence. Tail dependence measures dependence
between two random variables in the upper-right quadrant and in the lower-left
quadrant of their domain (Nelsen 2006). Tail dependence characterizes how large
(or small) values of one random variable appear when the other random variable has
large (or small) values. The parameter of asymptotic lower tail dependence \( \lambda_L \)
is the probability in the limit that one variable takes a very low value, given that the
other takes a very low value. Similarly, the parameter of asymptotic upper tail
dependence \( \lambda_U \) is the probability in the limit that one variable takes a very high
value, given that the other takes a very high value. The two parameters for tail
dependence are formally defined as:

\[
\lambda_L = \lim_{t \to 0^+} P(Y \leq \Phi^{-1}(t) \mid X \leq \Phi^{-1}(t)),
\]

(13)

\[
\lambda_U = \lim_{t \to 0^+} P(Y > \Phi^{-1}(t) \mid X > \Phi^{-1}(t)).
\]

(14)

where \( X \) and \( Y \) are two random variables, and \( G \) and \( F \) are the marginal cumulative
distribution functions of $X$ and $Y$, respectively. The asymptotic tail dependence parameters for copula function are shown as following (Nelsen 2006),

$$
\lambda_L = \lim_{t \to 0^+} \frac{C(t, t)}{t},
$$

$$
\lambda_U = 2 - \lim_{t \to 0^+} \frac{1 - C(t, t)}{1 - t}.
$$

For some of the families of copulas, the tail dependence parameters $\lambda_L$ and $\lambda_U$ can be easily evaluated, while for others, they can only be solved numerically. In the case of Gaussian copula and Student-t copula, the copula functions are symmetric, implying the asymptotic upper and lower tail dependences are identical.

For an Archimedean copula with generator $\varphi$, the tail dependence parameters are:

$$
\lambda_L = \lim_{t \to 0^+} \frac{\varphi^{-1}(2\varphi(t))}{t} = \lim_{x \to 0^+} \frac{\varphi^{-1}(2x)}{\varphi^{-1}(x)},
$$

$$
\lambda_U = 2 - \lim_{t \to 0^+} \frac{1 - \varphi^{-1}(2\varphi(t))}{1 - t} = 2 - \lim_{x \to 0^+} \frac{1 - \varphi^{-1}(2x)}{1 - \varphi^{-1}(x)}.
$$

The asymptotic parameters for some Archimedean copulas are summarized in Table 2.4.

<table>
<thead>
<tr>
<th>Family</th>
<th>$\lambda_L$</th>
<th>$\lambda_U$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Clayton, $\theta \geq 0$</td>
<td>$2^{-1/\theta}$</td>
<td>0</td>
</tr>
<tr>
<td>Frank</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Gumbel</td>
<td>0</td>
<td>$2 - 2^{1/\theta}$</td>
</tr>
</tbody>
</table>

Table 2.4 Lower and Upper Asymptotic Tail Dependence for Archimedean Copulas
Figure 2.1 shows the probability density contour plots of Clayton and Gumbel copulas. As shown in the figures, the two copulas allow for different asymmetric tail behavior. The Clayton copula has higher dependence on the lower-left quadrant, while Gumbel copula has higher dependence on the upper-right quadrant.

![Contour plots of Clayton and Gumbel copulas](image)

**Figure 2.1** Probability density contour plots of \((X, Y)\) with standard normal marginals and the Archimedean copulas

Whether the proposed index and the losses exhibit tail dependence is of fundamental interest in weather index insurance design. Weather index insurance pays indemnities based not on actual losses experienced by the policyholder but rather on realizations of a weather index that is highly correlated with actual losses (Barnett and Mahul, 2007). A trigger value is specified to establish the range of value over which or less than which indemnity will be paid. For example, the trigger value of
rainfall would be specified for the disasters caused by too much rain, say flood, or too little rain, say drought. Since the disasters often happen with abnormally large or small values of the index, and losses usually happen when the realized yield is much less than “ordinary” yield, the basis risk that the contract designers want to minimize depends not on the global dependence, but rather on the upper-right or lower-left quadrant tail dependence. For example, if the index insurance contract is designed for drought, the dependence of indemnity and losses is considered only on the lower tail of rainfall. A copula function that is used to model the dependence of rainfall and yield should exhibit higher dependence at the lower tail.

The nature of tail dependence among a series of indices is also of fundamental interest in index insurance design. Similar to the financial analysts who diversifies a olio by optimally combining individual stocks, insurers attempt to diversify their “portfolio” of insurance contracts by offering products in different areas. If the indices associated with different areas exhibit lower tail dependence, then they are more highly correlated for extreme low values than for mid-range or extreme high values. This would imply that the insurer’s portfolio of insurance contracts is riskier, and the insurer’s maximum probable losses would be higher, than would be indicated by conventional mean-variance theory.
2.5 Empirical Estimation and Selection of Copulas

2.5.1 Estimation of Copulas

Various estimation procedures based on maximum likelihood principles are available to estimate copulas. First, if one is not sure about the form of the marginal distributions, it is possible to estimate the copula's dependence parameter using the method of pseudo-maximum likelihood suggested by Genest and Favre (2007). When the copula $C$ is absolutely continuous with density $c$, the log-likelihood function is:

$$\ell(\theta) = \sum_{i=1}^{n} \log[c\left(\frac{R_i}{n+1}, \frac{S_i}{n+1}; \theta\right)],$$

(19)

where $\theta$ is the dependence parameter, $R_i = \sum_{i=1}^{n} 1(X_i \leq x)$ and $S_i = \sum_{i=1}^{n} 1(Y_i \leq y)$.

If one is willing to specify the form of the marginal distributions, it is possible to estimate the dependence parameter of the copula and the parameters of the marginal distributions jointly via the method of full-information maximum likelihood. The log-likelihood function is:

$$\ell(\beta_1, \beta_2, \theta) = \sum_{i=1}^{n} \log[c(F(X_i; \beta_1), G(Y_i; \beta_2); \theta)],$$

(20)

where $\beta_1$ and $\beta_2$ are the parameters for marginal distributions, $\theta$ is the dependence parameter, $F$ and $G$ are the marginal cumulative distribution functions for
$X$ and $Y$, respectively. Parameters $\theta$, $\beta_1$ and $\beta_2$ can be estimated simultaneously by maximizing the log-likelihood function $\ell(\beta_1, \beta_2, \theta)$.

The parameters of the copula and the parameters of the marginals can also be estimated via limited information maximum likelihood using a two step procedure known in the literature as the “inference function for marginals” method (Joe 1997). In the first stage, the marginal parameters $\beta_1$ and $\beta_2$ is estimated by maximizing the score functions from the two marginal distributions of the multivariate copula model, respectively. The estimates $\hat{\beta}_1$ and $\hat{\beta}_2$ are expressed as

$$\hat{\beta}_i = \arg \max_{\beta} \sum_{t=1}^{T} \log f_i(x_{it}; \beta), \quad i = 1, 2.$$  \hspace{1cm} (21)

where $f_i$ stands for the marginal density function of the $i^{th}$ marginal distribution $F_i$. $T$ represents the number of observations. In the second stage, given the marginal estimates $\hat{\beta}$, the copula parameter $\theta$ is estimated by solving the maximum likelihood function:

$$\ell(\theta) = \sum_{i=1}^{n} \log[c(F(X_i; \hat{\beta}_1), G(Y_i; \hat{\beta}_2); \theta)].$$  \hspace{1cm} (22)

In this chapter, the marginal distributions of rainfalls are examined by checking the histograms of rainfall data. For the purpose of convenience, I used two-step MLE. First, I estimated the marginal parameters and examined the errors. The log-normal distribution appeared to offer the best fit for the marginal distributions. I then fixed the estimates of marginal parameters and estimated the copula parameter by MLE in the second stage.
2.5.2 Selection of Copulas

Another important task in working with copulas in empirical applications is to choose among various candidate copulas. In statistics, a “goodness-of-fit” statistic is a measure of how well a statistical model fits a set of observations. Some tests based on goodness-of-fit statistics may be used to compare the performance of copulas. The Pearson’s chi-square test and Kolmogorov-Smirnov test are most widely used.

Pearson’s chi-square test is often used in analyzing the results of experiments with different treatments. The Pearson’s chi-square test statistic is

\[ X^2 = \sum_{i=1}^{n} \frac{(O_i - E_i)^2}{E_i}, \]

where \( n \) is the number of bins, \( O_i \) represents an observed frequency for the \( i \)th bin, \( E_i \) represents an expected frequency asserted by the fitted distribution for the \( i \)th bin, and statistic \( X^2 \) asymptotically approaches a chi-square distribution with degree of freedom equal to the number of observations, minus the reduction in degree of freedom.

The Kolmogorov-Smirnov statistic for a given cumulative distribution function \( F(x) \) is

\[ D_n = \sup_x |F_n(x) - F(x)|, \]

where \( F_n(x) = \frac{1}{n} \sum_{i=1}^{n} I(X_i \leq x) \) represents the empirical distribution function with the indicator function \( I(X_i \leq x) \). Under the null hypothesis that the sample comes from the fitted distribution \( F(x) \), \( \sqrt{n}D_n \) converges to the Kolmogorov distribution. The cumulative distribution function of
the random variable $K$ with Kolmogorov distribution is

$$
Pr(K \leq x) = 1 - 2 \sum_{i=1}^{\infty} (-1)^{i-1} e^{-2i^2 x^2}.
$$

(23)

The problem with Kolmogorov-Smirnov test is that it is complicated when it is applied to multivariate data. In the case of two-dimension, the maximum difference will differ depending on which of $Pr(x < X \land y < Y)$ or $Pr(x > X \land y < Y)$ or any of the other two possible arrangements is used.

In addition to Pearson’s chi-square test and Kolmogorov-Smirnov test, the Akaike’s Information Criterion (AIC) is another measure of the goodness-of-fit of a statistical model. The AIC is not a test in the sense of hypothesis testing, but rather a tool for model selection. In general, AIC is expressed as

$$
AIC = 2k + n \left( \ln \frac{RSS}{n} \right),
$$

where $k$ is the number of parameters in the statistical model, and $RSS$ is the residual sum of squares, $RSS = \sum_{i=1}^{n} \hat{\epsilon}_i^2$. Given a data set, several competing models may be compared according to their AIC. The one with the lowest AIC is the best. Since AIC prefers the model with higher degree of freedom, as the sample size increases there is an increasing tendency to accept the more complex model. This attribute of AIC makes the comparison of models using Monte Carlo Simulation unreliable.

Genest, Quessy and Remillard (2006) develop a Goodness-of-fit statistic for copulas and apply parametric bootstrapping to compare the fit provided by copulas. Suppose $F(x_1, x_2)$ is the joint distribution based on specific copula function and marginals. Let $K(\theta, t) = P\{F(x_1, x_2) \leq t\}$ with the copula parameter $\theta$. The
empirical version of $K(\theta, t)$ is defined as

$$K_n(t) = \frac{1}{n} \sum_{j=1}^{n} \mathbb{1}(V_j \leq t), \quad t \in [0,1], \quad (24)$$

where $n$ represents the size of sample, $V_j$ are pseudo-observations defined by

$$V_j = \frac{1}{n} \sum_{k=1}^{n} \mathbb{1}(X_{1k} \leq X_{1j}, X_{2k} \leq X_{2j}),$$

and $\mathbb{1}(V_j \leq t)$ refers to the indicator function that has the value of 1 when $V_j \leq t$ and the value of 0 when $V_j > t$. The Goodness-of-fit statistic for copulas is given by

$$S = \int_{0}^{1} |\gamma(t)| \, k(\theta,t) \, dt, \quad (25)$$

where $\gamma(t) = \sqrt{n}[K_n(t) - K(\theta,t)]$ and $k(\theta,t)$ represents the density function of $K(\theta,t)$. In chapter III the Goodness-of-fit statistic of Genest, Quessy and Remillard (2006) will be used to compare the performance of copulas in modeling tail dependence.
Chapter 3: Tail Dependence among Weather Indices: The Case of Iowa

County-Level Rainfalls

3.1 Introduction

In recent years, a number of papers have appeared in the finance literature that ask whether asset returns in spatially-separated markets are more highly correlated during a financial crisis, a phenomenon referred to as “spatial contagion” (Durante and Jaworski 2009; Bradley and Taqqu 2004; Bradley and Taqqu 2005a; Bradley and Taqqu 2005b). If stock returns are more highly correlated during financial crises than in normal times, stock portfolios will be riskier than predicted by conventional asset pricing models. The questions being addressed by financial analysts are analogous to those that must be addressed in index insurance design: in both cases, one is concerned with the degree of dependence exhibited by two or more random variables at the extremes of their distribution.

Tail dependence is an important issue in the design of index insurance products for three reasons. First, suppose an insurer offers a range of index insurance contracts written on a weather variable, say, rainfall at different locations within a defined geographical area. One of the key questions that must be addressed by the
insurer is where the rainfall stations on which insurance is offered should be located and, in particular, how many and how far apart they should be. The stations must be sufficient in number and geographical density to keep basis risk at an acceptable level, that is, it should be possible for a potential insured to purchase an index contract for a rainfall station sufficiently near to him that the rainfall at that station closely tracks his losses. However, the insurer also has an interest in keeping the number of stations to a minimum in order to reduce the costs of administering the insurance program. The trade-off between basis risk and administration cost can only be adequately assessed if there is a clear understanding of how rainfalls at different locations in the target geographical area are related to losses and to one another, particularly at the tails.

Second, the insurer will be interested in assessing the distribution of payouts of his entire portfolio of index insurance contracts in order to calculate the maximum probable loss associated with his entire book of business. If the underlying weather variables exhibit tail dependence, then standard portfolio risk assessments based explicitly or implicitly on normal distribution theory could result in serious underestimates of the riskiness of the portfolio, leaving the insurer exposed to greater business risk than he realizes. Faithfully capturing the tail dependence in these instances is essential for analyzing the performance of the book of business.

Third, an important task in index insurance design is to compute the expected indemnity associated with a given indemnity schedule. Indemnities, however, are
paid only when the index falls below a certain threshold, an event that occurs only infrequently. As such, the data available to support the calculation of this critical statistic is usually very limited. One way to address the paucity of data is to estimate the expected indemnities of multiple contracts jointly. This should lead to gains in efficiency that will depend primarily on the degree of dependence exhibited at the critical extremes of the underlying index distributions. In other words, in the presence of tail dependence, it may be possible to achieve substantial gains in efficiency by jointly estimating the expected indemnities of various contracts, provided the tail dependencies are faithfully captured.

In this chapter, I ask whether spatially separated weather variables commonly used in index insurance design, such as rainfall at different weather stations within a defined geographical area, are “more highly correlated at the tails”. As a case study, I will assess the degree of dependence at the tail of distribution exhibited by Iowa June county-level rainfalls, employing data for all 99 Iowa counties from 1954-2008 obtained from National Climatic Data Center (NCDC).

I employ two methods including a test based on Kendall’s tau, and estimation of bivariate copula functions. The first method compares the Kendall’s correlation coefficients computed for subsamples of lower and upper tail observations. It is used to preliminarily test whether two proposed indices are more highly correlated at the lower tail of their distributions than at the upper tail of their distributions. The second method tests for tail dependence by estimating alternate bivariate copulas and
comparing how well they fit the data. Specifically, I search among various candidate bivariate copulas with different tail dependence characteristics and, using goodness-of-fit tests, attempt to identify the copula structures that best explain the nature of dependence among rainfalls in adjacent counties.

3.2 Kendall’s Tau Test

Kendall’s tau is a measure of rank correlation first defined by Kendall (1938), which describes the difference between the frequency that a pair of observations is in the same order on both variables and the frequency that they are in the opposite order. Kendall’s tau is often used instead of linear correlation coefficient in measuring association when there is a monotonic but not necessarily linear relationship between two random variables. As a nonparametric statistic, Kendall’s tau can be used to make inferences about the degree of relation between variables without making assumption about the distribution underlying the samples that are observed (Cliff and Charlin, 1991). Therefore, Kendall’s tau is more reliable than linear correlation coefficient when the underlying distributions of the variables and the pattern of dependence between them are unknown.

Suppose I have observations on \( p \) random variables \( x_1, x_2, \ldots, x_p \). According to Kendall (1970), Kendall’s tau between variable \( x_i \) and \( x_j \) is formulated as
\[ \tau_{ij} = \Pr(x_{k,i} > x_{k,j}, x_{k,j} > x_{l,j}) - \Pr(x_{k,i} > x_{l,i}, x_{k,j} < x_{l,j}) \]  

(1)

for observations \( k, l \). The sample estimate of Kendall’s tau is

\[ \bar{\tau}_{ij} = \sum \sum t_{kl,ij} / n(n-1) , \]

(2)

where, \( t_{kl,ij} = \text{sign}(x_{k,i} - x_{l,i}) \times \text{sign}(x_{k,j} - x_{l,j}) \), which is 1 if \( k \) and \( l \) are in the same order on both variables, -1 if they are opposite, and 0 if they are tied on either. The general formula and unbiased estimate for the variance of Kendall’s tau is addressed by Cliff and Charlin (1991).

In large samples, the sample tau has asymptotically normal distribution and a complicated standard deviation (Cliff and Charlin, 1991). Thus, whether the two random variables are more highly correlated at the lower tails of their distributions than at the upper tail of their distributions may be tested by comparing Kendall’s taus computed for subsamples of lower and upper tail observations:

- Sort the paired observations \((x_{1t}, x_{2t})\) according to the values \(x_{1t}\).
- Delete the \(n/4\) middle observations, partitioning the observations into two sets of equal size, one containing the lower ranked observations and one containing the higher ranked observations.
- Compute the Kendall taus for each subset of observations, \( \tau_L \) and \( \tau_U \), and set \( U = \tau_U - \tau_L \).

For large samples, \( U \) also has asymptotically normal distribution with mean zero and variance calculated by the Cliff and Charlin, which allows us to test the hypothesis \( H_0: \tau_U = \tau_L \) or \( H_0: U = 0 \) against the alternative \( H_A: \tau_U \neq \tau_L \) or
\( H_A : U \neq 0 \) using a two-sided Student-t test.

Before conducting the test, a bootstrap method is applied to simulate the large sample of \( \tau \) with independent observations. First, a random sample is drawn from the original rainfall data of two counties with replacement. After the drawn sample is sorted through the 55 years by the average rainfall of the state, the first 21 years of data form the drought subsample, and the last 21 years of data form the normal/flood subsample. Second, I calculate the Kendall’s \( \tau \) between the pair of counties for drought subsample and normal/flood subsample, respectively. Third, I repeat this procedure 1000 times, and then obtain two samples of \( \tau \)s for the two counties, the drought \( \tau \) and the normal/flood \( \tau \). Fourth, a 90% confidence interval of \( \tau \) difference is created by kernel smoothing the \( \tau \)s and finding the lower critical value at the 5% quantile and higher critical value at the 95% quantile. If zero is within the confidence interval, then the upper tail dependence is not significantly different from the lower tail dependence. If zero is smaller than the lower critical value, then the upper tail dependence is significantly greater than the lower tail dependence. If zero is larger than the higher critical value, then the lower tail dependence is greater than the upper tail dependence. Repeat the test for each pair of the 99 counties.
<table>
<thead>
<tr>
<th>County</th>
<th>Dr</th>
<th>Fl</th>
<th>County</th>
<th>Dr</th>
<th>Fl</th>
<th>County</th>
<th>Dr</th>
<th>Fl</th>
<th>County</th>
<th>Dr</th>
<th>Fl</th>
</tr>
</thead>
<tbody>
<tr>
<td>Adair</td>
<td>0</td>
<td>1</td>
<td>Clay</td>
<td>0</td>
<td>11</td>
<td>Hancock</td>
<td>0</td>
<td>17</td>
<td>Madison</td>
<td>2</td>
<td>0</td>
</tr>
<tr>
<td>Adams</td>
<td>0</td>
<td>3</td>
<td>Clayton</td>
<td>2</td>
<td>1</td>
<td>Hardin</td>
<td>0</td>
<td>4</td>
<td>Mahaska</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>Allamakee</td>
<td>0</td>
<td>4</td>
<td>Clinton</td>
<td>0</td>
<td>2</td>
<td>Harrison</td>
<td>0</td>
<td>5</td>
<td>Marion</td>
<td>1</td>
<td>3</td>
</tr>
<tr>
<td>Appanoose</td>
<td>0</td>
<td>1</td>
<td>Crawford</td>
<td>0</td>
<td>5</td>
<td>Henry</td>
<td>1</td>
<td>2</td>
<td>Marshall</td>
<td>1</td>
<td>4</td>
</tr>
<tr>
<td>Audubon</td>
<td>0</td>
<td>1</td>
<td>Dallas</td>
<td>0</td>
<td>1</td>
<td>Howard</td>
<td>0</td>
<td>10</td>
<td>Mills</td>
<td>0</td>
<td>13</td>
</tr>
<tr>
<td>Benton</td>
<td>1</td>
<td>0</td>
<td>Davis</td>
<td>1</td>
<td>1</td>
<td>Humboldt</td>
<td>0</td>
<td>19</td>
<td>Mitchell</td>
<td>0</td>
<td>6</td>
</tr>
<tr>
<td>Black Hawk</td>
<td>0</td>
<td>5</td>
<td>Decatur</td>
<td>0</td>
<td>1</td>
<td>Ida</td>
<td>0</td>
<td>1</td>
<td>Monona</td>
<td>0</td>
<td>9</td>
</tr>
<tr>
<td>Boone</td>
<td>5</td>
<td>4</td>
<td>Delaware</td>
<td>1</td>
<td>1</td>
<td>Iowa</td>
<td>1</td>
<td>2</td>
<td>Monroe</td>
<td>0</td>
<td>2</td>
</tr>
<tr>
<td>Bremer</td>
<td>0</td>
<td>10</td>
<td>Des Moines</td>
<td>0</td>
<td>1</td>
<td>Jackson</td>
<td>0</td>
<td>2</td>
<td>Montgomery</td>
<td>0</td>
<td>5</td>
</tr>
<tr>
<td>Buchanan</td>
<td>0</td>
<td>4</td>
<td>Dickinson</td>
<td>0</td>
<td>2</td>
<td>Jasper</td>
<td>0</td>
<td>6</td>
<td>Muscatine</td>
<td>1</td>
<td>5</td>
</tr>
<tr>
<td>Buena Vista</td>
<td>0</td>
<td>11</td>
<td>Dubuque</td>
<td>1</td>
<td>5</td>
<td>Jefferson</td>
<td>0</td>
<td>0</td>
<td>Obrien</td>
<td>0</td>
<td>6</td>
</tr>
<tr>
<td>Butler</td>
<td>0</td>
<td>11</td>
<td>Emmet</td>
<td>1</td>
<td>6</td>
<td>Johnson</td>
<td>0</td>
<td>8</td>
<td>Osceola</td>
<td>0</td>
<td>14</td>
</tr>
<tr>
<td>Calhoun</td>
<td>0</td>
<td>5</td>
<td>Fayette</td>
<td>0</td>
<td>2</td>
<td>Jones</td>
<td>0</td>
<td>5</td>
<td>Page</td>
<td>0</td>
<td>5</td>
</tr>
<tr>
<td>Carroll</td>
<td>0</td>
<td>1</td>
<td>Floyd</td>
<td>0</td>
<td>3</td>
<td>Keokuk</td>
<td>0</td>
<td>3</td>
<td>Palo Alto</td>
<td>0</td>
<td>5</td>
</tr>
<tr>
<td>Cass</td>
<td>0</td>
<td>2</td>
<td>Franklin</td>
<td>0</td>
<td>12</td>
<td>Kossuth</td>
<td>0</td>
<td>8</td>
<td>Plymouth</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Cedar</td>
<td>0</td>
<td>2</td>
<td>Fremont</td>
<td>0</td>
<td>13</td>
<td>Lee</td>
<td>0</td>
<td>0</td>
<td>Pocahontas</td>
<td>0</td>
<td>6</td>
</tr>
<tr>
<td>Cerro Gordo</td>
<td>0</td>
<td>8</td>
<td>Greene</td>
<td>3</td>
<td>0</td>
<td>Linn</td>
<td>0</td>
<td>6</td>
<td>Polk</td>
<td>2</td>
<td>3</td>
</tr>
<tr>
<td>Cherokee</td>
<td>1</td>
<td>4</td>
<td>Grundy</td>
<td>0</td>
<td>5</td>
<td>Louisa</td>
<td>0</td>
<td>3</td>
<td>Pottawattamie</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Chickasaw</td>
<td>0</td>
<td>1</td>
<td>Guthrie</td>
<td>0</td>
<td>2</td>
<td>Lucas</td>
<td>2</td>
<td>1</td>
<td>Poweshiek</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>Clarke</td>
<td>0</td>
<td>1</td>
<td>Hamilton</td>
<td>0</td>
<td>3</td>
<td>Lyon</td>
<td>0</td>
<td>8</td>
<td>Ringgold</td>
<td>1</td>
<td>2</td>
</tr>
</tbody>
</table>

Note: Dr means that the number of taus significantly higher in drought years. Fl means that the number of taus significantly higher in normal/flood years.

Table 3.1 Number of taus with higher values in drought and flood years. Each county with all the other 98 counties
The test results, as listed in Table 3.1, show that the upper tail dependence is more significant than lower tail dependence. In all of the 99 counties, 86 counties have more taus with significantly higher values in normal/flood years than in drought years, while only 5 counties have more taus with significantly higher values in drought years than in normal/flood years. The 5 counties are Boone, Clayton, Greene, Madison and Mahaska. Davis County and Delaware County have the same number of taus with higher values in drought years and normal/flood years. Jefferson, Lee, Plymouth, Pottawattamie, Tama and Wayne County do not show any tail dependence. The rejection of null hypothesis of no tail dependence for most pairs of counties implies that tail dependence exists in the association of rainfall, while not all the associations have the same pattern: some of the associations tend to be stronger in drought years and others tend to be higher in normal/flood years.

Another similar test is conducted to investigate the association between a county and its adjacent counties. In all the 99 Iowa counties, 63 counties have all their adjacent counties inside the state and there are 36 counties having some of their adjacent counties outside the state. For purposes of simplicity, I only test the 63 counties having all their adjacent counties inside the state. As is shown in Table 3.2, 9 of the 63 counties have more taus with higher values in drought years, two counties have the same number of adjacent counties that have lower tail dependence and upper tail dependence, three counties do not have adjacent tail dependence, and 49 of the 63 counties have more taus with higher values in normal/flood years.
Furthermore, 44 of the 63 counties have no adjacent taus higher in drought years, and 9 of the 63 counties have no adjacent taus higher in normal/flood years. This result implies that in the relationship of a county and its adjacent counties, tail dependence is more significant than when all the associations between a county and the other counties in Iowa are considered. Similar to Table 3.1, Table 3.2 shows different patterns of tail dependence. For some counties, the upper tail dependence dominates, while the lower tail dependence dominates in some other counties.
<table>
<thead>
<tr>
<th>County</th>
<th>Ad</th>
<th>Dr</th>
<th>Fl</th>
<th>County</th>
<th>Ad</th>
<th>Dr</th>
<th>Fl</th>
<th>County</th>
<th>Ad</th>
<th>Dr</th>
<th>Fl</th>
</tr>
</thead>
<tbody>
<tr>
<td>Adair</td>
<td>7</td>
<td>0</td>
<td>1</td>
<td>Delaware</td>
<td>6</td>
<td>1</td>
<td>1</td>
<td>Madison</td>
<td>7</td>
<td>2</td>
<td>0</td>
</tr>
<tr>
<td>Adams</td>
<td>7</td>
<td>0</td>
<td>3</td>
<td>Fayette</td>
<td>9</td>
<td>0</td>
<td>2</td>
<td>Mahaska</td>
<td>6</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>Audubon</td>
<td>6</td>
<td>0</td>
<td>1</td>
<td>Floyd</td>
<td>8</td>
<td>0</td>
<td>3</td>
<td>Marion</td>
<td>6</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Benton</td>
<td>7</td>
<td>1</td>
<td>0</td>
<td>Franklin</td>
<td>8</td>
<td>0</td>
<td>6</td>
<td>Marshall</td>
<td>6</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>Black Hawk</td>
<td>7</td>
<td>0</td>
<td>5</td>
<td>Greene</td>
<td>6</td>
<td>3</td>
<td>0</td>
<td>Monroe</td>
<td>7</td>
<td>0</td>
<td>2</td>
</tr>
<tr>
<td>Boone</td>
<td>6</td>
<td>5</td>
<td>1</td>
<td>Grundy</td>
<td>7</td>
<td>0</td>
<td>5</td>
<td>Montgomery</td>
<td>7</td>
<td>0</td>
<td>5</td>
</tr>
<tr>
<td>Bremer</td>
<td>6</td>
<td>0</td>
<td>5</td>
<td>Guthrie</td>
<td>7</td>
<td>0</td>
<td>2</td>
<td>Obrien</td>
<td>8</td>
<td>0</td>
<td>6</td>
</tr>
<tr>
<td>Buchanan</td>
<td>7</td>
<td>0</td>
<td>4</td>
<td>Hamilton</td>
<td>7</td>
<td>0</td>
<td>3</td>
<td>Palo Alto</td>
<td>7</td>
<td>0</td>
<td>5</td>
</tr>
<tr>
<td>Buena Vista</td>
<td>8</td>
<td>0</td>
<td>5</td>
<td>Hancock</td>
<td>7</td>
<td>0</td>
<td>7</td>
<td>Pocahontas</td>
<td>8</td>
<td>0</td>
<td>6</td>
</tr>
<tr>
<td>Butler</td>
<td>8</td>
<td>0</td>
<td>5</td>
<td>Hardin</td>
<td>7</td>
<td>0</td>
<td>3</td>
<td>Polk</td>
<td>7</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>Calhoun</td>
<td>6</td>
<td>0</td>
<td>5</td>
<td>Henry</td>
<td>6</td>
<td>1</td>
<td>2</td>
<td>Poweshiek</td>
<td>7</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>Carroll</td>
<td>7</td>
<td>0</td>
<td>1</td>
<td>Humboldt</td>
<td>7</td>
<td>0</td>
<td>6</td>
<td>Sac</td>
<td>7</td>
<td>0</td>
<td>4</td>
</tr>
<tr>
<td>Cass</td>
<td>7</td>
<td>0</td>
<td>2</td>
<td>Ida</td>
<td>6</td>
<td>0</td>
<td>1</td>
<td>Shelby</td>
<td>6</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>Cedar</td>
<td>6</td>
<td>0</td>
<td>2</td>
<td>Iowa</td>
<td>7</td>
<td>1</td>
<td>2</td>
<td>Story</td>
<td>6</td>
<td>0</td>
<td>2</td>
</tr>
<tr>
<td>Cerro Gordo</td>
<td>8</td>
<td>0</td>
<td>6</td>
<td>Jasper</td>
<td>8</td>
<td>0</td>
<td>0</td>
<td>Tama</td>
<td>7</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Cherokee</td>
<td>8</td>
<td>1</td>
<td>4</td>
<td>Jefferson</td>
<td>6</td>
<td>0</td>
<td>0</td>
<td>Union</td>
<td>7</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>Chickasaw</td>
<td>7</td>
<td>0</td>
<td>1</td>
<td>Johnson</td>
<td>7</td>
<td>0</td>
<td>4</td>
<td>Wapello</td>
<td>7</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>Clarke</td>
<td>7</td>
<td>0</td>
<td>1</td>
<td>Jones</td>
<td>6</td>
<td>0</td>
<td>5</td>
<td>Warren</td>
<td>7</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>Clay</td>
<td>8</td>
<td>0</td>
<td>7</td>
<td>Keokuk</td>
<td>6</td>
<td>0</td>
<td>3</td>
<td>Washington</td>
<td>6</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>Crawford</td>
<td>8</td>
<td>2</td>
<td>3</td>
<td>Linn</td>
<td>7</td>
<td>2</td>
<td>5</td>
<td>Webster</td>
<td>7</td>
<td>2</td>
<td>5</td>
</tr>
<tr>
<td>Dallas</td>
<td>7</td>
<td>5</td>
<td>1</td>
<td>Lucas</td>
<td>7</td>
<td>2</td>
<td>3</td>
<td>Wright</td>
<td>8</td>
<td>0</td>
<td>7</td>
</tr>
</tbody>
</table>

Note: Ad is the number of adjacent counties. Dr is the number of taus significantly higher in drought years. Fl is the number of taus significantly higher in normal/flood years.

Table 3.2  Number of adjacent taus with higher values in drought and flood years

The Kendall’s tau test, as a preliminary test, demonstrates that Kendall’s correlation coefficients are different between lower tail and upper tail in most pairs of counties in Iowa, which is evidence for tail dependence. However, there are still some problems with this procedure. First, the comparison of Kendall’s taus does not give a specific measure of tail dependence, which is an important factor for contract design. Second, the test only deals with paired data and does not consider
multivariate dependence among the county rainfalls. Index insurance designers are more interested in the dependence structure among multiple indices. To solve the first problem, a copula method is used in order to model the dependence structure at tails and to measure the tail dependence. After that, a spatial structural model is used in the next chapter to reveal the variation of the spatial autoregressive lag at tails with multivariate variables.

3.3 Copulas and Goodness-of-fit Statistics

I compared how well five distinct copulas (Gaussian, Student-t, Frank, Clayton and Gumbel) fit pairwise Iowa county rainfall data. The two-step MLE is used to estimate copula parameters. Based on a visual inspection of the histograms of observed rainfalls, the rainfalls in each county are assumed to have log-normal distribution. After estimating the parameters of the log-normal marginals for each county, I use maximum likelihood to estimate the copula parameter for each of the five copulas and for each pair of adjacent counties using the fitted marginal distributions.

To compare the performance of these copulas, the Goodness-of-fit statistic for copulas developed by Genest, Quesy and Remillard (2006) is calculated for each fitted copula function and for each pair of adjacent counties. Suppose $F(x_1, x_2)$ is the joint distribution based on specific copula function as is shown in (1). Let
$K(\theta, t) = P\{F(x_1, x_2) \leq t\}$ with the copula parameter $\theta$. The empirical version of $K(\theta, t)$ is defined as

$$K_n(t) = \frac{1}{n} \sum_{j=1}^{n} 1(V_j \leq t), \quad t \in [0,1],$$

(10)

where $n$ represents the size of sample, $V_j$ are pseudo-observations defined by

$$V_j = \frac{1}{n} \sum_{k=1}^{n} 1(X_{1k} \leq X_{1j}, X_{2k} \leq X_{2j}),$$

and $1(V_j \leq t)$ refers to the indicator function that has the value of 1 when $V_j \leq t$ and the value of 0 when $V_j > t$. The Goodness-of-fit statistic for copulas is given by

$$S = \int_0^1 |\gamma(t)|^2 k(\theta, t)dt,$$

(11)

where $\gamma(t) = \sqrt{n}[K_n(t) - K(\theta, t)]$ and $k(\theta, t)$ represents the density function of $K(\theta, t)$.

In order to compute the Goodness-of-fit statistic based on the empirical process $\gamma$, I generate a large number of independent samples of size $n$ from the fitted copulas, and compute the corresponding values of the statistic $S$ for each copula and for each pair of adjacent counties. In this paper, I use Gaussian kernel density function to fit the empirical distribution. Specifically, the bootstrap procedure involves three steps. First, I fit a bivariate kernel density of the observations and calculate the cdf’s of bivariate kernel function at $N \times N$ grids in the area $[0,1]^2$. Here, I use $N = 50$. Second, I generate 1000 random samples of size $n = 55$, from the fitted copula function $\hat{C}$ with the estimated parameter $\hat{\theta}$. For each of these samples, I fit a bivariate kernel function and obtain the cdf’s of
bivariate kernel at the same grids as in step one. Third, for each of the 1000 samples, a Goodness-of-fit statistic for copulas, $S$, is computed based on the cdf’s in the first step and the second step and the kernel density function in the second step. I repeat the procedure for each of the five copulas. The means of the $S$ statistic in the 1000 samples generated from the five copulas are used to compare the performance of the five copulas.

<table>
<thead>
<tr>
<th>Rankings</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gaussian copula</td>
<td>13%</td>
<td>15%</td>
<td>23%</td>
<td>25%</td>
<td>24%</td>
</tr>
<tr>
<td>Student-t copula</td>
<td>16%</td>
<td>18%</td>
<td>28%</td>
<td>23%</td>
<td>15%</td>
</tr>
<tr>
<td>Frank copula</td>
<td>10%</td>
<td>13%</td>
<td>9%</td>
<td>29%</td>
<td>39%</td>
</tr>
<tr>
<td>Clayton copula</td>
<td>43%</td>
<td>33%</td>
<td>11%</td>
<td>6%</td>
<td>7%</td>
</tr>
<tr>
<td>Gumbel copula</td>
<td>18%</td>
<td>21%</td>
<td>29%</td>
<td>17%</td>
<td>15%</td>
</tr>
</tbody>
</table>

Table 3.3  Percentage of copulas’ rankings in adjacent counties

The comparison of the five copulas is conducted for each of the 297 adjacent pairs of counties using the Goodness-of-fit statistic for copulas. The performance of copulas is evaluated by the rankings of Goodness-of-fit statistic for each pair of counties. Table 3.3 shows the percentage of rankings for each of the five copulas. In all the 297 pairs of adjacent rainfalls, 43%, or 128 pairs, are best fitted by the
Clayton copula, 18%, or 53 pairs, are best fitted by the Gumbel copula, 16%, or 48 pairs, are best fitted by the Student-t copula, 13%, or 39 pairs, are fitted best by the Gaussian copula, and only 10%, or 29 pairs, are best fitted by the Frank copula. Considering the second best fit, 33% of the 297 pairs select the Clayton copula, and 39% of the pairs select Gumbel copula. When it comes to the worst fit with respect to the Goodness-of-fit statistic, 39% of the pairs list the Frank copula as the worst fit, and 29% of the pairs list it as the second worst fit. It is obvious that the Clayton copula performs best in fitting the rainfall data of adjacent counties, the Gumbel copula is the second best, and the Frank copula performs worst. Gaussian copula and Student-t copula perform better than the Frank copula but worse than the Clayton copula and the Gumbel copula.

By looking at Table 2.4, the Clayton copula is characterized by strong lower-tail dependence. The good performance of the Clayton copula, therefore, implies that for many pairs of adjacent counties, rainfalls are more strongly related when precipitation is abnormally low, which is strong evidence that lower tail dependence exists in many adjacent counties. For some pairs of adjacent counties, the Gumbel copula provides a better fit, suggesting that upper tail dependence also exists. It is possible that some pairs of adjacent rainfalls have both lower tail dependence and upper tail dependence. The Gaussian copula and the Student-t copula can also capture some degree of tail dependence. However, since they are symmetric copulas, they tend to underestimate lower tail dependence when, as is in
the case of Iowa rainfall, the correlation of rainfall in adjacent counties rises asymmetrically in drought years, but may only slightly rises in years of high precipitation.

The tail dependence parameters for each pair of counties can be computed by the functions shown in Table 2.4 using the estimated copula parameters. Table 3.4 reports the mean and standard deviation of estimated copula parameters for all the 297 pairs of adjacent counties and the related tail dependence. The estimated parameter of the Clayton copula has relatively highest variation across the adjacent counties. The average lower tail dependence of the 297 pairs of adjacent counties is 0.46 and the average upper tail dependence is 0.62, with standard deviation 0.17 and 0.06, respectively. The upper tail dependence tends to be more stable than lower tail dependence across all the pairs of adjacent counties. Therefore, a contract design that focuses on the correlation between the indemnity and losses caused by drought may require further investigation in the lower tail dependence among adjacent counties.
Table 3.4  Summary of estimates of parameters for three Archimedean copulas and
the average tail dependence of adjacent counties

<table>
<thead>
<tr>
<th>Copulas</th>
<th>Mean of $\hat{\theta}$</th>
<th>Std. of $\hat{\theta}$</th>
<th>Mean of $\lambda_\kappa$</th>
<th>Std. of $\lambda_\kappa$</th>
<th>Mean of $\lambda_\lambda$</th>
<th>Std. of $\lambda_\lambda$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Frank</td>
<td>6.69</td>
<td>1.64</td>
<td>0</td>
<td>-</td>
<td>0</td>
<td>-</td>
</tr>
<tr>
<td>Clayton</td>
<td>1.01</td>
<td>0.44</td>
<td>0.46</td>
<td>0.17</td>
<td>0</td>
<td>-</td>
</tr>
<tr>
<td>Gumbel</td>
<td>2.20</td>
<td>0.35</td>
<td>0</td>
<td>-</td>
<td>0.62</td>
<td>0.06</td>
</tr>
</tbody>
</table>

3.4 Conclusion

The existence of tail dependence between spatially separated agricultural indices such as rainfall is important for insurers who are interested in assessing the maximum probable loss associated with his portfolio, and who must estimate expected indemnities using limited extreme value data. The Pearson linear correlation coefficient, which is commonly used in measuring dependence, is generally inadequate for the task because it cannot describe nonlinear association and cannot distinguish between lower-tail and global dependence.

In order to test for and measure tail dependence among county-level June Iowa rainfalls, I estimated a variety of copula functions, including Archimedean copulas (Clayton, Gumbel, and Frank) and elliptical copulas (Gaussian and Student-t) for adjacent county pairs. The performance of the five copulas was assessed by comparing by the Goodness-of-fit statistic based on a nonparametric bootstrap
procedure. My results indicate that the Clayton copula fits the data best, which implies that lower tail dependence exists in most of adjacent county-level rainfalls in Iowa. The results suggest that accounting for tail dependence in the contexts where extreme events could substantially enhance the accuracy of loss assessment for agricultural index insurance portfolios. The results also suggest that patterns of tail dependence differ across counties. Some of the adjacent counties tend to have higher correlation when drought occurs, while some tend to have higher correlation in normal or abnormally wet years.
Chapter 4: Designing an Optimal Rainfall Index Insurance Contract:

The Case of Henan, China

4.1 Introduction

China is one of the largest agricultural producers in the world and one of the countries most vulnerable to serious natural disasters. Since 1949, average losses due to natural disasters have been over $1.25 billion\(^2\) per year, with the average affected crop growing area of over 40 million hectares and the average affected population of up to 200 million per year.\(^3\) The most frequent natural disasters affecting Chinese farmers are flood, drought, earthquake and typhoon. These events can cause serious damages that will lead affected farmers to fall deeper into poverty if not addressed by governmental authorities. Governmental disaster assistance suffers from high costs and can strain the national budget, given the systemic nature of catastrophic weather events. Moreover, government disaster assistance cannot always reach farmers quickly because of the need to assess damages. Needed is some form of insurance that can indemnify farmers quickly in case of a disaster at a relatively low cost.

\(^2\) I used the exchange rate in 2006: $1 = ¥8.

\(^3\) http://www.lxxf.gov.cn/xwzx/ShowArticle.asp?ArticleID=422
The insurance sector has been growing rapidly in China over the past few decades. Total premiums have increased from $57 million in 1980 to $70,517 million in 2006 and $87,923 million in 2007.\(^4\) However, the growth of agricultural insurance premiums has not kept pace with the growth of non-agricultural insurance. The relatively slow growth of agricultural insurance premiums can be explained by the prevalence of more severe moral hazard and adverse selection problems compared with other insurance products. The more severe asymmetric information problems necessitate the government subsidies when introducing traditional insurance.

Regular agricultural insurance service did not exist in China until 1982 when the People’s Insurance Company (Group) of China (PICC), the state-owned monopoly insurance company in China, began to offer agricultural insurance service. In 2006, only PICC and China Insurance had nationwide agricultural insurance service, and the agricultural insurance premium is $100 million, which is no more than 0.14% of total premium in the whole insurance industry, compared with $5.4 million and 1.68% in 1985.\(^5\) After realizing the importance of agricultural insurance, the Chinese government began to promote its development. In 2007 total agricultural insurance premiums jumped to $658.9 million, of which $268.8 million was paid by the central government, over $250 million was paid by local governments, and only about $137.5 million was paid by farmers.\(^6\) Table 4.1 shows the agricultural insurance policies sold

\(^6\) [http://www.agri.gov.cn/jpjs/t20080725_1091264.htm](http://www.agri.gov.cn/jpjs/t20080725_1091264.htm)
in China.

<table>
<thead>
<tr>
<th>Product(s)</th>
<th>Liability</th>
<th>Region(s)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Crop:</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Corn, Rice, Wheat, Soy Bean,</td>
<td>Multi-peril</td>
<td>National</td>
</tr>
<tr>
<td>Cotton, etc.</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Livestock:</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Pig, Dairy Cow, Chicken, etc.</td>
<td>Multi-peril</td>
<td>National</td>
</tr>
<tr>
<td><strong>Others:</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Watermelon, Rapes, Tobacco, etc.</td>
<td>Multi-peril</td>
<td>Some growing areas</td>
</tr>
<tr>
<td>Drought Index Insurance for Drought</td>
<td>Multi-peril</td>
<td>Small-scale pilot in Shanghai</td>
</tr>
<tr>
<td>Vegetables</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 4.1 Agricultural Insurance Sold in China in 2007

The shrinking share of agricultural insurance in the Chinese insurance market is mainly attributed to the much lower profitability of agricultural insurance compared with other insurance products and the reform of state-owned insurance companies. Crop yield is highly correlated with labor input and climate condition. Insured
farmers have an incentive to input less labor and cultivate the land carelessly, which is called ex ante moral hazard. After harvest, it is possible for farmers to conceal real yield and claim losses. This is called ex post moral hazard. Adverse selection refers to the phenomenon that farmers with high risk are more eager to buy insurance than those with low risk. Moral hazard and adverse selection lead to high costs and thus high premiums. As a result of higher costs, farmers are not willing to purchase agricultural insurance at market rates.

The asymmetric information problems that restrict the development of agricultural insurance in much of the world is more serious in developing countries such as China. Because of high density of population and limited cultivated land, the per capita arable land in China was only 0.09 hectare in 2006, which is less than one half of the per capita arable land area in the world. This leads to that has been referred to as the “small units” problem (Hazell and Skees, 2005). Small units lead to higher operation costs associated with screening and monitoring. On the other hand, the total claim and payment of agricultural insurance ran up to $75 million in 2005 and $132 million in 2006, respectively, equivalent to 85.7% and 75% of total premium. High risk and high costs render insurance companies unprofitable. The PICC used to be a state-owned monopoly that relied on government credit and government finance. However, with insurance industry commercialization, the number of insurance companies in China grew to 110 by the end of 2007. Facing competition, the PICC has to focus more on profitability, and as a result has reduced
its agricultural insurance portfolio.

To provide risk management opportunities for poor farmers, researchers and development organizations have been exploring weather index insurance. Weather index insurance pays indemnities based not on actual losses experienced by the policyholder but rather on realizations of a weather index that is highly correlated with actual losses (Barnett and Mahul, 2007). Generally, a weather index measures a specific weather variable, say rainfall or temperature. A threshold is specified to establish the range of values over or less than which indemnity will be paid.

Weather index insurance has several advantages relative to traditional insurance products. First, individual farmer cannot affect the underlying index so that moral hazard problem does not exist. Second, farmers whose individual yields are more correlated with underlying index, not those whose risk is higher, have more incentives to buy index insurance. There is little possibility that individual farmers have better information than the insurer about the underlying index, and thus little potential for adverse selection exists. Third, indemnities paid are based solely on the value of underlying index, which is usually provided by a public institution, so the insurer does not need to employ agricultural experts to estimate the actual loss house by house. In addition, there is no need to classify individual policyholders according to their risk exposure. Therefore, operation costs with index insurance are much lower than traditional insurance.

Weather index insurance has been introduced as pilot projects in various
developing countries. Target users range from micro-level clients such as individual farmers to macro-level institutional clients such as financial institutions and macro-level institutions such as national governments. Among available weather indices, rainfall is the most commonly used because of its high correlation with yields. However, there are also insurance products based on other indices such as river water levels and regional livestock mortality. One of the first index insurance pilot programs was introduced in Andhra Pradesh and Uttar Pradesh, India in 2003. As of 2007, more than 10,000 clients and five insurers were involved. In Ethiopia, the World Food Program took out a ‘livelihood loss’ index on behalf of government in 2006. In 2008 this ‘livelihood loss’ index insurance covered up to 6.7 million people. Malawi introduced rainfall index insurance in 2005 to protect farmers against drought.

One of the problems with designing index insurance contracts is accounting for the dependence structure between the index and yield risks. The modeling approach can utilize the procedures of generating pairs of random variables from a given pair of marginal distributions with a known degree of correlation (Zhu, Ghosh and Goodwin, 2008). In terms of correlation, Spearman’s rank correlation coefficient and Kendall’s tau coefficient are commonly used. However, Spearman’s rank correlation coefficient and Kendall’s tau coefficient are both global measure. To model the correlation at the tails of the distribution, a joint probability distribution is a better choice, since

---

association between any set of random variables at any place of their distributions can be fully described by knowing the multivariate joint distribution. Deng, Barnett and Vedenov (2007) claimed that a multivariate empirical distribution is probably the closest to preserving the information contained in the data without imposing distributional assumptions. However, applications of multivariate empirical distributions are limited by the paucity of data and the discontinuity of density functions. Vedenov (2008) argued that copulas are an alternative way to model joint distributions of random variables with greater flexibility both in terms of marginal distributions and the dependence structure.

4.2 Data

In this paper I design a drought insurance program for Henan Province, China. Henan province is the main winter wheat growing area of China. Wheat production in Henan Province is equal to 1/4 of total wheat production in China. Wheat yield data obtained from China Farming Information System and National Bureau of Statistics of China\(^8\) provide 56 annual average wheat yield values (10 tons/hectare) observed in Henan Province covering the period 1951-2006.\(^9\) Rainfall data obtained from the China Meteorological Data Sharing Service System provide monthly rainfall data (millimeter) observed in meteorological stations in Henan Provinces from 1951

\(^8\) http://www.stats.gov.cn/
to 2006.\textsuperscript{10}

Figure 4.1 illustrates historical Henan Province wheat yields and estimated quadratic trend

\[ Y_t = 254.6 - 0.3t + 6.81t^2 + \mu_t, \]

where \( Y_t \) is wheat yield at time \( t \), and \( \mu_t \) represents detrended residual. Time series analysis indicated that the detrended residuals are serially uncorrelated, stationary and close to normally distributed. The analysis also revealed no obvious structural changes. Figure 4.2 presents the empirical histogram of the detrended yields, the fitted normal kernel distribution (red), and the fitted parametric normal distribution (blue). The kernel and normal distributions are very close to each other, supporting the assumption that detrended Hunan wheat yields are normally distributed. Thus, I will use a normal distribution to model the marginal distribution of yield.
Figure 4.1  Wheat yields and trend in Henan

Figure 4.2  Histogram of detrended yield in Henan, 1951-2006
Figure 4.3 presents the empirical histogram of March rainfall at Henan Province station 53898. The autocorrelation function and partial autocorrelation function indicated no apparent autocorrelation in ACF or PACF. A unit root test confirmed that the rainfall time series are all stationary. A Box-Pierce test of dependency of rainfall measured at three distinct stations generated p-values of 0.9484. I could not, therefore, reject the null hypothesis that the time series are independent.

![Histogram of March rainfall in Henan, 1951-2006.](image)

Figure 4.3 Histogram of March rainfall in Henan, 1951-2006.
A simple linear regression confirms that March rainfall influences winter wheat yield significantly, which is consistent with farmers’ experience (see Winter Wheat Management Technology in Spring). The growing period for winter wheat is from October to June. Due to low temperature in the north of China in winter, winter wheat seedlings do not grow after December until March. Then, the seedlings begin to grow fast and get ready to blossom. The drought in March is called the Spring Drought, which, if it occurs, will reduce the winter wheat yield dramatically by delaying the growth or killing the seedlings.

4.3 Contract Design

Consider a stop-loss rainfall index insurance for drought under the assumption that area yield and rainfall index are both stationary. Suppose the average rainfall that affects the yield significantly in the area at time t is $\bar{x}_t$. A proportionate stop-loss insurance contract for $\bar{x}_t$ at time t is a contract that pays a loss-contingent indemnity

$$\tilde{n}_t = I(\bar{x}_t | a, b) = n_0 \min \{ \max \{ \frac{b - \bar{x}_t}{b - a}, 0 \}, 1 \},$$

(2)

where $n_0$ is the maximum indemnity the insurer will pay. It can be normalized to 1. $b$ is the “trigger”, or the maximum, of rainfall level that will induce an indemnity payment, and $a$ is the “stop”, or the rainfall level beyond which no additional indemnity is paid. The indemnity is a function of rainfall given parameters $a$ and $b$.

Naturally, $b > a$. This function means that for the wheat growing area suffering from drought, an indemnity will be paid if rainfall is less than $b$, and the indemnity is a decreasing function of rainfall until rainfall reaches the stop value of $a$, at which the indemnity is $n_0$. Figure 4.4 shows the relationship between rainfall and indemnity.

Suppose the rainfall variable $\tilde{x}$ has a probability distribution function $f(x)$ and cumulative distribution function $F(x)$. The fair premium rate of this contract, $\pi$, is equal to the expected value of indemnity

$$\pi = E\tilde{x} = \int I(x_\epsilon)f(x_\epsilon)dx_\epsilon$$

$$= \int f(x_\epsilon)dx_\epsilon + \int_0^b \frac{b-x_\epsilon}{b-a} f(x_\epsilon)dx_\epsilon + 0$$

$$= F(a) + \frac{1}{b-a}[b\int_a^b f(x_\epsilon)dx_\epsilon - \int_a^b x_\epsilon f(x_\epsilon)dx_\epsilon]$$

$$= F(a) + \frac{1}{b-a}[bF(b) - bF(a) - \int_a^b x_\epsilon F(x_\epsilon)dx_\epsilon]$$

$$= F(a) + \frac{1}{b-a}[bF(b) - bF(a) - \int_a^b x_\epsilon F(x_\epsilon)dx_\epsilon]$$

$$= \frac{1}{b-a} \int_a^b F(x_\epsilon)dx_\epsilon$$
4.4 Optimization Procedure

Given $T$ observations on rainfall $x_t$ and yields $y_t$, I construct a model of the joint distribution of $x$ and $y$ using a selected parametric copula family. Given the fitted distribution, one can use numerical integration methods to compute an estimate of the expected indemnity

$$\pi(\alpha) = EI(\tilde{x}; \alpha).$$

(4)

I assume the insured present risk neutral preferences represented by a constant relative risk aversion (CRRA) utility function $u(x) = \frac{x^{\rho}}{1-\rho}$, where $\rho$ is the coefficient of relative risk aversion, $\rho > 0$ and $\rho \neq 1$. Suppose $\rho$ equals 0.5. The farmers would like to buy the index insurance product only if the expected utility with insurance is greater than or equal to the expected utility without insurance. Then the
willingness to pay $\omega(\alpha)$ can be computed by numerically solving the nonlinear inequality

$$Eu(\tilde{y}) \leq Eu(\tilde{y} + I(\tilde{x}, \alpha) - \omega).$$  \hspace{1cm} (5)

Given a procedure to estimate the fair premium and the willingness to pay, I search for an optimal design, that is, for an optimal value of the contract parameters $\alpha$, by maximizing willingness to pay $\omega(\alpha)$ subject to constraint $\pi(\alpha) = \pi^*$, where $\pi^*$ is a target fair premium level, say 5%. Farmers’ willingness to pay is a function of contract parameters such that inequality (4) is satisfied. Thus, the procedure of estimate the fair premium and willingness to pay can be presented as:

$$\text{Max}_{\omega} \omega(\alpha) \hspace{1cm} \text{s.t.} E I(\tilde{x}; \alpha) = \pi^*,$$  \hspace{1cm} (6)

where, $\omega(\alpha)$ is solved by (5) when the equality is satisfied. The similar method to the previous section is used to solve this maximization problem. For every given value of $b$, one can find a value of parameter $a$ such that $a$ satisfies the restriction to the expectation of indemnity. For each pair of parameters $a$ and $b$, the willingness to pay $\omega$ can be fixed by numerically solving the equation (4) with the pseudo data generated by a specific underlying dependence. Then, the optimal contract design could be found by searching the maximal value of $\omega$ and relative contract parameters $a$ and $b$.

The approach above can be repeated under different assumptions regarding the underlying dependency between rainfall and yields. In particular, one can perform the procedure using conventional assumptions of symmetric tail dependence and then
allowing for varying degrees of tail dependence. A hypothesis to be tested is that conventional approaches lead to significant underestimates of the value of index insurance due to the incidence of substantial lower tail dependency. Table 4.2 reports the highest willingness to pay and relative contract parameters $a$ and $b$ using conventional method approaches, Student-t copula, Clayton copula, Frank copula and Gumbel copula. Figure 4.5-4.7 shows the relationship between willingness to pay and the contract parameter $b$ under different assumptions of underlying dependence.

<table>
<thead>
<tr>
<th></th>
<th>Willingness to pay</th>
<th>$a$</th>
<th>$b$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Conventional method</td>
<td>0.034</td>
<td>49</td>
<td>295</td>
</tr>
<tr>
<td>Clayton copula</td>
<td>0.144</td>
<td>66</td>
<td>515</td>
</tr>
<tr>
<td>Frank copula</td>
<td>0.142</td>
<td>63</td>
<td>545</td>
</tr>
<tr>
<td>Gumbel copula</td>
<td>0.139</td>
<td>71</td>
<td>510</td>
</tr>
</tbody>
</table>

Table 4.2  Willingness to pay and contract parameters

The estimation results show that using copulas to model the underlying dependence structure leads to higher willingness to pay and therefore lower basis risk. By looking at Table 4.2, it is obvious that the willingness to pay of conventional method is only 0.034, which is much lower than the willingness to pay of copulas. The willingness to pay of each of the four copulas is around 0.14. In the three copulas, the Clayton copula has the greatest willingness to pay, 0.144. Since the
Clayton copula can capture the lower tail dependence between rainfall and yield, the high willingness to pay indicates the existence of asymmetric dependence structure and high correlation of rainfall and yield at the lower tail of their distribution. In this case, the conventional methods that apply the Pearson linear correlation coefficient to contract design underestimate the risk at the lower tail, or when drought happens. The threshold of indemnity payment, $b$, is lower with the conventional method than with copulas. The contract based on the conventional methods, therefore, does not pay any indemnity until severe drought happens, which results in quite a lot of risk not covered by the insurance and the basis risk is high. However, when the copulas are used, the tail dependence, asymmetric or symmetric, is modeled, and the pseudo data generated by copulas decline the effect of small sample at the tails. The drought risk is appropriately covered by the insurance contract based on copulas. Thus, the basis risk is lower and farmers’ willingness to pay is higher.
Figure 4.5  Clayton copula

Figure 4.6  Frank copula
Figure 4.7  Gumbel copula

4.5 Conclusion

Index insurance outperforms traditional agricultural insurance in that 1) it effectively reduces asymmetric information problem; 2) it is easy to operate with low costs, and thus low premium; 3) the indemnity can be paid quickly; 4) clients include farmers and institutions. According to the development history and current situation of agricultural insurance in China, conventional agricultural insurance cannot exist without government subsidy. Therefore, index insurance is a good choice for China’s agricultural insurance market.

The challenge for the design of index insurance is how to decline the basis risk
with small sample. Joint distribution is considered to take good use of information in small sample and obtain an optimal contract design. Copulas provide a mechanism to estimate joint distribution of risk variable and index using historical data. Based on the results in previous chapters, the Clayton copula works well in modeling lower tail dependence and then Clayton copula is chosen to design rainfall index insurance for drought in Henan Province, China. Results show that different marginals lead to different contract parameters. The nonparametric marginals and the combination of exponential for rainfall and normal for detrended yield result in similar contract parameters. However, the application of normal distribution for both rainfall and yield tends to underestimate the risk for insurance companies when drought happens.

In the test of willingness to pay, the contracts based on copulas and exponential rainfall and normal yield all have higher willingness to pay than the contract based on conventional methods. In the four copulas I tested, Clayton copula has the highest willingness to pay and therefore has lowest basis risk. This result indicates that lower tail dependence is an important property that cannot be ignored in insurance contract design, and that the appropriate copula function can capture the asymmetric tail dependence and reduce the basis risk. The index insurance based on appropriate copulas overcomes the drawbacks of traditional agricultural insurance in pricing, and the drawbacks of index insurance based on conventional linear correlation coefficient in willingness to pay. Therefore, compared with the
traditional agricultural insurance products, the new index insurance based on copulas will be easier to be applied in poor developing countries.
Chapter 5: Tail Dependence and Spatial Error Model with Variant

Autoregressive Parameters

5.1 Introduction

Tail dependence among Iowa county-level rainfalls was empirically detected among county pairs using bivariate copula functions in Chapter 3. In this chapter, we address whether multivariate tail dependence exists among county-level rainfalls and how this tail dependence varies geographically.

Multivariate copulas may be used to model complicated tail dependence among multivariate data. However, the multivariate Archimedean copulas, which capture the asymmetric tail dependence, are very complicated. Along with the increase of dimension, the number of copula parameters can also increase. Each of Iowa’s 99 counties has four to seven adjacent counties. As a result, an assessment of the performance of copulas requires to compare not only different copula families, but also the copulas with different number of parameters, which complicates the analysis.

When we examine tail dependence among non-adjacent counties, it is natural to ask whether tail dependence is related to how far the two counties are from each other. One would expect counties that are farther apart to exhibit less correlation than the counties that are closer. The degree to which this is so measures the ability of insurers to spatially diversity their portfolio of insurance contracts. If correlations fall of quickly with distance, portfolios can be effectively geographically
diversified; if correlations do not fall off quickly with distance, the opposite is true.

The role of space and distance has attracted increased attention in many economic areas in recent years. For example, geographic distance is considered to influence long-run outcomes and trade (Diamond 1997; Anderson and van Wincoop 2004), and climatic variations tend to affect per-capita incomes (Sachs 2003). Empirical study of such spatial-related economic problems typically requires the specification and estimation of spatial econometric models. The most widely used spatial econometric models include the spatial autoregressive lag and error model.

The standard taxonomy of spatial autoregressive lag and error models were first applied in spatial econometrics to model the spatial dependence (Anselin 1988). Thereafter this approach was extended in Anselin and Bera (1988) and Anselin (2001 and 2003) for the structural form of a spatial regression.

In spatial autoregressive lag and error models, interactions between cross-sectional units are typically modeled by some function of the distance between them. The distance function typically involves a weighting matrix between the sectional units. The most widely used weighting matrices are based on distance, inverse of distance, inverse of distance square, distance thresholds and combinations of distance and distance threshold. Weighting matrices based on distance may be used to model the interactions whose strength increase with distance; weighting matrix based on inverse distance may be used to model interactions whose strength decrease with distance. Weighting matrices can be specified such that an interaction is effective only if the distance between two units is less than or greater than some specified threshold distance. Combined weighting matrices allow us to capture interactions that vary with distance, but are effective only beyond or within a specified
No matter what kind of weighting matrix is used, the spatial autoregressive parameter in conventional spatial autoregressive lag and error model is assumed to be constant. This invariant autoregressive parameter is based on the assumption that the correlation between spatially separated rainfalls does not change. Even in the spatial-temporal model that takes time series into account, the spatial autoregressive parameters are usually set to be a scalar (Kapoor, Kelejian and Prucha, 2006).

However, a model with invariant autoregressive parameter is not appropriate for modeling data with tail dependence. In the case of Iowa county-level rainfalls, the correlation between spatially separated rainfalls tends to be higher in drought years than in normal or flood years. The model with constant autoregressive parameter could underestimate the correlation in drought years and overestimate the correlation in normal or flood years. One way to solve this problem is to estimate spatial autoregressive lag or error model with variant autoregressive parameters. If the parameter is significantly negatively related with the systemic variable, say state average rainfall, we can say that the rainfalls in different counties are more highly correlated in drought years. If the parameter is significantly positively related with the state average rainfall, the rainfalls in different counties are shown to be more highly correlated in flood years.

The spatial autoregressive model is usually estimated by a two-stage least squares procedure or by a generalized moments estimator (Kelejian and Prucha, 1998, 2001 and 2004; Bell and Bockstael, 2000; Kim, Phipps and Anselin, 2003; Lee, 2005; Kapoor, Kelejian and Prucha, 2007). When the variant autoregressive parameter is inserted, the model can also be estimated by the generalized spatial two-stage least
squares method.

In this chapter, I first analyze the autoregressive model with a variant autoregressive parameter. Then, I use maximum likelihood to estimate the cross-sectional model in which the spatial autoregressive parameter is assumed to be a logistic function of the systemic variable. Thereafter, the panel data SEM model with the variant spatial autoregressive parameter is formulated and tested using Monte Carlo simulation. A case study of Iowa county-level rainfalls is also conducted.

5.2 Model Specification

5.2.1 Conventional Spatial Error Model (SEM)

Consider the conventional cross-sectional spatial autoregressive model:

\[
\begin{align*}
  y &= \gamma x + u \\
  u &= \rho Wu + \varepsilon,
\end{align*}
\]

where, \( y \) is the \( N \)-by-1 vector of county level rainfall, \( N \) is the number of counties, \( x \) is the systemic risk, \( W \) is an \( N \times N \) spatial weighting matrix, \( \gamma \) is the systemic parameter, \( \rho \) is the scalar autoregressive parameter, \( u \) is an \( N \)-by-1 vector of regression disturbance, and \( \varepsilon \) is an \( N \)-by-1 vector of innovations. The variable \( Wu \) is usually called “spatial lags” of \( u \). The elements in \( W \), say \( w_{ij} \), represent the distance between meteorological stations \( i \) and \( j \). The diagonal elements of \( W \) are all zeros. The elements in \( W \) are standardized so that the sum of elements in each row is 1.

According to Ord (1975) and Anselin (1988), the spatially lagged dependent variable is correlated with the disturbance term, which leads to the ordinary least
squares estimator typically not consistent. Therefore, model (1) is estimated by a two-stage least squares procedure (2SLS) (Kelejian and Prucha, 2004). The 2SLS procedure is based on the following assumptions:

Assumption 1: All diagonal elements of the spatial weights matrix \( W \) are zero.

Assumption 2: The matrix \( (I - \rho W) \) is nonsingular with \( |\rho| < 1 \).

Assumption 3: The row and column sums of the matrices \( W \) and \( (I - \rho W)^{-1} \) are bounded uniformly in absolute value.

Assumption 4: The observations of innovations \( \{\varepsilon_i : 1 \leq i \leq n, n \geq 1\} \) are distributed independently and identically with \( E(\varepsilon_i) = 0 \) and \( E(\varepsilon_i^2) = \sigma^2 \), where \( \sigma^2 < \infty \).

There are three tests that can be used to detect the presence of spatial autocorrelation in the residuals of a least square model (LeSage, 1999). The first test is to use Moran’s \( I \) statistic. Given that the weighting matrix \( W \) is standardized, Moran’s \( I \) statistic takes the form (LeSage, 1999)

\[
I = e'W/e'e, \tag{2}
\]

where \( e \) represents the residuals of least square regression. According to Cliff and Ord (1972, 1973, 1981), Moran’s \( I \) statistic has the asymptotic normal distribution. Moran’s \( I \) statistic tests whether spatial correlation exists in the least square residuals, and whether the spatial error model is appropriate. The critical value for the Moran’s \( I \) test with confidence level 0.05 is 1.96. If Moran’s \( I \) statistic is greater than 1.96, then the null hypothesis of no spatial correlation is rejected. Otherwise, the null hypothesis cannot be rejected.

The second test is the likelihood ratio test. The likelihood ratio statistic (\( LR \)) is
based on the difference between the log likelihood of the spatial error model and the log likelihood of a least square regression. $LR$ belongs to the chi square distribution with the degree of freedom 1. The critical value of the likelihood ratio test with confidence level 0.05 is 6.635. If $LR$ is greater than 6.635, then the null hypothesis of no spatial correlation is rejected. Otherwise, the null hypothesis cannot be rejected.

The third test of spatial correlation is the Lagrange Multiplier test. The Lagrange Multiplier statistic ($LM$) takes the form

$$LM = \frac{1}{\text{tr}(W+W')*W^e} [e^T We / \sigma^2]^2,$$

where $e$ represents the residuals of a least square regression, and $\sigma$ represents the standard deviation of the residuals. $LM$ also belongs to the chi square distribution with degree of freedom 1 and also has the critical value 6.635. If $LM$ is greater than 6.635, then the null hypothesis of no spatial correlation is rejected.

Table 5.1 shows the test results of the three statistics using four different weighting matrices and the cross-sectional data. $W_1$ represents the weighting matrix based on inverse distance. $W_2$ is based on the inverse distance square. $W_3$ is based on the threshold value 100 miles. Its elements are equal to 1 if the distance between the row county and column county is less than or equal to 100 miles, and the elements are equal to 0 if the distance is greater than 100 miles. $W_4$ is the combination of $W_1$ and $W_3$. The elements are equal to the inverse distance if the distance between the row county and column county is less than or equal to 100 miles, and the elements are equal to 0 if the distance is greater than 100 miles. The tests are conducted for each year from 1954 to 2008 using different weighting matrices. The cells of Table 5.1
report the number of years that have the statistics greater than the critical values, or the number of years in which spatial correlation exists among the county-level rainfalls in Iowa.

<table>
<thead>
<tr>
<th>Statistics</th>
<th>$W_1$</th>
<th>$W_2$</th>
<th>$W_3$</th>
<th>$W_4$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Moran’s $I$</td>
<td>55</td>
<td>55</td>
<td>54</td>
<td>55</td>
</tr>
<tr>
<td>$LR$</td>
<td>49</td>
<td>54</td>
<td>54</td>
<td>54</td>
</tr>
<tr>
<td>$LM$</td>
<td>52</td>
<td>54</td>
<td>54</td>
<td>54</td>
</tr>
</tbody>
</table>

Table 5.1. The number of years in 1954-2008 in which statistics are greater than the critical values.

By using Moran’s $I$ statistic, almost all the years show significant spatial correlation, no matter which weighting matrix is used. The statistics $LR$ and $LM$ are more sensitive to the weighting matrix. In Table 5.1, both of the statistics show that spatial correlation exists in 54 of the total 55 years by using $W_2$, $W_3$ and $W_4$. However, based on $W_1$ there are only 49 years and 52 years with significant spatial correlation by using $LR$ and $LM$, respectively. In other words, $LR$ and $LM$ statistics indicate that the spatial correlation declines fast with the increasing of distance between counties. Table 5.1 also implies that $W_2$ and $W_4$ are more appropriate than $W_1$ and $W_3$ in modeling spatial correlation by spatial error model.
5.2.2 The SEM with Variant Spatial Autoregressive Parameter

In model (3), $\rho$ measures the correlation between rainfall in county and spatial lags. Suppose the spatial structure of county-level rainfalls in Iowa is appropriately modeled by the SEM as in (3). If there is tail dependence, $\rho$ will not be constant, but rather vary with the systemic variable $x$. Higher correlation at the lower tail of the distribution would be indicated if $\rho$ is decreasing in $x$. According to Assumption 2, the value of $\rho$ is within the range (-1, 1). Here I suppose that the relationship between the absolute value of $\rho$ and $x$ is expressed by a logistic regression

$$|\rho(x)| = \frac{1}{1+e^{-(\beta_0 + \beta_1 x)}} ,$$

(4)

where $\beta_0$ and $\beta_1$ are the logistic parameters.

Similar to the conventional SEM, there are some assumptions concerning the function of $\rho(x)$:

Assumption 1: All diagonal elements of the spatial weighting matrix $W$ are zero.

Assumption 2: The matrix $(I - \frac{1}{1+e^{-(\beta_0 + \beta_1 x)}} W)$ is nonsingular with

$$\left|\frac{1}{1+e^{-(\beta_0 + \beta_1 x)}}\right| < 1.$$

Assumption 3: The row and column sums of the matrices $W$ and $(I - \frac{1}{1+e^{-(\beta_0 + \beta_1 x)}} W)^{-1}$ are bounded uniformly in absolute value.

Assumption 4: The observations of innovations $\{\varepsilon_i : 1 \leq i \leq n, n \geq 1\}$ are
distributed independently and identically with $E(\epsilon_i) = 0$ and $E(\epsilon_i^2) = \sigma^2$, where $\sigma^2 < \infty$.

5.3 The Estimation of SEM with Variant Autoregressive Parameter

5.3.1 Estimation Approaches

Kelejian and Prucha (1998) states that in the conventional SEM the disturbance terms have mean zero and variance-covariance matrix

$$
\Omega_u = E(\epsilon u') = \sigma^2 I - \rho W \sigma^2 (I - \rho W)\sigma^2. \quad (5)
$$

The disturbance terms are both spatially correlated and heteroskedastic. Although the ordinary least squares estimator is unbiased, it is typically not consistent. Therefore, the SEM cannot be consistently estimated by ordinary least square.

To estimate the SEM model consistently, Kelejian and Prucha (1998) introduce a two-stage least squares procedure (2SLS) with the instrumental variables. In the first step the least square estimates are constructed and the associated residuals $u_i$ is calculated by:

$$
\hat{u} = y - \hat{\gamma}X. \quad (6)
$$

In the second step he spatial autoregressive model is usually estimated by a generalized spatial two-stage least squares procedure or by a generalized moments estimator. Then, the estimate of $\gamma$ is modified by the estimate of $\rho$.

Another way to estimate the SEM is to use an iterative approach based on the maximum likelihood estimation (LeSage, 1999). First, the least square estimates are constructed and the associated residuals are calculated. Second, given the estimates
of $\gamma$, the estimate of $\rho$ is found by maximizing the log likelihood function. Third, the least square estimates of $\gamma$ is updated using the estimate of $\rho$ determined in step two. This process is continued until convergence in the residuals.

The two-stage least square procedure by Kelejian and Prucha (1998) and the iterative approach based on the maximum likelihood estimation by LeSage (1999) can both estimate the conventional spatial error model consistently. However, when it comes to the SEM with variant spatial autoregressive model, the second step that involves GMM in Kelejian and Prucha (1998) becomes complicated because $\rho$ is a nonlinear function of $x$. Hence, I apply LeSage’s iterative approach to estimate the SEM with a variant spatial autoregressive parameter.

5.3.2 Estimation Results

The analysis of the SEM with a variant spatial autoregressive parameter involves two steps. The first step is to estimate the conventional cross-sectional SEM for each year using LeSage’s iterative approach. The second step is to check whether there is a trend of the spatial autoregressive parameter $\rho$ with respect to the systemic variable $x$ and estimate the logistic regression of $\rho$ on $x$.

The relationship between $\rho$ and $x$ is shown in Figure 5.1. Each plot represents a pair of $(\rho, x)$ in each of the 55 years from 1954 to 2008. By looking at Figure 5.1, most of $\rho$’s are within the range [0.5, 0.9]. There is an obviously decreasing trend of $\rho$ on the state average rainfall $x$. When the state average rainfall is less than 600, almost one half of the $\rho$’s are larger than 0.7. However,
when the state average rainfall is greater than 600, only one \( \rho \) out of the ten \( \rho \)'s is larger than 0.7. Although the linear model can capture the relation between \( \rho \) and \( x \), considering the limitation of \( \rho \), I use the logistic regression to model the relation between \( \rho \) and \( x \). The parameters \( \beta_0 \) and \( \beta_1 \) are estimated using the four weighting matrices. The estimates are reported in Table 5.2.

<table>
<thead>
<tr>
<th>Coefficients</th>
<th>( W_1 )</th>
<th>( W_2 )</th>
<th>( W_3 )</th>
<th>( W_4 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>( \hat{\beta}_0 )</td>
<td>1.34*** (0.22)</td>
<td>1.43*** (0.23)</td>
<td>1.41*** (0.23)</td>
<td>1.44*** (0.20)</td>
</tr>
<tr>
<td>( \hat{\beta}_1 )</td>
<td>-0.001* (0.0004)</td>
<td>-0.001* (0.0004)</td>
<td>-0.001* (0.0004)</td>
<td>-0.001* (0.0004)</td>
</tr>
</tbody>
</table>

***: significant at level 0.001. *: significant at level 0.05

Table 5.2 Estimates of parameters in Equation (5)
Figure 5.1 Plots of the autoregressive parameter $\rho$ with respect to $x$

By looking at Table 5.2, it is obvious that the spatial structural model have the disturbance term $u$ is inversely related to the systemic variable $x$. The regression results do not differ much when different weighting matrices are used. According to the analysis of model (3), if $\rho$ is decreasing with $x$, then the higher correlation at the lower tail of the distribution can be concluded. The negative estimates of the spatial parameter imply that lower left tail dependence exists in county-level rainfall of Iowa. In widespread drought years when the average precipitation of Iowa is low,
the correlation between counties becomes lower than that in normal years when the average precipitation of Iowa is around its mean. The estimation of model (3) with function $\rho(x_t)$ is a strong evidence for higher correlation at the lower tail. The conventional SEM that set the spatial autoregressive parameter fixed may underestimate the spatial correlation in the case of Iowa rainfall when the widespread drought occurs.

### 5.4 Panel Data Model

#### 5.4.1 The Model

In Section 5.3, the cross-sectional SEM is estimated for each of the 55 years, and the spatial autoregressive parameter $\rho$ is assumed to be a logistic function of the systemic variable $x$. The next question is whether $\rho(x_t)$ can be plugged into the SEM and estimated using the panel data. Kapoor, Kelejian and Prucha (2007) estimate the conventional SEM using panel data. The panel data model employed by Kapoor, Kelejian and Prucha (2007) has a constant spatial autoregressive parameter. Suppose there are $N$ counties. By introducing the variant spatial autoregressive parameter as a function of systemic variable $x$, the panel data model is expressed by

$$y_N(t) = \gamma x(t) + u_N(t)$$

$$u_N(t) = \rho(x(t))Wu_N(t) + \varepsilon_N(t),$$

$$|\rho(x(t))| = \frac{1}{1 + e^{-(\beta_0 + \beta x(t))}}$$

where $W$ is an $N \times N$ weighting matrix of known constants which does not involve $t$. $y_N(t)$ denotes the $N \times 1$ vector of observations on the dependent variable in period $t$. $u_N(t)$ denotes the $N \times 1$ vector of errors in period $t$. $\varepsilon_N(t)$ denotes the $N \times 1$ vector of i.i.d. errors in period $t$. $\rho(x(t))$ is the spatial autoregressive parameter as a function of systemic variable $x$. $\gamma$ is the systemic coefficient. $\beta_0$ is the systemic intercept. $\beta$ is the systemic coefficient vector.
\( x(t) \) represents the systemic risk in period \( t \). \( u_N(t) \) and \( E_N(t) \) are both the \( N \times 1 \) vectors representing the disturbance and innovations in period \( t \), respectively.

Following Kapoor, Kelejian and Prucha (2007), I allow the innovations to be correlated over time by assuming the following error component structure:

\[
E_N = (e_T \otimes I_N) \mu_N + v_N, 
\]

where \( \mu_N \) is the vector of unit specific error components, and \( v_N = [v_N(1),...v_N(T)]' \) represents the error components that vary over both the cross-sectional units and time periods. In scalar notation,

\[
E_{i,t,N} = \mu_{i,N} + v_{i,t,N}. 
\]

There are three extra assumptions. First, \( v_{i,t,N} \) are independently and identically distributed with zero mean and finite variance \( \sigma_v^2 \). Second, \( \mu_{i,N} \) are independently and identically distributed with zero mean and finite variance \( \sigma_{\mu}^2 \). Third, \( v_{i,t,N} \) and \( \mu_{i,N} \) are independent. These assumptions imply that \( E v_{i,t,N} = 0 \), and

\[
E(\varepsilon_{i,t,N}\varepsilon_{j,s,N}) = \begin{cases} 
\sigma_v^2 + \sigma_{\mu}^2, & \text{if } i = j, t = s \\
\sigma_{\mu}^2, & \text{if } i = j, t \neq s \\
0, & \text{otherwise}
\end{cases}
\]

The innovations \( \varepsilon_{i,N} \) are autocorrelated over time, but are not spatially correlated across counties.

5.4.2 A Monte Carlo Simulation

In the following, I report on a Monte Carlo study of the panel data SEM with variant spatial autoregressive parameters. I consider the estimates of the spatial
autoregressive parameters $\beta_0$, $\beta_1$ and variances $\sigma^2_\mu$ and $\sigma^2_\nu$, and corresponding parameter $\gamma$. I also consider LeSage’s iterative process using maximum likelihood estimation. For purposes of comparison, I use $W_2$ to generate data and use $W_1$, $W_2$, $W_3$ and $W_4$ to estimate the parameters.

Similar to Kapoor, Kelejian and Prucha (2007) simulation, all of my Monte Carlo experiments $N = 100$ and $T = 5$. The data are generated according to (7) and (8). The observations on $x$ correspond to the state average June rainfall in Iowa over the period 2004-2008. Suppose $\rho(x(t))$ is positive. The parameter $\gamma$ is taken to be equal to one. I assume that both of the error components, $\mu_N$ and $\nu_N$, are normally distributed. The generation of $\mu_N$ requires the specification of $\sigma^2_\mu$, $\sigma^2_\nu$, $\beta_0$, $\beta_1$ and the weighting matrix $W$. I set $\sigma^2_\mu = \sigma^2_\nu = 1$ in all of my Monte Carlo experiments. $\beta_0$ takes three values, -1, 0 and 1. $\beta_1$ takes five values, -0.9, -0.5, 0, 0.5 and 0.9. In the procedure of generating data, $W$ takes the form $W_2$, the inverse distance square. To summarize, my Monte Carlo simulation contains 15 cases, which range from three selections of $\beta_0$ and five selections of $\beta_1$. I also consider four weighting matrices as in the earlier study to model the generated data. $W_1$ represents the weighting matrix based on inverse distance. $W_2$ is based on the inverse distance square. $W_3$ is based on the threshold value 100 miles. Its elements are equal to 1 if the distance between the row county and column county is less than or equal to 100 miles, and the elements are equal to 0 if the distance is greater than 100 miles. $W_4$ is the combination of $W_1$ and $W_3$.

Table 5.3 and 5.4 present measures of dispersion for the pseudo sample distribution of estimates $\hat{\beta}_0$ and $\hat{\beta}_1$ using the four different weighting matrices.
The measure of dispersion I adopt is the Kapoor et al. (2007) root mean squared error (RMSE). The difference between Kapoor et al. (2007) RMSE and the standard measure of RMSE is that Kapoor et al.’s RMSE is based on quantiles, and the standard RMSE is based on moments. The reason why I use Kapoor et al. (2007) RMSE is that quantiles are assured to exist. Kapoor et al.’s RMSE is defined as

\[
RMSE = \left[ bias^2 + \left( \frac{IQ}{1.35} \right)^2 \right]^{1/2},
\]

(11)

where \(bias\) is the difference between the median and the true value of the parameter, and \(IQ\) is the interquantile range defined as the 0.75 quantile minus the 0.25 quantile. If the distribution is normal, then the median is the mean and \(IQ/1.35\) is the standard deviation, and then Kapoor et al.’s RMSE is the same as the standard RMSE. Table 5.3 and 5.4 are the results of Kapoor et al.’s RMSE based on a 100 replications for each of the 15 cases and each of the four weighting matrices.
<table>
<thead>
<tr>
<th>Parameter values</th>
<th>RMSE</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$\hat{\beta}_{0,1}$</td>
</tr>
<tr>
<td>-1</td>
<td>0.07</td>
</tr>
<tr>
<td>-1</td>
<td>0.09</td>
</tr>
<tr>
<td>-1</td>
<td>0.10</td>
</tr>
<tr>
<td>-1</td>
<td>0.08</td>
</tr>
<tr>
<td>-1</td>
<td>0.08</td>
</tr>
<tr>
<td>0</td>
<td>0.07</td>
</tr>
<tr>
<td>0</td>
<td>0.07</td>
</tr>
<tr>
<td>0</td>
<td>0.09</td>
</tr>
<tr>
<td>0</td>
<td>0.08</td>
</tr>
<tr>
<td>0</td>
<td>0.09</td>
</tr>
<tr>
<td>1</td>
<td>0.05</td>
</tr>
<tr>
<td>1</td>
<td>0.06</td>
</tr>
<tr>
<td>1</td>
<td>0.07</td>
</tr>
<tr>
<td>1</td>
<td>0.05</td>
</tr>
<tr>
<td>1</td>
<td>0.05</td>
</tr>
</tbody>
</table>

Note: $\hat{\beta}_{0,1}$ is the maximum likelihood estimator of $\beta_0$ based on weighting matrix $W_1$; $\hat{\beta}_{0,2}$ is the maximum likelihood estimator of $\beta_0$ based on weighting matrix $W_2$; $\hat{\beta}_{0,3}$ is the maximum likelihood estimator of $\beta_0$ based on weighting matrix $W_3$; $\hat{\beta}_{0,4}$ is the maximum likelihood estimator of $\beta_0$ based on weighting matrix $W_4$.

Table 5.3  RMSEs of the estimates of $\beta_0$

The results in Table 5.3 and Table 5.4 suggest that the RMSEs of $\hat{\beta}_0$ and $\hat{\beta}_1$ based on four weighting matrices are quite similar. The RMSEs of $\hat{\beta}_0$ based on $W_2$ are, on average, approximately 10% smaller than those of $\hat{\beta}_0$ based on $W_1$, $W_3$ and $W_4$. The RMSEs of $\hat{\beta}_1$ based on $W_2$ are, on average, approximately 8% smaller than those of $\hat{\beta}_1$ based on $W_1$, $W_3$ and $W_4$. The similarity of the RMSEs of the four
weighting matrices suggests that, at least for the 15 parameters combination cases, the
selection of weighting matrix is not “very important” in determining the efficiency of
the corresponding maximum likelihood estimators. Table 5.3 and 5.4 also indicate
that RMSEs are smaller when the value of $\rho$ is greater. For example, when
$(\hat{\beta}_0, \hat{\beta}_1)$ is equal to (-1, -0.009) and the corresponding $\rho$ equals 0.004, the RMSEs
of $\hat{\beta}_0$ are around 0.07 and the RMSEs of $\hat{\beta}_1$ are around 0.0001. When $(\hat{\beta}_0, \hat{\beta}_1)$ is
equal to (-1, -0.005) and the corresponding $\rho$ equals 0.03, the RMSEs of $\hat{\beta}_0$ are
around 0.09 and the RMSEs of $\hat{\beta}_1$ are around 0.0003.
<table>
<thead>
<tr>
<th>Parameter values</th>
<th>RMSE</th>
</tr>
</thead>
<tbody>
<tr>
<td>( \beta_0 )</td>
<td>( \hat{\beta}_{1,1} )</td>
</tr>
<tr>
<td>-1</td>
<td>0.0001</td>
</tr>
<tr>
<td>-1</td>
<td>0.0003</td>
</tr>
<tr>
<td>-1</td>
<td>0</td>
</tr>
<tr>
<td>-1</td>
<td>0.0004</td>
</tr>
<tr>
<td>0</td>
<td>-0.009</td>
</tr>
<tr>
<td>0</td>
<td>-0.005</td>
</tr>
<tr>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>0</td>
<td>0.0007</td>
</tr>
<tr>
<td>1</td>
<td>-0.009</td>
</tr>
<tr>
<td>1</td>
<td>-0.005</td>
</tr>
<tr>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>1</td>
<td>0.005</td>
</tr>
</tbody>
</table>

Note: \( \hat{\beta}_{1,1} \) is the maximum likelihood estimator of \( \beta_1 \) based on weighting matrix \( W_1 \); \( \hat{\beta}_{1,2} \) is the maximum likelihood estimator of \( \beta_1 \) based on weighting matrix \( W_2 \); \( \hat{\beta}_{1,3} \) is the maximum likelihood estimator of \( \beta_1 \) based on weighting matrix \( W_3 \); \( \hat{\beta}_{1,4} \) is the maximum likelihood estimator of \( \beta_1 \) based on weighting matrix \( W_4 \).

Table 5.4 RMSEs of the estimates of \( \beta_1 \)

5.4.3 The Case of Iowa County-Level Rainfalls

In this section, I employ maximum likelihood to estimate multivariate tail dependence among Iowa county-level rainfalls using the panel data SEM with variant spatial autoregressive parameters. The weighting matrix is \( W_2 \), the inverse distance square. In the first step, I estimate the least square regression of county-level rainfall
y on the systemic variable $x$. The residuals $u$ are calculated. In the second step, the maximum likelihood function is constructed based on equation (9) and (10), given that the innovations $\epsilon_i$ is the sum of $\mu_i$ and $\nu_i$ are both normally distributed with mean zero and variance $\sigma_\mu$ and $\sigma_\nu$, respectively. Then, the parameter of systemic risk which has been estimated in the first step is replaced by the new estimate based on the maximum likelihood estimation in step two. These procedures are conducted iteratively until convergence. The estimation results are reported in Table 5.5.

<table>
<thead>
<tr>
<th>Parameters</th>
<th>$\hat{\beta}_0$</th>
<th>$\hat{\beta}_i$</th>
<th>$\hat{\gamma}$</th>
<th>$\hat{\sigma}_\mu$</th>
<th>$\hat{\sigma}_\nu$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Estimates</td>
<td>1.13</td>
<td>-0.0009</td>
<td>1.04</td>
<td>0.006</td>
<td>0.0001</td>
</tr>
</tbody>
</table>

Table 5.5 The Estimation Results

5.5 Conclusion

In this chapter, I model multivariate tail dependence using a spatial autoregressive error model with a variant spatial autoregressive parameter. Given that the spatial autoregressive parameter must lie within (-1, 1), I model the parameter as a logistic function of the systemic variable and test whether the spatial autoregressive parameter is significantly related to the systemic available. By checking the plot of estimates of spatial autoregressive parameter based on cross-sectional SEM for each year, there is strong evidence of the correlation between spatial autoregressive parameter and the systemic variable. These results indicate
that the spatial correlation among county-level rainfalls in Iowa is not constant, but
rather varies with the systemic weather events. The results also imply that the
conventional spatial autoregressive model with constant spatial autoregressive
parameter is not appropriate in modeling the spatial variables with tail dependence.
In addition, the autoregressive parameter if found to be inversely related to the
systemic variable, which implies lower tail dependence. When a widespread
drought occurs, the spatial correlation among county rainfalls tends to be higher than
in more normal years. The selection of weighting matrix does not have significant
effect on the variation of spatial correlation.

I used a Monte Carlo approach to analyze a panel data SEM with variant
spatial autoregressive parameters. By comparing the performance of panel data
model with four different weighting matrices using the measure of dispersion, RMSE,
I found little difference between the weighting matrices. The increasing of spatial
autoregressive parameter tends to increase the dispersion of the estimates. After the
Monte Carlo simulation, I estimate the panel data model with the Iowa county-level
rainfalls by the maximum likelihood approach. The negative coefficient of systemic
variable with respect to the spatial autoregressive parameter indicates that lower tail
dependence exists among the rainfalls. That is, when there is widespread drought,
the spatial correlation among county-level rainfalls tends to increase. This result is
consistent with my conclusion in previous chapter about the variation of correlation
between adjacent counties using bivariate copulas.

The analysis of SEM with a variant spatial autoregressive parameter, using
cross-sectional and panel data, is based on the assumption of particular functional
form of the spatial autoregressive parameter. Therefore, the existence of tail
dependence that is identified by the SEM model is also based on the structural assumption. The possible problem with this assumption is misspecification problem. Then the next step would be to test the robustness of the conclusion with different structural assumptions.

There are several suggestions for future research. First, it would be of interest to extend the panel data SEM to the panel data spatial autoregressive lag model. In doing this it would be of interest to consider the variant spatial autoregressive lagged parameter. Second, another area of interest would be to extend to the nonparametric or semi-parametric model, which do not rely on the structural assumption and therefore do not have misspecification problem. Third, the investigation of correlation among index insurance indices could be extended to the correlation between indices and yields or losses.
References


Bryla, E., and J. Syroka. 2007. Developing Index-Based Insurance for Agriculture in Developing Countries. United nations Department of Economic and Social Affairs Devison for Sustainable Development Policy Integration and Analysis Branch Innovation Brief, New York, NY.


http://www3.interscience.wiley.com/journal/122544792/abstract?CRETRY=1&SRETRY=0


