ESTIMATION OF PREMIUM RATES WITH SCARCE DATA: A SIMULATION STUDY ON AQUACULTURE INSURANCE

by

Zhihua Li

A Thesis Submitted to the Faculty of the

DEPARTMENT OF AGRICULTURAL AND RESOURCE ECONOMICS

In Partial Fulfillment of the Requirements
For the Degree of

MASTER OF SCIENCE

In the Graduate College

THE UNIVERSITY OF ARIZONA

2005

STATEMENT BY AUTHOR

This Thesis has been submitted in partial fulfillment of requirements for an advanced degree at The University of Arizona.

Brief quotations from this thesis are allowable without special permission, provided that accurate acknowledgement of source is made. Request for permission for extended quotation from or reproduction of this manuscript in whole or in part may be granted by the head of the major department or the Dean of the Graduate College when in his or her judgment that proposed use of the material is in the interests of scholarship. In all other instances, however, permission must be obtained from the author.

SIGNED:_____

APPROVED BY THE THESIS DIRECTOR

This thesis has been approved on the date shown below:

_____ _____

Alan P. Ker Date

Department Head and Associate Professor

Agricultural and Resource Economics

ACKNOWLEDGEMENTS

First, I would like to express my deep and sincere gratitude to my major advisor, Dr. Alan P. Ker for his constructive guidance, patience, sincere encouragement, and expertise throughout the writing of this thesis. I truly know that without him this thesis would not have been accomplished.

I would also like to thank my committee members, Dr. Gary D. Thompson and Dr. Paul Wilson. Their excellent comments and advice helped to improve this thesis.

My dear husband, my parents, and my family encouraged me throughout this thesis. Their love and support accompanied me during the difficult time.

Finally, many thanks go to all my fellow students, faculty and staff members of the Department of Agricultural and Resource Economics for their unforgettable support.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ABSTRACT

The general purpose of this thesis is to investigate the small sample performances of various rating methodologies in estimating premium rates for catfish insurance policy. To accomplish this objective, we simulate the yield data by modeling the frequency, severity, temporal and spatial correlations of twenty major risk factors in catfish production. In order to increase the applicability of this simulation study, twelve scenarios are considered to see how the methodologies will perform under different assumptions. Simulation results show that, under these data generating processes, estimators that use extraneous data generally perform better than those that only use individual data when the sample size is small.

1. INTRODUCTION

1.1 Introduction

Aquaculture is defined as the farming of aquatic organisms, including fish, mollusks, crustaceans and aquatic plants, where farming implies some form of intervention in the rearing process to enhance production, such as regular stocking, feeding, protection from predators, etc. (FAO 1997). Although aquaculture has existed for thousands of years, it only became a specialized agricultural business in the United States in the 1950's. Aquaculture species grown in the United States include finfish (catfish, trout, salmon, striped bass, tilapia, baitfish, ornamental fish, and others), crustaceans (crawfish, shrimp, and others), mollusks (oysters, clams, mussels, and others), aquatic plants (algae, seaweeds, water chestnuts, hyacinths, and others), and some reptiles such as alligators and turtles (Hanfman, 1993).

For the past two decades, aquaculture has been the fastest growing segment of agriculture in the United States. In 1983, aquaculture production was 308 million pounds with a final sales value of 259 million dollars. In 2001, aquaculture production has exceeded 800 million pounds valued at 935 million dollars (see Figure 1.1). Carlberg et al. (2001) explained several factors that have contributed to this phenomenal growth. First, per capita consumption of seafood has increased significantly as consumers become more aware of the health and nutritional benefits of fish and the fact that seafood is a good source of animal protein. For example, the U. S. per capita consumption of seafood has risen from 12.5 pounds in 1980 to 15.6 pounds in 2002, a 25% increase (NMFS 2002). Furthermore, as the population of the United Stated continues to grow and capture

fisheries are approaching their maximum harvest levels, aquaculture will be the major source of additional seafood supply in meeting the increasing consumer demands. The U. S. Congress summarized the importance of aquaculture in the National Aquaculture Act of 1980: "… aquaculture has the potential for reducing the United States trade deficit in fisheries products, for augmenting existing commercial and recreational fisheries, and for producing other renewable resources, thereby assisting the United States in meeting its future food needs and contributing to the solution of world resource problems. It is, therefore, in the national interest, and it is the national policy, to encourage the development of aquaculture in the United States."

Catfish farming originated in the Mississippi Delta region in the late 1960's and early 1970's (Dean et al., 2003). Today, the catfish industry is the largest sector in the U.S. aquaculture industry. Production of farm-raised catfish has grown rapidly to approximately 597 million pounds in 2001 and accounted for more than 70% of the annual aquaculture production in the United States (see Figure 1.2). Farm-raised catfish generated a final sales value of 386 million dollars and accounted for 41% of the total sales of aquaculture products (see Figure 1.3). It is now the fifth most popular fish in the United States behind shrimp, tuna, salmon, and Alaska pollock. Table 1.1 shows the top 10 fish and shellfish consumption in the United States. Per capita consumption of catfish has doubled since 1990, reaching an all time high of 1.16 pounds in 1999 (see Figure 1.4). The popularity of farm-raised catfish is due to its consistent quality, delicate flavor, firm texture, versatility, year-around availability, and nutritional value (Robinson and Avery, 2000).

Catfish production is concentrated in the southern United States consisted of Alabama, Arkansas, Louisiana, and Mississippi. These states have warm climates, abundant water and heavy clay soils for pond construction, which are conditions favorable to commercial catfish production. These four states account for 95% of catfish production. Mississippi dominates the other three states by producing 70% of the total. The industry provides over 13,000 jobs in production, processing, feed manufacturing, and related support industries, and contributes more than $4 billion to the four states' economy annually (Robinson and Avery, 2000). The catfish industry has become the major source of economic activities and employment in these states. For example, in Mississippi, the catfish industry employs over 7000 direct employees with an annual payroll of $102 million (Dean et al., 2003). Engle (2003) stated that the overall impact of this industry is even greater because it is centered in a region of the country that is characterized by low levels of economic development and high unemployment rates.

As with other agriculture enterprises, catfish producers also face a variety of production hazards. The major perils include diseases, water related problems, off-flavor, bird predation etc., which significantly affect the profitability of the industry and hinder its further development. Given the importance of the catfish industry in terms of its economic value, the Risk Management Agency of the United States Department of Agriculture has begun to investigate the feasibility of providing insurance tools for catfish producers against losses.

Due to the nature of the catfish aquaculture production practices, the implementation of aquaculture insurance to the catfish industry will present a number of

challenges (Shaik, 2001). The first challenge is related to insurability issue. It is difficult to differentiate random peril from management events, and epidemic from endemic hazards. Also, as with crop insurance, there exist the problems of adverse selection and moral hazard. The second challenge is the measurability problem. It is difficult to determine the numbers or pounds of fish in the pond. Difficulties are due to a number of factors: there are multiple batches in a pond at the same time; ponds may be in continuous operation for several years before completely drained; the majority of mortalities are unseen; and auditing practices are inconsistent (Avery, 2002). Catfish insurance contracts require the verification of the numbers (or pounds) of fish to be insured, and the numbers (or pounds) of fish lost, when a claim is made. For a contract that is made for a specific peril such as off-flavor, it also requires verification whether the cause of loss is associated with that peril. All of these are not easy to achieve. The third challenge is related to the actuarial issues. In crop insurance contracts, premium rates are determined based on historical yield data available at the time of rating. Catfish insurance is just a pilot program; the industry does not have the long-term production data. Estimated catfish yields have not been systematically measured, and some of the information needed to calculated aggregate yields has only recently been collected (Kazmierczak and Soto, 2001). Therefore, the data to estimate the probability and magnitude of losses is not available and the potential for subjective data to be collected is also low.

The contribution of this thesis is to conduct simulations to generate the yield data based on some possibly relevant data generating processes, and to evaluate the

performance of various parametric and nonparametric approaches in determining premium rates. These data generating processes involve the various major risk factors associated with catfish production. After identifying the risks and making assumptions about them, yield data can be simulated by modeling the frequency, severity, as well as the temporal and spatial correlations of all those risk factors. Then, various parametric and nonparametric techniques are employed to estimate actuarially fair premium rates. These approaches are considered in an attempt to minimize inefficiencies or inequities in the catfish insurance program, to both the insurer and insured. Recovering accurate premium rates is essential to an actuarially sound catfish insurance program. If the premium rates are overestimated or underestimated, program losses will increase because only producers whose rates are underestimated will participate in the program while producers with overestimated rates will either not be able to purchase insurance (too expensive) or will purchase insurance at a higher cost relative to a fair level. Of course, these adverse selection losses cannot be eliminated. Even with abundant data, one still could not estimate the premium rates without any errors. However, the losses may be minimized by appropriate choices of estimation methodologies. Hence, the performances of the different methodologies are compared based on the mean squared errors of the estimated rates.

The remainder of this chapter contains three sections. The next section outlines the problems of moral hazard and adverse selection. These problems are common in any insurance contracts and are the major causes of program losses. Hence the construction of actuarially fair premium rate is very important. Section 1.3 discusses the objectives of

this thesis, which is to improve the accuracy of premium rates under the situation of no data available. The last section outlines the structure of this thesis.

1.2 Moral Hazard and Adverse Selection

As with crop insurance contracts, catfish insurance contracts are exposed to moral hazard and adverse selection problems. Both problems are aspects of asymmetric information. Asymmetric information arises due to differential information concerning production practices and growing conditions held by the insured (catfish producers) and the insurers. Existence of information asymmetry increases the costs and challenges the efficacy of insurance.

Moral hazard arises when the insurer faces a fixed payment scheme but the insured can affect the probability of risks occurring by hidden actions that are not observed by the insurer. Without insurance, a catfish producer will try his best to reduce the likelihood of undesirable outcomes. However, after purchasing insurance, the producer loses some of the incentive to reduce the probability of adverse outcomes and hence may increase the probability of risks occurring. For example, the insurer cannot observe production practices throughout the production season. A catfish producer may fail to provide treatment for diseases or aeration equipments for catfish production. Through actions unknown to the insurer at the time of the contract, the producer has altered the yield distribution. The effect of moral hazard is that the post-insurance risk has increased.

Adverse selection arises when the insured has better information about the risk probabilities than the insurer when setting the premium rate. In most cases, catfish producers know more about their yield distributions than insurers. Producers whose expected loss is larger than the insurance premium will tend to buy the insurance whereas producers whose expected loss is smaller than the insurance premium will tend not to buy insurance. Thus, the premiums calculated based on the information of all potential clients tends to be too low to cover the indemnity payments, resulting in a loss to the insurer.

Moral hazard and adverse selection problems have resulted in large losses in the crop insurance program (Ker, 1996). To avoid the same losses in catfish aquaculture insurance program, the construction of actuarially fair premium rate is crucial.

1.3 Objective of the Study

The main objective of this thesis is to evaluate the small sample performances of various rating methodologies under the situation of no data available, and come up with some appropriate methodologies that may improve the accuracy of premium rates. Accurate premium rates require proper representation of the conditional yield distributions. In satisfying this objective, this thesis conducts simulations to generate yield data and employs parametric and nonparametric approaches to determine premium rates for two time periods through fifty time periods. These approaches to modeling yield distributions include: normal distribution, beta distribution, kernel density estimation, Bayesian nonparametric estimation, etc. This simulation study allows us to

evaluate the different approaches at various levels of simulated historical yield data. If historical yield data were available, one could employ time series models to predict future yields and use parametric or nonparametric approaches to construct the premium rates. However, catfish insurance, or aquaculture insurance is just a pilot program. While there might be data on country-level and state-level, there is no data on county-level or farm-level. Therefore, simulations are needed to generate possible yield data under relatively reasonable assumptions. Comparison of the different methodologies is based on mean squared errors (MSE). The MSE is an error metric that captures both the bias and the variance of an estimator. In order to increase the practicability of this simulation study, different simulation scenarios are considered.

1.4 Plan of the Study

The remainder of this thesis is organized as follows.

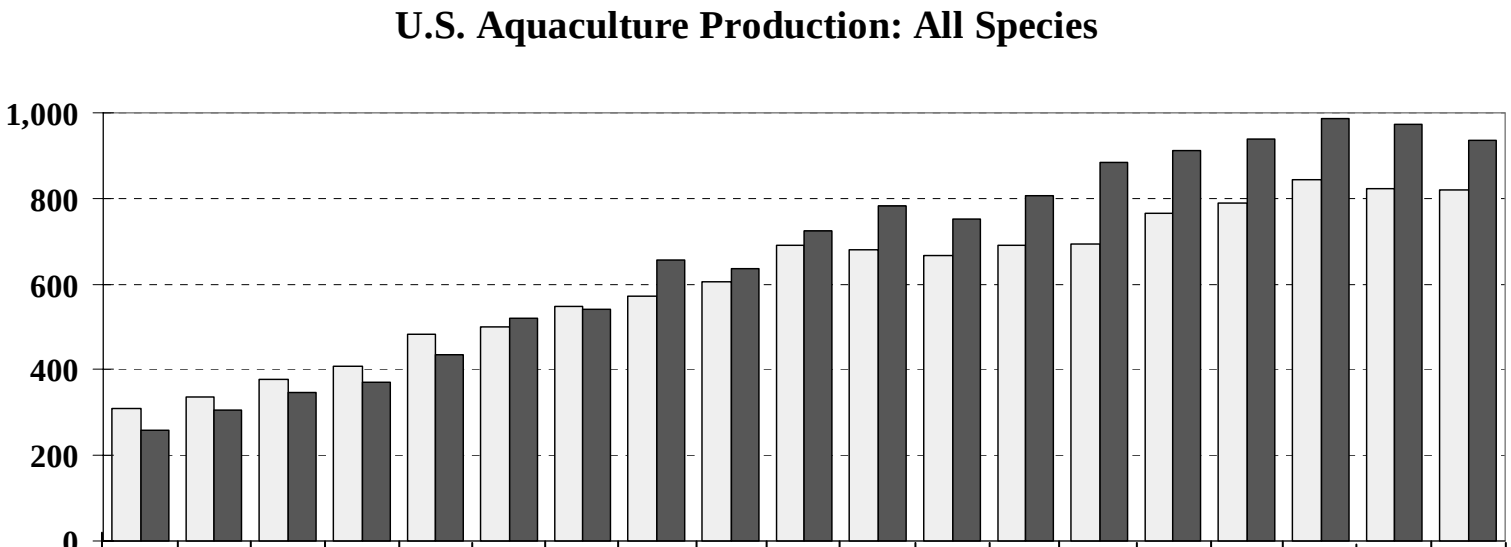
Chapter 2 first reviews the catfish aquaculture farming practices and associated production risks; then discusses how the major risks are modeled based on the characteristics of the risks. The production process of catfish involves three stages: egg and fry production, fingerling production and food fish production. The major production risks include: infectious diseases, water quality related risks, off-flavor, bird predation, extreme weather conditions, etc. Twenty risk factors will be examined and modeled in the simulations, based on the information about the frequency and severity of each risk factor as well as their temporal and spatial correlations.

Chapter 3 elaborates on various parametric and nonparametric rating methodologies. The parametric methodologies include normal distribution and beta distribution. The nonparametric methodologies include empirical rates, kernel density estimation, Bayesian nonparametric kernel density estimation, and estimation of possibly similar densities.

Chapter 4 presents the designs of twelve simulation scenarios and the accompanying MSE results. It first describes the yield data generating process in the base scenario and then considers three variations from the base scenario to design other scenarios. The performances of the twelve methodologies are compared both horizontally (in each scenario) and vertically (across scenarios, the first scenario being the baseline).

Finally, Chapter 5 presents the concluding remarks and directions of further research.

Figure 1.1: US Aquaculture Production (data are in millions)



	1983	1984	1985	1986	1987	1988	1989	1990	1991	1992	1993	1994	1995	1996	1997	1998	1999	2000	2001
Pounds	308	336	376	409	484	500	548	573	604	691	679	666	690	694	767	790	842	823	819
Dollars	259	307	348	372	437	520	540	655	636	724	782	751	806	886	910	939	987	973	937

Source: 1. 2002 Fisheries of the United States, Fisheries Statistics & Economics Division
2. http://www.msstate.edu/dept/crec/aquallspec.html

Figure 1.2: Components of Aquaculture (Thousand Pounds)

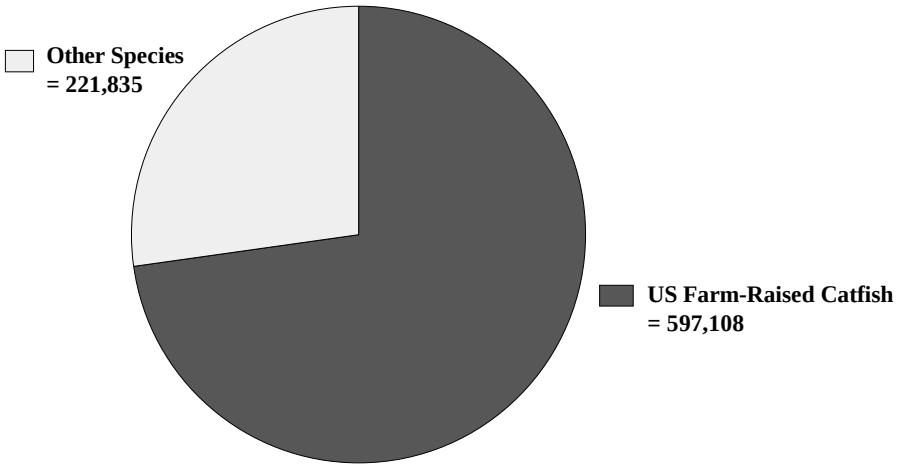
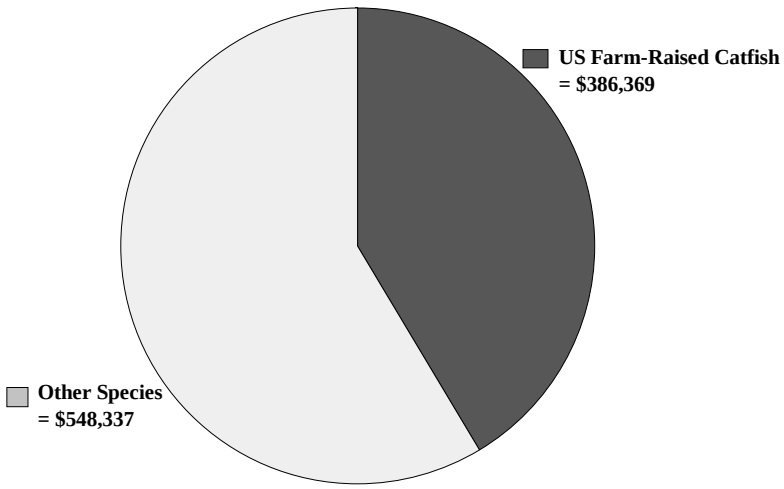
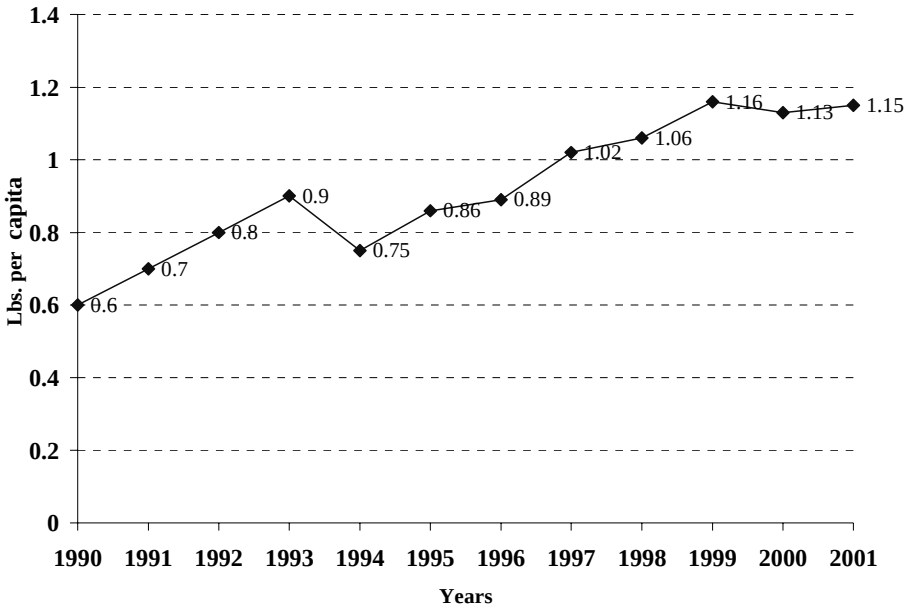


Figure 1.3: Components of Aquaculture (Sales in thousand $)



Source: 2002 Fisheries of the United States, Fisheries Statistics & Economics Division

Figure 1.4: Per Capita Consumption of Catfish



Source: 1. **http://www.catfishinstitute.com/About/Fact.asp**

2. Economic Impact of the Mississippi Farm-Raised Catfish Industry at the year of 2003

Table 1.1: Top 10 Fish and Shellfish Consumption in the US (2001, Edible Meat Basis)

Species	Per Capita Lb/Yr.
Shrimp(27% from Aq.)	3.40
Tuna	2.90
Salmon (50% from Aq.)	2.02
Alaska Pollock	1.21
Catfish (100% from Aq.)	1.15
Cod	0.56
Clams	0.46
Crabs	0.44
Flatfish	0.39
Tilapia	0.35

Source: Economic Impact of the Mississippi Farm-Raised Catfish Industry at the Year of 2003 Aq. = Aquaculture

2. CATFISH AQUACULTURE FARMING PRACTICES AND RISKS

2.1 Channel Catfish

There are at least 39 species of catfish in North America, but only six of them have been cultured or have the potential for commercial production (Wellborn, 1988). The channel catfish, *Ictalurus punctatus*, dominates in farming because it has the best combination of characteristics for commercial production (Jensen, 1997). Therefore, this thesis will focus on the culture of channel catfish. The word *catfish* refers to the channel catfish unless indicated otherwise.

Channel catfish are warm water fish native to central North America (Tucker et al., 2004). The fish is slender and scaleless, with a gently sloping dorsal profile anterior of the dorsal fin and deeply forked tail. They prefer a substrate of sand and gravel and usually dwell at the bottom of the water. The fish grow efficiently at 80 to 85°F in water. Growth is limited when water temperature is less than 45°F or greater than 95°F (Morris, 1993). At lower temperature, since the metabolic rate is reduced, they eat less and hence grow slowly. If the temperature is too low, the immune system of the fish will be impaired and they are more vulnerable to diseases. On the other hand, at higher temperature, the respiration rate of fish is increased. Because fish need more energy to maintain respiration, feed conversion and hence fish growth is reduced. If the temperature is too high they can die (Jensen, 1997).

2.2 Production Ponds

Catfish can be grown in ponds, cages, and raceways. Pond culture is the most popular method. Other methods require more management efforts and cost more and thus are less common. Based on water supplies and terrain, there are two types of ponds: levee ponds and watershed ponds.

Levee ponds are built on flat land by excavating the pond area to a shallow depth and using the soil obtained to build levees around the perimeter of pond. Levee ponds have the advantages that catfish producers can harvest fish by seine without draining and oxygen can reach all the way to the bottom of the ponds. The disadvantage is that it is more expensive to build. Levee ponds use ground water supplied by wells or surface water such as springs and streams (Beem, 1998).

Watershed ponds are constructed by building dams across ravines or valleys. They are less costly to build than levee ponds and producers are able to make use of steeper sites. They can also serve as reservoirs and help to reduce land erosion. One disadvantage is that ponds cannot be refilled at will because they depend on rainfall for water supply. In addition, watershed ponds tend to stratify which can result in a phenomenon called turn-over that may cause oxygen depletion. Runoff from rainfall is the main source of water for watershed ponds because the rainwater can be stored behind the dams built across valleys (Whitis, 2002).

Levee ponds are usually built on flat land like the delta areas of western Mississippi, southeastern Arkansas or northeastern Louisiana, while watershed ponds are more common in hilly regions of Mississippi, Tennessee, Alabama, Georgia and Illinois

(Whitis, 2002). Levee ponds account for 91.2% of the total acreage of catfish production, while watershed ponds account for the remaining 8.8% (Avery, 2002).

2.3 Production Process

Information contained in this section depends heavily on publications by Robinson and Avery (2000), Morris (1993) and Lewis (1994). The production cycle of catfish can be divided into three phases: egg and fry production, fingerling production, and food fish production.

2.3.1 Egg and Fry Production

The production process begins with careful selection and mating of quality brood stocks to produce eggs. The brood stocks are placed into ponds for free spawning. The spawning season is usually in the spring when the water temperatures reach about 68° F. This is generally around May in southeastern United States. Spawning containers are placed in 2 to 3 feet of water, 1 to 10 yards apart to serve as nesting sites. Female brood fish lay eggs in the containers and the male fertilize the eggs. After the eggs are fertilized, depending on the preference of the producer, fertilized eggs may be left in the containers for parental (male) hatching in the pond, or transferred to the hatchery. The first method is cheaper but unreliable because the number of fry successfully hatched is not immediately known and the survival rate is usually low. The most efficient way is to hatch catfish eggs in a hatchery. The hatchery provides a controlled environment with good water agitation and adequate quality. Troughs made of wood, fiberglass or metal

are used to incubate the eggs. Paddles or aerators are used to circulate water in and around the egg masses in order to provide adequate oxygenation. Depending on water temperature, channel catfish eggs generally remain in the hatchery for 5 to 10 days. Hatching begins when water temperature is around 75 to 85° F.

Immediately after the hatching, the young fish are called “sac fry” because they have a yolk sac attached to their abdomens that serves as a nutrition source. After a couple of days, the yolk sac is used up and the fry turn black and swim up to the surface in the hatching troughs with their mouths open and heads moving back and forth, searching for food. Feeding begins at this stage. Depending on the preference of the producer, the fry may be left in the hatchery troughs for a few days and fed a 50% protein diet, or moved to nursery ponds. The nursery pond method is commonly practiced. The size of fingerlings desired at harvest determines the stocking rate of fry. For example, stocking 10,000 fry per acre will produce fingerlings of 7 to 10 inches in about 150 days, while stocking 100,000 fry per acre will yield fingerlings of 3 to 5 inches in about 150 days.

2.3.2 Fingerling Production

The fry remain in schools after placed in the nursery ponds, where natural foods in the pond are their main source of nourishment. The nursery ponds usually have been fertilized 2 to 3 weeks before being stocked. The fertilization is necessary because it can produce a large number of zooplankton (small animals) for the young fish as food. The young fish are also fed a high-protein, powdered feed. But these feeds serve more as a

fertilizer for pond zooplankton than as a direct source of food for the fry. In 2 to 3 weeks after stocking, the fry swim up to the water surface and daily feeding of fry begins. They are fed food pellets that consist of 35 to 40% protein. These feeds are also made up of a mixture of soybeans, corn, wheat, vitamins and minerals, which not only help in producing healthier fish, but also cleaner, milder tastes. When the size of the fry reaches 2 to 3 inches, they are commonly called fingerlings. Fry stocked in the summer can grow to 5 to 6 inches long by late fall or early winter. They are then transferred to grow-out ponds.

2.3.3 Food Fish Production

The number of fingerlings to stock into a grow-out pond depends on several factors such as the surface acreage of the pond and management ability of the producers. For experienced producers, the stocking rate can be 5,000 to over 10,000 fingerlings per acre. The fingerlings are fed a high-quality 28 to 32% protein once a day for 150 to 180 days before harvest. The best feeding period in the southeastern United States is from May to October. When the fish are about 18 months old and averaging 1 to 1.5 pounds in weight, they can be harvested for processing.

Once a crop of fish reaches the proper size to be harvested, two types of cropping systems can be considered. The first type of cropping system only harvests fish of suitable size for processing and lets smaller fish remain in the pond. This is done 2 to 3 times a year. Once the ponds are partially harvested, new fingerlings of equal number are restocked into the ponds to replace those that were removed. This cycle of incomplete

harvest and restocking is repeated for a few years without draining the ponds. Therefore, several different year-classes of fish are in the same pond at any given time. That's why this method is sometimes called multiple batch. This is a more popular production system. The second type of cropping system involves removing all fish from the pond, draining and refilling the ponds for restocking. Because only a single crop of fish are involved in the production cycle, this method is sometimes called single batch. The annual draining required to remove all fish can significantly increase production cost, so this method is used less extensively. Catfish are harvested in large seines and then transported to processing plants alive, in which they are made into fillets, steaks, nuggets and whole-gutted fish.

2.4 Associated Production Risks

This section addresses the primary production risks associated with commercial catfish production. The risk factors listed here follows from an invited presentation at the 2002 National Risk Management for Aquaculture Workshop (Avery, 2002). Information presented here is drawn largely from the Southern Regional Aquaculture Center, Mississippi State University Cooperative Extension Service, other extension services and organizations.

2.4.1 Infectious Diseases

Infectious disease is one of the major perils facing catfish aquaculture production. Ever since the start of commercial catfish production, diseases have caused significant

economic losses and affect the profitability of the catfish industry. In the last decade, as the culture practices have become more intensive, previously rare diseases have spread among the catfish population, making the problem of diseases even more severe. It is estimated that disease related losses account for approximately 45-50% of all losses incurred on farms annually and may account for as much as $100 million annually in direct economic losses (USDA 2003). The normal mortality rate is 20-35% for fry to 5-inch fingerlings, and 18-24% for larger fish (Avery, 2002).

Stress plays an important role in channel catfish diseases. Stress usually predisposes catfish to diseases. Common stress conditions include: rough handling, drastic water temperature fluctuations, low dissolved oxygen and other poor water quality problems, insufficient nutrition, and overcrowding. Measures to minimize stress as much as possible can reduce the severity, frequency and duration of disease outbreaks (Lewis et al., 1994).

Infectious diseases are mainly caused by bacteria, parasites and viruses. The following sections will discuss some of the major diseases in channel catfish production.

2.4.1.1 Enteric Septicemia of Catfish (ESC)

Information contained in this section is based on SRAC Publication No. 477 (Hawke et al., 1998) and the Catfish 2003 info sheet. Enteric septicemia of catfish (ESC) is caused by the gram-negative bacterium *Edwardsiella ictaluri*. It is one of the most important infectious bacterial diseases of farm-raised channel catfish. Approximately 30% of all disease cases submitted to fish diagnostic laboratories in the southeastern United States are ESC. In Mississippi, it has been reported at frequencies as high as 47%

of the yearly total and causes millions of dollars of economic losses to the catfish industry yearly. Outbreaks of ESC typically occur in the spring and fall when the water temperatures are warm (68 to 82°F). Fish are more susceptible to ESC when they are under stress. The transmissions of ESC can occur in three ways. First, the transmission can be from fish to fish through water contamination with bacteria shed in feces, or by cannibalism of dead or infected fish. Second, birds that pick up infected fish and drop them into another pond can spread the disease to another pond. Third, wet nets and equipments can transfer the disease pond to pond.

2.4.1.2 Columnaris

Information contained in this section is based on SRAC Publication No. 479 (Durborow et al., 1998). Columnaris is caused by a bacterium called *Flavobacterium columnare*. It is the second leading disease that causes fish deaths in the southeastern United States. Channel catfish are susceptible to this disease when they are under some type of environmental stress and when the water temperatures are in the range of 25 to 32°C (77 to 90° F) in the spring, summer and fall. Affected fish usually have brown to yellowish-brown lesions on their gills, skin and / or fins. The gill function is disrupted because the bacteria attach and spread over the gill and finally cover individual gill filaments, resulting in cell death. Also, the bacteria can produce enzymes that erode portions of the gills. Damages in the skin and fins may result in essential salts and fluids release. The effect of columnaris is more devastating because it may expose the fish to secondary infection or other diseases. For example, columnaris is often followed by

winter saprolegniosis (another common bacterial disease that will be discussed later). In a case study, 80% of catfish ponds infected with winter saprolegniosis also experienced columnaris outbreaks in the preceding summer or fall.

2.4.1.3 Proliferative Gill Disease (Hamburger Gill Disease)

Information contained in this section is based on SRAC Publication No. 475 (Mitchell et al., 1998). Proliferative gill disease is a common disease in farm-raised channel catfish caused by a myxosporean parasite called *Aurantiactinomyxon sp*. This parasite causes severe damage to the gills. The gills of infected fish swell and appear mottled red and white like raw hamburger meat. Therefore, PGD is sometimes referred to hamburger gill disease. The affected gills cannot remove sufficient oxygen from the water, causing catfish to suffocate and die, even when the level of dissolved oxygen is high enough. The most severe outbreaks of PGD are observed in the spring, but it can also occur in the fall and winter. The effect of PGD varies. It can kill a few dozen fish over a couple of days, or up to 100% of the fish in less than 3 days. Newly stocked fish are extremely vulnerable to this disease and account for the majority of losses associated with PGD.

2.4.1.4 Winter Fungus (Saprolegnia)

Information contained in this section is based on SRAC Publication No. 4700 (Durborow et al., 2003) and Mississippi State University Extension Service Information

Sheet 1392. Many fungi cause diseases that can infect and kill channel catfish. The causative agent mostly belongs to the family *Saprolegniaceae*, so fungal diseases in channel catfish are commonly called saprolegniasis. Winter fungus, also called winter kill, is the most common and economically important fungal disease of farm-raised channel catfish. It usually occurs among fish of harvestable size in colder months between October and March. Two factors lead to the occurrence of Winter Kill: a rapid decrease in the water temperature and a large number of motile fungal spores in the water. During colder months, if fish are unable to adapt to sudden fluctuation in the water temperature, rapid drops in the water temperature can impair the fish's immune system, resulting in a loss of mucus from the skin and temporary suppression of mucus production. Mucus protects the skin of fish from the contact and infection of fungal spores. Without mucus, fugal spores can penetrate and damage the skin and muscles of fish, causing fish death. The severity of winter fungus is variable and usually results in chronic and smaller losses. However, high mortalities and significant losses have been observed.

2.4.1.5 Ich (White Spot Disease)

Information contained in this section is based on SRAC Publication No. 476 (Durborow et al., 1998). Ich is a common name for the protozoan parasite *Ichthyophthirus multifiliis* and the disease that it causes. This parasite can kill a large number of fish in a short time. Ich is usually transmitted into a pond by a carrier fish, other animals, or man. It can also come from a river or stream that are used as a water

source for the pond. In the pond, this parasite goes through three stages of its life cycle (tomont, theront, trophont) and survives in a fish host, where they feed and mature. In advanced stages of infection, Ich is found under the mucus and epithelium (top layer of cells) in the fish's gills or skin. Ich cells are about the size of salt granules (1 mm or 1/32 inch across). Infected fish may have small white spots on their skin as if they were sprinkled with salt. That's why Ich is sometimes called white spot disease. Ich can also cause the fish to slough off large amounts of mucus on their skin producing a stucco-like appearance. This parasite causes fish kills in three ways. First, respiration of fish is hindered. The thickening of the epithelium as a reaction to the Ich invasion, the deformation of the lamellae (the respiratory folds of the gills), and the Ich organisms covering the gills can reduce oxygen transfer. Second, the epithelial layer of the gill may separate and results in loss of electrolytes, nutrients and fluids. Third, the infection of Ich can cause the fish more susceptible to other diseases. This disease usually occurs in spring and fall, and does not cause problems in warm summer months.

2.4.1.6 Trematode

Information contained in this section is based on SRAC Publication No. 1801 (Terhune et al., 2003) and the Catfish 2003 info sheet. Trematodes are parasites that infest many types of fish and are common in fish ponds frequented by fish-eating birds. Recently, one species of trematodes, *Bolbophorus sp.* has caused significant losses to catfish producers from Louisiana, Mississippi and Arkansas. The life cycle of the *Bolbophorus* is very complex, which involves one final host (the American white pelican)

and two intermediate hosts (the ram's horn snail and catfish). The life cycle begins when the mature trematode lays eggs in the gastrointestinal tract of the American white pelican, which are then released into the pond together with the bird's feces. The eggs hatch and infest the ram's horn snail. Infected snails release larval trematodes, which then infest and encyst in fish. The life cycle is completed when a pelican catches and eats the infected fish. Studies show that the American white pelican and the ram's horn snail are the only final and intermediate hosts for *Bolbophorus*, respectively. Without these hosts, catfish could not be infested with *Bolbophorus*. Transmission of trematodes from fish to fish is not possible. Researches have shown that this trematode causes massive damage to the excretory system such as kidneys and liver of infected catfish. Although mortalities are usually in smaller fish, food-sized fish that survive still suffer from anorexia and poor growth, making them unmarketable. The easiest way to control *Bolbophorus* infection is to reduce the number of ram's horn snail in the ponds and keep fish-eating birds off ponds.

2.4.1.7 Channel Catfish Virus Disease

Information contained in this section is based on SRAC Publication No. 4702 (Camus, 2004). Channel catfish virus disease (CCVD) is the only important viral disease in channel catfish production. It exists in all catfish growing regions of the United States and causes high mortalities in fingerlings and fry. The causal agent is a herpesvirus. This virus causes damages to the kidneys, spleen, liver, intestinal tract, pancreas and brain of the fish, resulting in kidney failure, destruction of blood-forming tissues and

hemorrhage. Although this disease primarily affects fry or fingerlings less than one year old and less than six inches long, larger fish such as the brood stock, if infected, may transmit the disease to fry via eggs or semen. In addition to this vertical transmission, this virus can be transmitted horizontally, i.e., from fish to fish via the water or by direct contact. The outbreaks of CCVD usually occur in the summer. The overall impact of CCVD on the catfish industry is small, which accounts for only 1 to 2% of the total disease losses, collectively. However, the effects on individual producers can be significant. In some production units, the mortalities even approach 100%.

2.4.1.8 Channel Catfish Anemia

Information contained in this section is based on the annual report of the Animal & Dairy Science Department, University of Georgia (Burtle, 1997) and the catfish genetics research annual report (USDA and ARS 2003). Channel catfish anemia (CCV) is a disease that causes mortalities in market size catfish, characterized by severe and acute anemia. It has been reported across the southeastern U. S. since about 1981. During the disease outbreak, fish swim to the pond banks and begin to die. In affected fish, red blood cells account for only 1 to 5% of the blood. Since the red blood cells take the role of carrying oxygen from the gills to other important tissues and organs of the fish, scarcity of red blood cells would impair the oxygen carrying capacity and therefore hinder the normal functioning of other tissues. Therefore, affected fish may have light pink gills, pink or white internal organs and white mouth. That's why sometimes this disease is called the white lip disease. Preliminary findings show that the cause of this

disease may be an interruption of the normal maturation sequence of red blood cell precursors, but a complete characterization of the disease is not yet available. Other factors that may increase the number of fish kills from anemia are low levels of dissolved oxygen and rough handling that stresses the fish.

2.4.1.9 Visceral Toxicosis of Catfish

Visceral toxicosis of catfish (VTC) is a newly recognized problem. It was first noticed in Mississippi and Arkansas during the early spring of 1998 (APHIS 2000). At that time, a few channel catfish producers reported catastrophic mortality events in their brood or harvest-size fish. Most of the infected fish did not have gross external lesions, but necropsies revealed extensive visceral lesions such as congested spleens, pale proximal intestines where the blood vessels were prominent, multiple intestinal intussusceptions, fat effusion, and a reticular pattern to the liver due to vascular congestions (Khoo et al., 2003). This disease affected food fish and brood fish every spring and fall since 1998, causing extremely heavy losses to channel catfish producers. Experiments have suggested that the cause of death by this disease might be some kind of toxin.

2.4.2 Water Quality Related Risks

Poor water quality is one of the most serious threats to catfish production. Failure to maintain good water quality may result in massive losses. The primary water quality

concerns in channel catfish production include: low dissolved oxygen, nitrite toxicosis and toxic algae.

2.4.2.1 Low Dissolved Oxygen

Jensen (1997) states that low dissolved oxygen is by far the most common water-quality related problem in catfish production ponds. Oxygen is necessary for the survival of catfish. Oxygen concentrations should be maintained above 4 ppm (parts per million) for catfish to grow well. Chronically low oxygen not only reduces catfish growth, but also causes stress in catfish and lowers their resistance to diseases.

Aquatic plants such as algae produce oxygen during the daylight hours as a by-product of photosynthesis. This is the main source of oxygen in ponds. Under normal situations, photosynthesis produces adequate oxygen for respiration of aquatic animals, plants and the decomposition of wastes by bacteria. But when oxygen demand exceeds supply, oxygen depletion in a pond occurs. Excessive demands for oxygen usually occur when there are very dense algae blooms that require oxygen for respiration, especially at nighttime, and decomposition activities from algae bloom die-offs. Excessive demand may also result from a phenomenon called turn-over which is related to weather changes such as rain, wind and cold, and causes the algae to die off and oxygen to be removed rapidly through bacterial decompositions. Furthermore, reduced sunlight and rapid reduction in algae population from die-offs will inhibit oxygen production from photosynthesis. Lack of agitation from wind will reduce the amount of oxygen dissolved in the ponds. All these situations can reduce the supply of oxygen and result in low levels of dissolved oxygen.

Because warm water does not contain as much oxygen as cold water, most low-oxygen problems occur between May and September when temperatures are high. Thus, during warm weather months, it is more important to monitor the level of dissolved oxygen in the ponds.

2.4.2.2 Nitrite Toxicosis

This section depends largely on Jensen (1997) and Mississippi State University Extension Service Information Sheet 1390. Catfish, like other animals, produce nitrogenous wastes from the digestion of the protein feeds. Ammonia is the principal nitrogen waste product. Ammonia is also produced from bacterial decomposition of uneaten feed and dead animal or plant, including algae. Ammonia, although toxic to fish, are nutrient source for algae and certain aerobic (oxygen-requiring) bacteria. These bacteria use ammonia in a process called nitrification, during which ammonia is decomposed into nitrite and nitrate. Nitrite is toxic to fish while nitrate is not. Under normal conditions, nitrite can be converted to nontoxic nitrate, thus nitrite does not accumulate to toxic levels. But if the bacterial decomposition (nitrification) is disrupted, nitrite can build up and reach toxic levels. Nitrite enters the bloodstream through the gills and attaches to hemoglobin of the blood, forming methaemoglobin which turns the blood to chocolate-brown color. That's why nitrite toxicosis is also called brown blood disease. Methoemoglobin cannot carry oxygen through the bloodstream. Affected fish may suffocate and die, while fish that survive are susceptible to other stress related diseases.

Most nitrite problems occur during fall and spring when fluctuating temperatures may disrupt the bacterial decomposition.

2.4.2.3 Toxic Algae

Information contained in this section is based on SRAC Publication No. 466 (Brunson et al., 1994). The two most common types of algae in catfish ponds belong to the green and blue-green families. Algae in catfish ponds play a twofold role. On one hand, algae produce most of the oxygen in the pond through photosynthesis and assimilate much of the ammonia, thus helping to maintain the oxygen level and alleviate the problem of nitrite toxicosis in the water. On the other hand, algae may create a severe problem for catfish. As mentioned before, algae blooms die-off can result in low dissolved oxygen. Also, toxin-producing species of blue-green algae are common in catfish ponds that might result in fish kills. These blue-green algae are often related to the problem of off-flavor, which will be discussed below.

2.4.3 Off-flavor

Information contained in this section is based on Jensen (1997), Catfish 2003 info sheet, Shaik (2001) and Avery (2002). Catfish producers are often faced with a problem called off-flavor that causes undesirable tastes in the fish's flesh. The flavor may be so intense that it makes the fish unmarketable. Research has indicated that chemical compounds produced by blue-green algae are the cause of most common off-flavors. These compounds are 2-methylisoborneol (MIB) and geosmin. When catfish absorb

these compounds, MIB causes a musty off-flavor and geosmin causes a muddy off-flavor. Because blue-green algae are most abundant in the summer and fall, off-flavor during this time is most severe, with 50-75% of ponds experiencing this problem. Depending on temperature and other weather conditions, the duration of off-flavor varies from 2 weeks to over 6 months (Avery, 2002). The economic consequences of off-flavor are significant. First, it delays the harvest time of market-sized fish and increases production cost. Fish have to remain in the pond longer until the off-flavor goes off, which requires extra feeds and management efforts, thus increases the overall cost of producing them. The extension in production may increase risk of losses due to diseases and other problems. Furthermore, off-flavor delays sales of fish, which may prevent the producer from selling fish at a high price, resulting in fewer total sales revenue during a given year. Finally, additional economic losses may arise from delays in stocking the next crop of catfish. At the producer level, off-flavors increase catfish production costs by approximately $15 to $23 million annually (Catfish 2003), or 4 to 7 cents/pound (Avery, 2002).

2.4.4 Bird Predation

With the growth in catfish production in recent years, piscivorous birds have become an increasingly serious problem. Bird predation of catfish, especially that of double-crested cormorants and American white pelicans, causes millions of dollars losses to catfish producers. Studies in the National Wildlife Research Center (NWRC) showed that captive cormorants consumed 7-9 catfish/bird/day, which resulted in a 30% reduction

in fish abundance and a 23% loss in biomass in 2 ponds, with an annual impact of near $5 million to Mississippi Delta region alone. NWRC research also indicated that 99.6% of the diet of pelicans collected in northwestern Mississippi is catfish, and pelicans consume up to 3 pounds of catfish per foraging session. Great egrets and blue herons are also predators of catfish. Great egrets usually eat smaller catfish, while great blue herons have greatest impact on fish that are near the surface of the water. In addition to eating fish, birds would wound fish that they don't eat, resulting in potential loss. Furthermore, these birds can serve as hosts of tramotode infections and spread diseases from pond to pond. Because most birds that cause problems are protected under the federal Migratory Bird Treaty Act (MBTA), the most common control measures are harassment techniques to frighten birds away from ponds (Jensen, 1997).

2.4.5 Extreme Weather Conditions

Extreme weather conditions can pose risks to catfish farms. Examples are extreme hear or cold, flooding, ice-over and drought (Avery, 2002). Temperature, an important factor in catfish production, depends heavily on weather condition. Therefore, extreme heat or cold will affect the growth of catfish severely. Ice can block the water supply. Flooding not only causes physical damage to the farm structures, but may also change the quality of water, especially if the floodwater contains pesticide residues from nearby agricultural farms, the consequence is even more serious. Drought reduces the water supply of the ponds due to high evaporation and no water replacement. Fish and other aquatic plants depend on water for oxygen. As the water supply declines, the

oxygen it can carry decreases and finally the fish may die. Catfish producers should be alert to any of the extreme weather conditions.

2.4.6 Management Error

Management errors such as human error, mechanical failures and power outages are also causes of fish kills. For example, pond-specific or farm-wide power loss in the summer time can lead to aerator shutdown and oxygen depletion.

2.4.7 Theft and Vandalism

Theft and vandalism normally do not cause large-scale problems in catfish farms. This is because of the good security procedures that catfish farmers put in place and the difficulty of accessing fish farm facilities without detection.

2.5 Environmental Issue

The remarkable growth of aquaculture has given rise to growing concerns about its impact on the environment. According to Boyd et al. (2000), recent environmental concerns include wetland destruction, conversion of agricultural land to ponds, water pollution, loss of biodiversity, competition for water use, use of toxic or bioaccumulative chemicals and negative social impacts. These concerns are mainly targeted at marine shrimp farming and cage culture of salmon. The main environmental concern about channel catfish production is water pollution caused by the discharge of pond effluents; other environmental concerns seem less problematic. Aquaculture effluents, including

catfish farming effluents, are regulated under the National Pollutant Discharge Elimination System (NPDES) of the Clean Water Act. This Act designates the U.S. Environmental Protection Agency to administer and enforce the NPDES. In addition to the federal regulation, states can develop and carry out their own programs (Tucker, 1999). With more and more states considering developing regulatory procedures, a lot of research has been done to evaluate the impact of pond effluents on the environment.

In a document prepared for the United States Environmental Protection Agency (EPA), Tucker and Hargreaves (1998) stated that, the current production practices enable catfish farming to have minimal impact on the environment. First, because EPA regulates the use of chemicals in catfish pond, catfish pond effluents contain environmentally insignificant amounts of pesticides and therapeutants. Second, the concentrations of nutrients and organic substances are reduced by the longer hydraulic residence time in catfish ponds. Therefore, any water that is discharged finally is diluted and has less environmental impact. Third, the volumes of water that can be discharged from catfish ponds are low, due to some water conservation measures such as reusing water for multiple fish crops before it is discharged and managing the pond water levels to capture the most rainfall. Finally, most water discharged from catfish ponds occurs during the winter and spring periods. During that time, the high precipitation can greatly dilute any water discharged, thus the effluent water quality is at its seasonal best. Research done by Boyd et al. (2000) also showed that, channel catfish farming is not harmful to the environment and is conservative of water, land, feeds, energy and other resources. Tucker (1999) summarizes some management practices to reduce the impact

of pond effluents on the environment, such as using effluents for irrigation of soybeans, treating pond effluents using constructed wetlands and grass filter strips.

2.6 Modeling Risks

We will consider twenty major risk factors associated with the production of catfish, which have been discussed in previous sections. The effects of the risks are represented in the realized yields. We can model the risks by assuming that risks affect yields in four aspects: frequency; severity; persistence across time; and persistence across space. Frequency is defined as how often (in terms of probability) a particular risk occurs in a given production cycle, while severity is defined as how much loss (in percentage of yields) is caused by a particular risk in a given production cycle. Persistence across time is the temporal correlation of a particular risk, which measures the relationship of risk occurring events between two production cycles. Persistence across space is the spatial correlation of a particular risk, which measures the relationship of risk occurring events between two experimental units. The definition of experimental units is very broad. They can be catfish ponds or farms; they can be in the same county or in different counties in the same state, or even in different states. But we are assuming they are equally spaced and in an order such that the second experimental unit is closest to the first experimental unit and the last experimental unit is the farthest.

Unfortunately, precise information about these four aspects of each risk factor is not available. However, four aquaculture specialists provided their opinions on the characteristics of the risks, which helped to build a very general structure to our

simulations.[1] A summary of their opinions is illustrated in Table 2.1. Although the information they provided is subjective, it at least provides a starting point. A formal survey will be sent out to the industry at the end of 2004. The result of the survey will provide more accurate and precise information about the characteristics of the risks in catfish production, and thus will improve the applicability of our simulation study, as the parameters are refined.

Each of these four characteristics is categorized into four levels: high, medium, low and negligible. The levels are quantified as follows. High frequency means the probability of a particular risk occurring in a given production cycle is assumed to be between 0.4 and 0.5. Medium frequency assumes a probability between 0.2 and 0.3. Low frequency assumes the probability is less than 0.1, and negligible frequency assumes a probability of less than 0.01. High severity assumes a 20% to 30% loss in yields. For example, if the riskless yield is 100 pounds/acre, then the realized yield is only 70 to 80 pounds/acre when the risk factor occurs. Medium severity means a 10% to 15% loss, while low severity represents loss of less than 5%, with negligible severity meaning a less than 1% loss. Both the temporal and spatial correlation coefficients for high, medium, low and negligible levels are assumed to be 0.25, 0.1, 0.01 and 0 respectively. The quantifications of the characteristics of risks are summarized in Table 2.2.

Temporal and spatial correlations are modeled together using correlation matrices, which have blocks that represent spatial correlations of different experimental units and

[1] They are: Jimmy L. Avery, extension professor and extension aquaculture leader, James A. Steeby, assistant extension professor and extension aquaculture specialist, both from Mississippi State University Extension Service; Kevin M. Fitzsimmons, associate professor, Dept. of soil, water, environmental science, University of Arizona; and another epidemiologist from the industry.

temporal correlations of different time periods (or production cycles) for the same experimental unit. Specifically, let ρ denote spatial correlation coefficient and α the temporal correlation coefficient. If d_{ti} is the risk number representing a particular risk factor at time *t* for experimental unit *i*, where $t = 1, 2, \ldots, T$, $i = 1, 2, \ldots, N$, then, for a series of temporally and spatially correlated risk numbers $d_{11}, d_{12}, \ldots, d_{1N}$, $d_{21}, d_{22}, \ldots, d_{2N}, \ldots, d_{T1}, d_{T2}, \ldots, d_{TN}$, the correlation matrix is:

$$\begin{pmatrix} \Omega & \omega_1 & \omega_2 & \cdots & \omega_{T-1} \\ \omega_1 & \Omega & \omega_1 & \cdots & \omega_{T-2} \\ \omega_2 & \omega_1 & \Omega & \cdots & \omega_{T-3} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \omega_{T-1} & \omega_{T-2} & \omega_{T-3} & \cdots & \Omega \end{pmatrix}_{NT \times NT}$$

where Ω is a $N \times N$ matrix

$$\Omega = \begin{pmatrix} 1 & \rho & \rho^2 & \cdots & \rho^{N-1} \\ \rho & 1 & \rho & \cdots & \rho^{N-2} \\ \rho^2 & \rho & 1 & \cdots & \rho^{N-3} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \rho^{N-1} & \rho^{N-2} & \rho^{N-3} & \cdots & 1 \end{pmatrix}_{N \times N}$$

and ω_j (*j=1,…,T-1*) is also a $N \times N$ matrix

$$\omega_j = \begin{pmatrix} \alpha^j & 0 & 0 & \cdots & 0 \\ 0 & \alpha^j & 0 & \cdots & 0 \\ 0 & 0 & \alpha^j & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & \alpha^j \end{pmatrix}_{N \times N}$$

One can see that Ω represents the spatial correlation among the *N* experimental units at a given time period and ω_j represents the temporal correlation of an

experimental unit at different time periods. Note that we are assuming there is no serial correlation among different experimental units. That is, the correlation between d_{ti} and d_{sj} $(t \neq s, i \neq j)$ is zero, resulting in off-diagonal elements of ω_j equal to zeros. For example, let *N*=2, *T*=3, and assume high spatial correlation ($\rho = 0.5$) and medium temporal correlation ($\alpha = 0.1$), then the correlation matrix would be:

$$\begin{pmatrix} 1 & 0.25 & 0.1 & 0 & 0.1^2 & 0 \\ 0.25 & 1 & 0 & 0.1 & 0 & 0.1^2 \\ 0.1 & 0 & 1 & 0.25 & 0.1 & 0 \\ 0 & 0.1 & 0.25 & 1 & 0 & 0.1 \\ 0.1^2 & 0 & 0.1 & 0 & 1 & 0.25 \\ 0 & 0.1^2 & 0 & 0.1 & 0.25 & 1 \end{pmatrix}_{6\times 6}$$

Once the correlation matrix of a particular risk factor is constructed, a series of temporally and spatially correlated risk numbers (for different experimental units and time periods) can be generated using the Cholesky decomposition of the correlation matrix. To be more specific, let **r** ~ multivariate standard normal (**0**, I_n) where **r** is a $n\times 1$ vector and I_n is an *n*-dimensional identity matrix. Let **A** be the correlation matrix such that $\mathbf{A}=\mathbf{L}^{\mathrm{T}}\mathbf{L}$ where **L** is an upper triangular matrix. Therefore, $\mathbf{L}^{\mathrm{T}}\mathbf{r}$ is a vector of multivariate normal random numbers with mean zero and correlation matrix equal to A. A matrix must be symmetric and positive-definite to be decomposed. Although our correlation matrix is symmetric by construction, sometimes it may not be positive-definite, depending on the temporal and spatial correlation coefficients. Fortunately, negative-definite correlation matrices do not occur in the simulation. Therefore, we

reserve the discussion of the Cholesky decomposition of non-positive-definite matrix in Appendix A.

Once a series of temporally and spatially correlated risk numbers are generated, we can model the frequency of the risk factor. To model the probability that a particular risk factor occurs, a uniformly distributed random number is required. Given that the value of cumulative distribution function (CDF) is uniformly distributed between 0 and 1, we can take the CDF of those multivariate normal random numbers and generate some probability numbers. If the probability number is less than the quantification level of frequency, it means the risk factor occurs. The severity of the risks is modeled using a scaling factor on the yields. If the risk factor occurs, then a scaling factor is defined according to the quantification level of severity. Otherwise, the scaling factor will be 1. For example, if the severity level is medium, then the scaling factor will be between 85% and 90%. That means, if the risk factor occurs, the realized yield is only 85-90% of the riskless yield.

The same modeling procedure is implemented for the twenty major risk factors and twenty scaling factors come up. Assuming the effects of the twenty risks on the yields are independent, the final simulated yield will be the average yield times all these scaling factors.

Table 2.1: Summary of the characteristics for each risk

	Risk items	Frequency	Severity	Temporal Correlation	Spatial Correlation
1	Enteric septicemia of catfish	medium	medium	low – medium	negligible – low
2	Columnaris	medium – high	medium	medium – high	low
3	Proliferative gill disease	low	medium – high	low	medium
4	Winter fungus	low – medium	medium	low – medium	low – medium
5	Channel catfish virus	low	low	low – medium	negligible – low
6	Channel catfish anemia	low – medium	medium – high	negligible – low	low – medium
7	Ich	negligible – low	negligible – low	negligible – low	negligible – low
8	Trematode	low	low – medium	medium	low – medium
9	Visceral toxicosis of catfish	low	medium – high	medium – high	low – medium
10	Nitrite Toxicosis	low – medium	low – medium	negligible – low	negligible – low
11	Toxic Algae	low	medium	negligible – low	low
12	Low Dissolved Oxygen	medium – high	low – medium	medium – high	negligible – low
13	Off-flavor	medium – low	medium	medium	medium
14	Predators	medium – high	low – medium	medium – high	medium – high
15	Ice-over	negligible – low	negligible – low	negligible – low	low
16	Flooding	negligible	negligible	negligible	negligible – low
17	Extreme heat / cold	low – medium	low – medium	medium	medium
18	Drought	low	low – medium	low	medium – high
19	Management Error	medium	high	negligible – low	low
20	Theft and vandalism	negligible – low	negligible	negligible – low	negligible – low

Table 2.2: Summary of the quantification of risks for the base scenario

Characteristics of risk	High	Medium	Low	Negligible
Frequency (probability)	0.4-0.5	0.2-0.3	<0.1	<0.01
Severity (scaling factor)	0.2-0.3	0.1-0.15	<0.05	<0.01
Persistence Across Time (correlation)	0.25	0.1	0.01	0
Persistence Across Space (correlation)	0.25	0.1	0.01	0

3. RATING METHODOLOGIES

Accurate premium rates are essential to the actuarial soundness of the catfish aquaculture insurance program. Determination of accurate premium rates requires precise measurement of yield risks, which in turn depends on appropriate estimation of the distribution of yields. In the crop insurance literature, there are various approaches to modeling conditional yield distributions and hence, constructing premium rates. These various approaches can be segmented into two primary groups, depending on whether they use a known parametric distribution or nonparametric techniques. Most yield distribution models are of a parametric nature. Under this approach, a specific parametric distribution could be selected a priori and parameters of the distribution are estimated using observed yield data. Conventional approaches to estimating conditional yield distribution and rating crop insurance contracts have typically used the normal distribution (Botts and Boles, 1957). Gallagher (1987) used a gamma distribution function in attempts to capture the asymmetry and negative skewness of soybean yields. Nelson (1990) confirmed negative skewness in county mean yield distributions for corn and thus used a beta distribution. These approaches have the limitation that they relied on a priori specification. If the distributional priori specification is incorrect, it could lead to inaccurate predictions and misleading inferences. Therefore, a variety of nonparametric approaches to estimating yield distributions have been developed to overcome some of the problems associated with the parametric approaches. Goodwin and Ker (1998), Ker and Coble (1997,1998), and Turvey and Zhao (1993) used univariate kernel density estimators to estimate yield densities. Ker and Goodwin (1998) developed

an empirical Bayesian nonparametric kernel density estimator that exploits the similarities among the county yield densities.

Most of the methods in rating crop insurance contracts are also applied to rating catfish aquaculture insurance contracts due to the large similarities of crop insurance and aquaculture insurance. The first section of this chapter discusses the general formula for calculating premium rates and introduces the rating methodologies. The next section presents the parametric approaches and the resulting premium rates. Following this, nonparametric approaches are delineated and their resulting premium rates are reported.

3.1 Premium Rates

Premium rates can be expressed as expected loss as a percentage of total liability (Ker, 1996). For a contract that guarantees $\lambda \times 100\%$ of the predicted yield, say y^e, the actuarially fair premium rate is given as

$$\text{Premium Rate} = \frac{Expected\ Loss}{Total\ Liability}$$

$$= \frac{\text{Prob}(Y < \lambda y^e)E(\lambda y^e - Y \mid y < \lambda y^e)}{\lambda y^e}$$

$$= \frac{\text{Prob}(Y < \lambda y^e)[\lambda y^e - E(Y \mid y < \lambda y^e)]}{\lambda y^e}$$

where $0 \le \lambda \le 1$ and $\lambda y^e - Y$ represents a loss. For example, if λ is 0.8, y^e is 5000 pounds/acre, then the guaranteed level, or total liability would be $0.8 \times 5000 = 4000$ pounds/acre. If a producer's realized yield is less than 4000 pound/acre, then loss happens and the producer will get an indemnity. If the probability

of yield less than 4000 pound/acre is 0.3, and the expected yield given a loss has happened is 3000 pound/acre, then premium rate is $0.3\times(4000-3000)/4000=7.5\%$. Accuracy in premium rates is dependent on accuracy in determining the probability that a loss will occur and the amount of loss that occurs. Both are given by the area under the probability density function between 0 and λy^e. Thus, precise measurement of the yield density is crucial. Both parametric and nonparametric approaches attempt to recover $\hat{f}(y)$, the estimated yield density or curve. In all our simulations, λ is chosen to be 0.75, and y^e is chosen to be the average yield of the first experimental unit.

Recall the main goal of this thesis is to undertake a performance comparison of a number of estimators based on simulated yield data. Twelve rating methodologies are considered here. They are: (1) the empirical rate only for the first experimental unit; (2) the empirical rate for all experimental units; (3) assuming normal distribution only for the first experimental unit; (4) assuming normal distribution for all experimental units; (5) kernel density estimation only for the first experimental unit; (6) kernel density estimation for all experimental units; (7) kernel density estimation for all experimental units with transformation of both location and scale parameters of the first experimental unit; (8) kernel density estimation for all experimental units with transformation of the location parameter of the first experimental unit; (9) Bayesian nonparametric kernel density estimation; (10) assuming beta distribution only for the first experimental unit; (11) assuming beta distribution for all experimental units; (12) nonparametric estimation of similar densities.

In addition to the parametric or nonparametric nature of these methodologies, another distinction is whether they utilize only individual data or pooled data. Pooled data means combining the data from all experimental units.[2] When estimating the densities for an experimental unit of interest, the standard approach is to use only the yield data of that unit. However, lack of historical yield data is one of the most fundamental obstacles in rating catfish aquaculture insurance contracts. Although yield data by county or farm tend to be extremely scarce, the number of counties or farms can be large. In some circumstances, the individual densities are believed to have some structural similarities that can be utilized to produce improved estimates. For example, in the same reporting area, weather patterns, water sources and technology use among other factors could be considered similar across units. These similarities provide a reasonal basis for assuming that the densities of the different units are related even though the magnitude of such relationships is unknown. Therefore, an alternative to the standard approach is to include the yield data from other experimental units in the estimation process for potential efficiency gains. In this thesis, given the relatively short length of time series for each experimental unit in the simulation (only 2 to 50 time periods), in determining the premium rate for the first experimental unit, it seems reasonable to use yield information from other experimental units. Eight out of the twelve methodologies use extraneous data in the estimation process. Extraneous data refers to the data from other experimental units.

[2] In the following discussion, we will use the term “pooled data” quite often. It is defined as the data set that consists of yield data from all experimental units.

3.2 Parametric Approaches

3.2.1 Assuming Normal Distribution for the First Experimental Unit

Under this approach, catfish yields of the first experimental unit are assumed to follow the normal distribution. The mean and variance of the distribution are estimated using the sample moments. Specifically, if $y_{11},...,y_{T1}$, is a sample of yields for the first experimental unit, then the sample mean and variance are given by:

$$\hat{\mu}_1 = \frac{\sum_{t=1}^{T} y_{t1}}{T} \quad \text{and} \quad \hat{\sigma}_1^2 = \frac{\sum_{t=1}^{T} (y_{t1} - \hat{\mu}_1)^2}{T-1}$$

Thus, the yields follow a normal distribution with mean $\hat{\mu}_1$ and variance $\hat{\sigma}_1^2$. The premium rate based on a normal distribution with mean $\hat{\mu}_1$ and variance $\hat{\sigma}_1^2$ is given by:

$$\text{Rate} = \frac{\Phi\left(\frac{\lambda y^e - \hat{\mu}_1}{\hat{\sigma}_1}\right)(\lambda y^e - \hat{\mu}_1) + \phi\left(\frac{\lambda y^e - \hat{\mu}_1}{\hat{\sigma}_1}\right)\hat{\sigma}_1}{\lambda y^e}$$

where Φ is the normal cumulative density function, and ϕ is the normal probability density function. This is derived from the first moment of the truncated normal distribution. If z ~ Normal (μ, σ^2) and δ is a constant, then $E[z \mid z < \delta] = \mu - \sigma\phi(\alpha)/\Phi(\alpha)$ where $\alpha = (\delta - \mu)/\sigma$ (Greene, 2000).

3.2.2 Assuming Normal Distribution for All Experimental Units

Under this approach, the yields of all experimental units are assumed to follow one single normal distribution. The mean and variance of this normal distribution are

estimated as the averages of the sample moments of each experimental unit. If there are *N* experimental units, and $y_{1i},...,y_{Ti}$ (*i*=1 to *N*) is a sample of yields for the i^{th} experimental unit, then its sample mean and variance would be:

$$\hat{\mu}_i = \frac{\sum_{t=1}^{T} y_{ti}}{T} \quad \text{and} \quad \hat{\sigma}_i^2 = \frac{\sum_{t=1}^{T} (y_{ti} - \hat{\mu}_i)^2}{T-1}$$

The mean and variance for the pooled data are given by:

$$\hat{\mu} = \frac{\sum_{i=1}^{N}\sum_{t=1}^{T} y_{ti}}{NT} \quad \text{and} \quad \hat{\sigma}^2 = \frac{\sum_{i=1}^{N}\sum_{t=1}^{T} (y_{ti} - \hat{\mu}_i)^2}{N(T-1)}$$

In this case, the pooled data are assumed to follow a normal distribution with mean $\hat{\mu}$ and variance $\hat{\sigma}^2$. The premium rate based on this distribution is given by:

$$\text{Rate} = \frac{\Phi\left(\frac{\lambda y^e - \hat{\mu}}{\hat{\sigma}}\right)(\lambda y^e - \hat{\mu}) + \phi\left(\frac{\lambda y^e - \hat{\mu}}{\hat{\sigma}}\right)\hat{\sigma}}{\lambda y^e}$$

3.2.3 Assuming Beta Distribution for the First Experimental Unit

The general formula for the probability density function of the beta distribution is

$$f(x) = \frac{(x-a)^{p-1}(b-x)^{q-1}}{B(p,q)(b-a)^{p+q-1}} \qquad a \le x \le b;\ p,q > 0$$

where *p* and *q* are the shape parameters, *a* and *b* are the lower and upper bounds, respectively, of the distribution, and *B*(*p*,*q*) is the beta function. The lower bound of the yields is assumed to be 0. Maximum likelihood estimation is used to recover the other three parameters. Once the parameters are known, the yield densities are known. The premium rate based on the beta distribution is given by:

$$\text{Rate} = \frac{P(Y < \lambda y^e)(\lambda y^e - E(Y \mid y < \lambda y^e))}{\lambda y^e} = \frac{P(Y < \lambda y^e)E(\lambda y^e - Y \mid y < \lambda y^e)}{\lambda y^e}$$

$$= \frac{P(Y < \lambda y^e)\int_0^{\lambda y^e}(\lambda y^e - Y)P(Y \mid y < \lambda y^e)dy}{\lambda y^e}$$

$$= \frac{P(Y < \lambda y^e)\int_0^{\lambda y^e}(\lambda y^e - Y)\frac{f(Y)}{P(Y < \lambda y^e)}dy}{\lambda y^e}$$

$$= \frac{\int_0^{\lambda y^e}(\lambda y^e - Y)f(Y)dy}{\lambda y^e}$$

For beta distribution, there is no closed form for this integral and as such, a numerical approximation is used to recover the estimated premium rate.

3.2.4 Assuming Beta Distribution for All Experimental Units

In this case, the yield data from all experimental units are combined and assumed to follow a single beta distribution. Maximum likelihood estimation is used to recover the four parameters of this beta distribution based on the pooled data. Premium rates are estimated in the same way as in the previous section.

3.3 Nonparametric Approaches

Parametric approaches assume that the distribution of the catfish yields follows a known functional form. Nonparametric density estimation techniques do not assume a particular functional form for the yield distributions. Instead, they allow the data to

"speak for themselves". Thus nonparametric approaches are fully flexible and essentially nest alternative parametric specifications.

3.3.1 Empirical Rate for the First Experimental Unit

The simplest approach to nonparametrically estimating a probability density function is the histogram. The empirical premium rate is analogous to a histogram where no smoothing is undertaken. An empirical rate is simply the average loss realization. It represents the expected loss if the sample size is large enough. If yield data were abundantly available, the empirical rate would recover a reasonably accurate estimate. The empirical rate for experimental unit 1 is given by:

$$\text{Rate} = \frac{\sum_{t=1}^{T} Max(\lambda y^e - y_{t1}, 0)}{T} \Big/ \lambda y^e$$

3.3.2 Empirical Rate for All Experimental Units

When we incorporate the yield data from other experimental units, the premium rate is given by:

$$\text{Rate} = \frac{\sum_{i=1}^{N}\sum_{t=1}^{T} Max(\lambda y^e - y_{ti}, 0)}{NT} \Big/ \lambda y^e$$

3.3.3 Kernel Density Estimation for the First Experimental Unit

One of the limitations of empirical rate is that the density estimate is discontinuous because no smoothing is undertaken. An alternative is to smooth between observations to build a continuous density estimate. Kernel density estimation techniques

offer a consistent approach to smoothing observations to build a continuous density estimate. For greater details about kernel density estimation, please see Appendix B.1. For *T* independent and identically distributed observations of a univariate series of yields from experimental unit 1, $Y_{t1} = (Y_{11},...,Y_{T1})$, the kernel density estimate at support point *y* is defined as:

$$\hat{f}(y) = \frac{1}{Th}\sum_{t=1}^{T} K\left(\frac{y - Y_{t1}}{h}\right)$$

where $K(\cdot)$ is chosen to be a symmetric probability density function centered at zero, and *h* is the smoothing parameter or bandwidth. The kernel density estimator places a bump or individual kernel at each observation. Intuitively, the estimate of the density at any support point is simply the sum of the height of the bumps, or kernels at that particular point. In regions where there are a lot of observations, the estimates will be large because the closeness of the data points raises their weights represented in the kernels, while in regions with few observations, the spread of the data points decreases the weights of the support point, resulting a small density estimate.

$K(\cdot)$ is also called the kernel function. In most cases it is chosen to be the standard normal distribution although a variety of alternatives may be used, such as the Epanechnikov kernel. The individual kernel being a probability density function guarantees that the kernel estimate itself is a density. The standard normal kernel is used throughout this thesis because of ease of use. Choosing the proper smoothing parameter is another important issue in nonparametric kernel density estimation. This parameter determines the weight to assign to neighboring observations in constructing the density

and thus corresponds to the amount of smoothing to be done. A larger bandwidth will assign more weight to neighboring observations and thus will result in a flatter, smoother density function, while a smaller bandwidth will yield a rough and irregular density. A variety of methods are available to estimate the bandwidth, including cross-validation, Silverman's "rule-of-thumb", and other plug-in approaches. For computational consideration, a rule-of thumb approach is used to determine the bandwidth parameter for all kernel methods discussed in this thesis. The rule-of-thumb bandwidth is given by:

$$h = 0.79 * \sigma * T^{-(1/5)}$$

where σ is the standard deviation and T is the number of observation. In the simulations, σ can be estimated by the sample variance. Recall the actuarially fair premium rate is given by:

$$\text{Rate} = \frac{P(Y < \lambda y^e)(\lambda y^e - E(Y \mid y < \lambda y^e))}{\lambda y^e} = \frac{\int_0^{\lambda y^e} (\lambda y^e - Y) f(Y) dY}{\lambda y^e}$$

Once the density estimates for a set of support points in the domain of (0, λy^e) are recovered, numerical integration can be used to recover the estimated rate.

3.3.4 Kernel Density Estimation for All Experimental Units

When yield data from other experimental units are added in the estimation, the pooled data are: $Y_{ti} = (Y_{11},...,Y_{T1},Y_{12},...,Y_{T2},...,Y_{1N},...,Y_{TN})$. The kernel density estimate at support point y is defined as:

$$\hat{f}(y) = \frac{1}{NTh} \sum_{i=1}^{N} \sum_{t=1}^{T} K\left(\frac{y - Y_{ti}}{h} \right)$$

Once the density estimates for a set of support points in the domain of (0, λy^e) are recovered, premium rates are constructed in the same way as in section 3.3.3.

3.3.5 Kernel Density Estimation for All Experimental Units with Transformation of both Location and Scale Parameters of the First Experimental Unit.

This is a slight variation from the method discussed in section 3.3.4. We are trying to use the yield data from other experimental units in estimating the conditional yield density for experimental unit 1, based on assumption of some kind of similarity among the densities of all units. The concept of similarity is loosely used here because the extent to which the set of yield densities is similar is unknown. If the set of curves were indeed similar in shape, estimators that use extraneous data would improve greatly in terms of efficiency. However, when the curves are not similar as assumed, then bias will result. One of the reasons that the density curves are dissimilar is because they come from distributions with different location or scale parameters. Although yields are assumed to be normal, the location and scale parameters of other experimental units' yields may be different from those of the first experimental unit, which may introduce bias. A transformation of the yield data from all experimental units may be invoked such that the set of different densities collapses to a single density.

A very nice property of kernel estimators is that they are invariant to transforming the data. Invariance allows us to estimate the density with the transformed data, and then take the inverse transformation on the estimated density to retrieve the original density. To ensure our density estimates have mean and variance of the first experimental unit,

yields from each experimental unit are first standardized according to its own parameters prior to entering the density estimators, thus having mean zero and variance one. Assuming y_{ti} is the yield of the i^{th} unit at time *t*, the standardization process is given by:

$$z_{ti} = \frac{y_{ti} - \hat{\mu}_i}{\hat{\sigma}_i}$$

where $\hat{\mu}_i$ and $\hat{\sigma}_i$ are defined in section 3.2.2. The standardized yield data of each unit are then combined together and kernel density estimation is undertaken. The kernel density estimate at support point z is defined as:

$$\hat{f}(z) = \frac{1}{NTh} \sum_{i=1}^{N} \sum_{t=1}^{T} K\left(\frac{z - z_{ti}}{h} \right)$$

Finally, the location and scale parameters of the first experimental unit, $\hat{\mu}_1$ and $\hat{\sigma}_1$, are used to transform these estimated densities back. The support point and the corresponding density estimate are given by:

$$y = z * \hat{\sigma}_1 + \hat{\mu}_1 \qquad \text{and} \qquad \hat{f}(y) = \hat{f}(z) / \hat{\sigma}_1$$

Once the supports and densities of yields are available, premium rates are constructed in the same way as in the previous sections.

3.3.6 Kernel Density Estimation for All Experimental Units with Transformation of the Location Parameter of the First Experimental Unit.

This method is slightly different from the previous one in that it just uses the location parameter of the first experimental unit in the transformation. The steps still involve standardization of the pooled data and then kernel density estimation. But in the

transformation step, we use the location parameter of experimental unit 1 ($\hat{\mu}_1$) and the scale parameter of the pooled data ($\hat{\sigma}$, defined in section 3.2.2). The support point and the corresponding density estimate after the transformation are given by:

$$y = z * \hat{\sigma} + \hat{\mu}_1 \qquad \text{and} \qquad \hat{f}(y) = \hat{f}(z) / \hat{\sigma}$$

Premium rates are constructed in the same way as in the previous sections.

3.3.7 Bayesian Nonparametric Kernel Density Estimation

Ker (1998) derived an empirical Bayes nonparametric kernel density estimator that exploits possible similarities among the set of unknown densities that are to be estimated. Rather than placing a prior on a parameter space, the estimator uses empirical Bayes techniques on the estimated densities from the standard kernel-type density estimators discussed in previous sections. As Ker and Ergun (2003) point out, the main strengths of the empirical Bayesian nonparametric kernel density estimator result from using kernel density estimator as the basis. First, since the Bayesian estimator depends on the kernel estimator, all the variations of the kernel estimator, such as higher order kernels, variable kernel methods and transformation-kernel density estimators, are also applied to the Bayesian estimator. Second, when the set of densities is not identical, Ker (1998) shows that the empirical Bayesian estimator converges in probability at a rate of $O_p(T^{-(4/5)})$ to the standard kernel density estimator, which is faster than the rate that the kernel density estimator converges to the unknown density of interest, $O_p(T^{-(2/5)})$. As such, the Bayesian estimator inherits the same asymptotic properties as the standard

kernel density estimator and converges to the unknown density at the optimal rate of $O_p(T^{-(2/5)})$. Finally, the Bayesian estimator does not require any specification as to the degree or form of similarities among the set of densities of the experimental units, which is usually unknown in practice.

Under this approach, not only a single conditional yield density, but the entire set of conditional yield densities, one for each experimental unit, will be considered. In our analysis, we have *N* experimental units with densities { $f_1, f_2, ..., f_N$ } and random samples $Y_{1i}, Y_{2i}, ..., Y_{Ti}$ from f_i for $i = 1,2,...,N$. Denote the kernel density estimate at support point y_j for experimental unit *i* as $\hat{f}_{ij}$. Based on the pointwise limiting distribution of kernel density estimators, Ker (1998) proposed the following hierarchical model:

$$\hat{f}_{ij} \mid \mu_{ij} \sim Normal(\mu_{ij}, \sigma_{ij})$$

$$\mu_{ij} \sim Normal(\mu_j, \tau_j^{\,2})$$

where $\mu_{ij} = f_{ij} + \beta_{ij}$, f_{ij} is the unknown density value for experimental unit *i* at support point y_j, β_{ij} is the bias for experimental unit *i* at support point y_j, $\sigma_{ij}^{\,2}$ is the variance of the kernel density estimate for experimental unit *i* at support point y_j, μ_j is the mean value of the densities across experimental units at support point y_j, and $\tau_j^{\,2}$ is the variance of the densities across experimental units at support point y_j.

The empirical Bayesian nonparametric kernel density estimator at support point y_j for experimental unit *i* is:

$$\tilde{f}_{ij} = \hat{f}_{ij}\left(\frac{\tau_j^{\,2}}{\tau_j^{\,2} + \sigma_{ij}^{\,2}}\right) + \hat{\mu}_j\left(\frac{\sigma_{ij}^{\,2}}{\tau_j^{\,2} + \sigma_{ij}^{\,2}}\right)$$

where the unknown parameters ($\mu_j, \tau_j^2, \sigma_{ij}^2$) must be estimated.

Bootstrapping methods are usually used to estimate the variance σ_{ij}^2, but require a lot of computation time. Therefore, an alternative estimate of σ_{ij}^2 is obtained by the asymptotic variance formula. It is easy to show that (see Appendix B.2):

$$\hat{\sigma}_{ij}^2 = \hat{f}_{ij}/(2\sqrt{\pi}Th)$$

Estimates of μ_j and τ_j^2 are obtained using the following method of moment estimators:

$\hat{\mu}_j = \frac{1}{N}\sum_{i=1}^{N}\hat{f}_{ij}$ and $\hat{\tau}_j^{\,2} = \hat{s}_j^{\,2} - \frac{1}{N}\sum_{i=1}^{N}\sigma_{ij}^{\,2}$ where $\hat{s}_j^{\,2} = \frac{1}{(N-1)}\sum_{i=1}^{N}(\hat{f}_{ij}^{\,2} - \hat{\mu}_j)^2$ (see Appendix B.2). For $\hat{\tau}_j^{\,2}$, we use the positive part estimator.

Intuitively, as the estimated variance of the kernel estimates across experimental units increases ($\hat{\tau}_j^{\,2}$ becomes larger), $\tilde{f}_{ij}$ will shrink less toward the overall mean ($\hat{\mu}_j$). Conversely, the larger the estimated variance of the kernel estimate for a given experimental unit ($\hat{\sigma}_{ij}^{\,2}$), the more $\tilde{f}_{ij}$ will shrink toward the overall mean ($\hat{\mu}_j$). The greater the estimated variance within the experimental units relative to the estimated variance across the experimental units, the greater $\tilde{f}_{ij}$ will shrink toward the overall mean, which implies greater potential efficiency gains. Ker (1998) indicates that the empirical Bayes nonparametric kernel estimator may offer the largest efficiency gains in

small samples where the variance within experimental units tends to be relatively high as compared to the variance across experimental units.

In our analysis, we are interested in estimating the premium rate for the first experimental unit. The empirical Bayesian nonparametric kernel density estimator at support point y_j for the first experimental unit is:

$$\tilde{\hat{f}}_{1j} = \hat{f}_{1j}\left(\frac{\hat{\tau}_j^{\,2}}{\hat{\tau}_j^{\,2} + \hat{\sigma}_{1j}^{\,2}}\right) + \hat{\mu}_j\left(\frac{\hat{\sigma}_{1j}^{\,2}}{\hat{\tau}_j^{\,2} + \hat{\sigma}_{1j}^{\,2}}\right)$$

where $\hat{f}_{1j}$ is the kernel density estimates only for the first experimental unit, which is derived in section 3.3.3. Premium rates are constructed in the same way as in the previous sections. Appendix B.2 contains a more detailed description of this method.

3.3.8 Nonparametric Estimation of Similar Densities

Ker (2002) developed an estimator that offers greater efficiency if the set of densities are similar while not losing much if the set of densities are dissimilar. This method has the same objective as the Bayesian estimator discussed in the previous section in that they are designed to exploit any similarities among the sets of densities. The difference is that this method does not require a hierarchical model and thus the need to know the relationship among the densities is circumvented.

If the densities were identical, one would pool the *N* samples and estimate a single density, as we did in section 3.3.4. However, if the densities are not identical, this estimator is inconsistent. In that case, a nonparametric estimator that combines a kernel

estimate based on the pooled data with a kernel estimate based on the individual data may be considered.

The idea is to pool the data and obtain the kernel density estimates, denoted $\hat{g}(y)$ (we derived this in section 3.3.4), and then multiply an individual correction function $r_i(y) = f_i(y)/\hat{g}(y)$ in order to adjust for individual effects. The correction factor function is itself estimated nonparametrically by:

$$\hat{r}_i(y) = \frac{1}{Th}\sum_{t=1}^{T} K\left(\frac{y - Y_{ti}}{h}\right) / \hat{g}(Y_{ti}).$$

Thus, the estimator of possibly similar densities for the i^{th} experimental unit is given by:

$$\tilde{f}_i(y) = \hat{g}(y)\hat{r}_i(y) = \frac{1}{Th}\sum_{t=1}^{T} K\left(\frac{y - Y_{ti}}{h}\right)\frac{\hat{g}(y)}{\hat{g}(Y_{ti})}$$

where $(Y_{1i}, Y_{2i}, ..., Y_{Ti})$ is the sample observations and y is a support point. Ker (2002) shows that $\tilde{f}_i$ is biased and the leading term of the bias is:

$$E\tilde{f}(y) - f(y) = \frac{1}{2}h^2\mu_2\left(\frac{f}{g}\right)''(y)$$

Let $\hat{f}(y)$ be the standard kernel density estimates based on individual data. Recall the leading term of its bias is:

$$E\hat{f}(y) - f(y) = \frac{1}{2}h^2\mu_2 f''(y)$$

Clearly, one can decrease the bias by reducing the curvature of $f(y)$. It is easy to see that the bias of the estimator $\tilde{f}(y)$ is not a function of the curvature of the unknown true density as it is for the estimator $\hat{f}(y)$. Rather, it is a function of the second derivative of

the correction function $r(y) = f(y)/g(y)$. If the start, $\hat{g}(y)$, is close to the true density, then the correction function $r(y)$ will have less global curvature than that of individual curves. Hence, $\tilde{f}(y)$ may have less bias.

Ker (2002) enumerates the advantages of this estimator. First, it starts from a nonparametric kernel density estimator, which does not assume any functional form and hence avoids any wrong specification about the underlying density. Second, when the sets of densities are similar, it has a lower bias due to the correction factor and more efficiency due to the pooling of data. Also, estimating the correction function nonparmetrically can make it fluctuate around 1 and thus the curvature will be close to zero. As a result, the total curvature that is estimated with the individual sample data may be significantly reduced.

In our analysis, we are interested in the premium rate for the first experimental unit. The estimator at support point *y* is:

$$\tilde{f}_1(y) = \hat{g}(y)\hat{r}_1(y) = \frac{1}{Th}\sum_{t=1}^{T} K\left(\frac{y - Y_{t1}}{h}\right)\frac{\hat{g}(y)}{\hat{g}(Y_{t1})}$$

where $\hat{g}(y)$ is the kernel density estimates based on pooled data, which is derived in section 3.3.4. Premium rates are constructed in the same way as in the previous sections.

4. SIMULATION ANALYSIS AND RESULTS

The main objective of this thesis is to investigate the small sample performances of various methodologies for estimating premium rates. In satisfying this objective, we undertake a performance comparison of different rating methodologies based on simulated yield data. This chapter focuses on simulations and the accompanying results. In section 2.5, we have discussed how the risks are modeled based on assumptions about the frequency and magnitude of each risk factor as well as their temporal and spatial correlations. In the simulations, we also need to make some assumptions, such as the number of experimental units, the quantification of the characteristics of the risk factors and the ranges of the means and variances of the yield distributions. It is important to see how the methodologies will perform under different assumptions: we developed one base scenario and eleven alternative scenarios.

The first section of this chapter presents the yield simulator and the base scenario. Section 2 discusses the relationship between sample size and similarities among experimental units. The designs of the remaining scenarios will be addressed in section 3. The last section presents the accompanying results of the twelve scenarios and the performances of the methodologies are compared. The designs, MSE results, performance comparisons of all twelve methodologies for each scenario can be found in Appendix C, D and E, respectively.

4.1 Yield Simulator and Base Scenario

Under the base scenario, we simulate the yields of 30 experimental units (e.g. farms or ponds) over a 50-year period (or 50 production cycles). Without loss of generality, we focus on estimating the actuarially fair premium rate for the first experimental unit. The yield simulation process is as follows.

(1) *Unconditional* yields are assumed normally distributed. *Unconditional* means that none of the risk factors have been applied.

(2) In order to randomize the average yields of different experimental units, the means of the yields of the 30 experimental units are drawn from U [50000, 100000].[3] These numbers are chosen for the following reason. The pond is our major experimental unit of interest, so the mean yield should be based on the average yield of a pond. Catfish production is about 5,000 pounds/acre.[4] The average size of a pond is 10 to 20 acres (Robinson and Avery, 2000). As such, the average yield of a pond is 50,000 (5,000 times 10) to 100,000 (5,000 times 20) pounds.

(3) The standard deviations of the yields of the 30 experimental units are drawn from U[10000, 15000].

(4) The information provided by the aquaculture specialists about the characteristics of the 20 risk factors is summarized in Table 2.1 and quantified in Table 2.2. The levels of the frequency and severity are operationalized using uniform distributions. We do not assign a fixed number to represent a certain level of frequency or severity of each

[3] U [50000, 100000] refers to the uniform distribution between 50,000 and 100,000. In all of the following discussions, U [a, b] means a uniform distribution between a and b.
[4] This figure comes from Mississippi State University Extension Services.

risk factor. Instead, we use a range. For example, medium frequency assumes that the probability of a particular risk occurring is between 0.2 and 0.3. To operationalize this probability, we take a random draw from *U* [0.2, 0.3]. Likewise, medium severity assumes a 10-15% loss to the *unconditional* yields. That means, if the risk occurs, the realized yields are only 85-90% of the *unconditional* yields. Hence the operationalized percentage loss in the simulation that represents medium severity will be a random draw from *U* [0.1, 0.15].

(5) Based on the information and quantification about the temporal and spatial correlations, a unique correlation matrix is constructed for each of the 20 risk factors.

(6) For each of the 20 risk factors, a series of temporally and spatially correlated risk numbers (for each of the 30 experimental units at each of the 50 time periods) are generated using the Cholesky decomposition of the correlation matrix. This involves multiplying the decomposition matrix with a series of multivariate standard normal random numbers. To be more precise, we first draw a series of multivariate standard normal random numbers. For example, let $\mathbf{r} \sim$ multivariate standard normal $(\mathbf{0}, \boldsymbol{I_n})$ where $\mathbf{r}$ is a $n \times 1$ vector and $\boldsymbol{I_n}$ is an n-dimensional identity matrix. Then we decompose the correlation matrix. Let $\mathbf{A}$ be the correlation matrix such that $\mathbf{A}=\mathbf{L}^{\mathrm{T}}\mathbf{L}$ where $\mathbf{L}$ is an upper triangular matrix. Finally, the decomposition matrix is multiplied by the vector of multivariate standard normal random numbers, i.e., $\mathbf{L}^{\mathrm{T}}\mathbf{r}$ is a vector (with dimension n) of multivariate normal random numbers with mean zero and correlation matrix equal to $\mathbf{A}$. Therefore, $\mathbf{L}^{\mathrm{T}}\mathbf{r}$ is made up of a series of temporally and spatially correlated risk numbers.

(7) To model the probability that an event occurs, a random number from U [0, 1] is drawn. For example, if p is such a random number, then $p<0.1$ means the probability that an event occurs is 0.1. In our simulations, given the fact that the cumulative distribution function (CDF) is distributed as U [0, 1], we can use the CDF at each of those multivariate normal random risk numbers as an indicator of risk occurring. If the CDF is less than the operationalized probability defined in step (4), it means that particular risk factor occurs in an experimental unit at a given time period. Consider an example. If the risk number for the 10^{th} experimental unit at the 25^{th} time period is –0.67, then the CDF of a standard normal distribution at –0.67 will be 0.25. If the frequency level for that risk factor is medium, then the operationalized probability that represents medium frequency will be a random draw from U [0.2, 0.3]. Suppose it is 0.26. Because 0.25 is less than 0.26, then this risk factor occurs at the 10^{th} experimental unit in the 25^{th} time period. If the CDF is greater than the operationalized probability, it means the risk factor does not occur.

(8) If the risk factor occurs, a scaling factor will be defined according to the severity level operationalized in step (4). The scaling factor is actually one minus the percentage loss. For example, if the severity level is medium, it assumes a 10-15% loss to the *unconditional* yields. Then the scaling factor will be a number between 0.85 and 0.9. In this case, the operationalized scaling factor is a random draw from U [0.85, 0.9]. If the risk factor does not occur, the scaling factor will be equal to 1. The same procedure is done for the 20 major risk factors and 20 scaling factors come up.

(9) *Unconditional* yields are drawn from a normal distribution with mean and variance specified in the step (2) and (3).

(10) Scaling factors are applied to the *unconditional* yields. Assuming the effects of the twenty risks on the yields are independent, the final realized yields will be the *unconditional* yields multiplied by all those twenty scaling factors.

(11) Premium rates are estimated for 14 specific time periods, based on yield data up to that time period. That is, in constructing the premium rate for time *t*, yield data from 1 to *t* are used.

(12) The simulation process is replicated 1000 times. For greater detail about the simulation process, readers are directed to Appendix F where a pseudo code in SAS-IML is given.

(13) The performances of the twelve methodologies are compared based on the mean squared error (MSE), which is given by:

$$MSE = \sum_{i=1}^{1000} (estimated\ rate_i - true\ rate)^2 / 1000$$

(14) The true rate is recovered by replications using the empirical rate method, where 25,000 yields are simulated and utilized.

4.2 Relationship between Sample Size and Similarity among Experimental Units

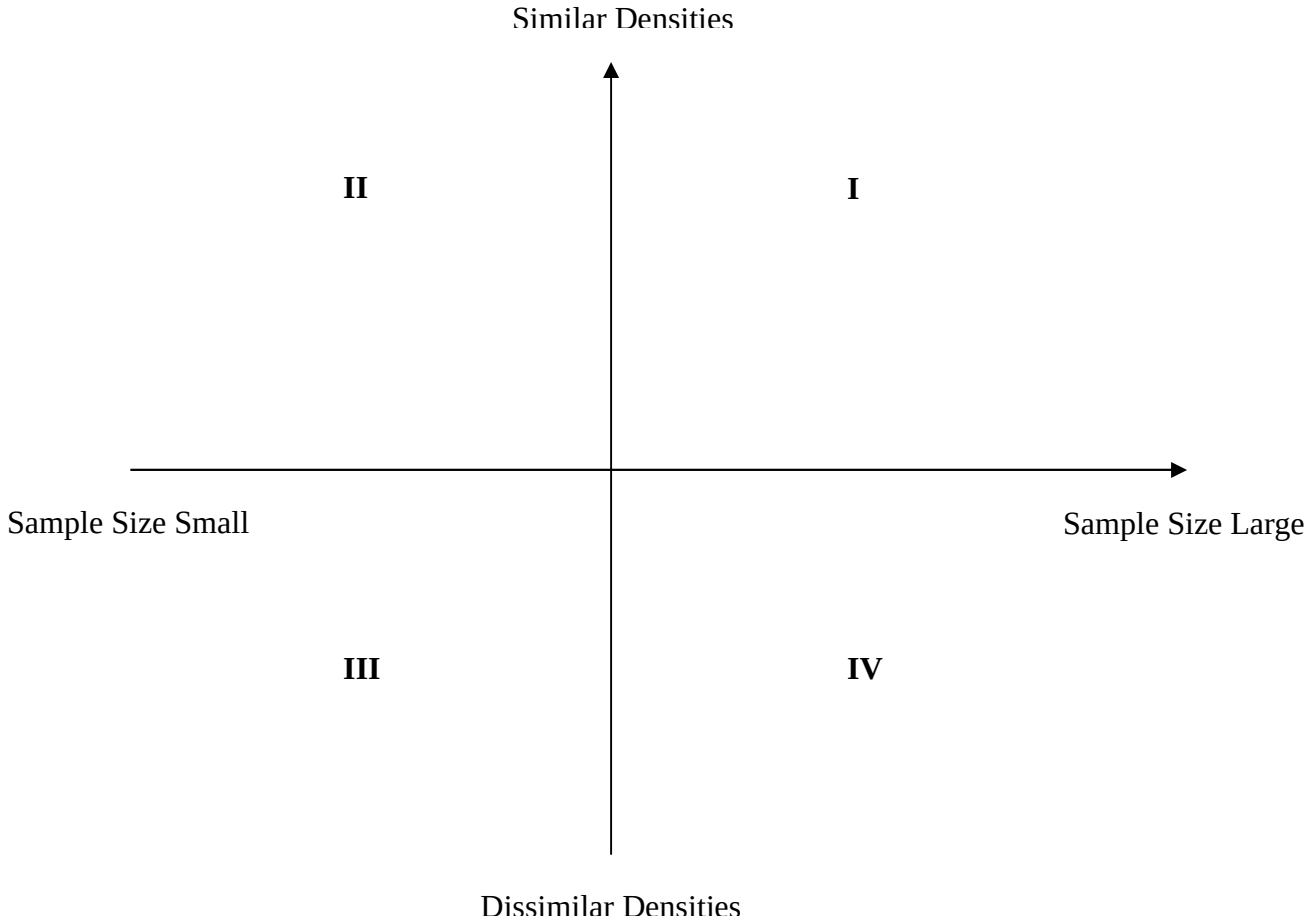
When estimating the yield density for the first experimental unit, the standard approach is to use only the yield data from that experimental unit. However, yield data from a particular experimental unit may not be abundant for estimation of the density. Therefore, yield data from other experimental units are incorporated in the estimation

process in hopes of potential efficiency gains. The incorporation of extraneous information is based upon the assumption that other experimental units are similar in structure to the first experimental unit. If the experimental units are, in fact, structurally dissimilar, the use of extraneous data could lead to an efficiency loss. Hence there is a trade-off in making use of extraneous data, depending on the availability of data and the similarity among the experimental units. The relationship between the sample size and the similarity among the experimental units is summarized in Figure 4.1.

In the second quadrant where the sample size is relatively small and the degree of similarity is relatively high, the extraneous data are very important. In this case, the available individual data is sufficiently scarce that estimators based on those data will be relatively inefficient. However, because the densities of the experimental units are similar, incorporating the data from other experimental units could significantly increase the efficiency. If the set of densities are not exactly similar bias will result, but the decrease in variance may be sufficiently large so that gains in efficiency result.

An opposite case is in quadrant IV, where the sample size is relatively large and the degree of similarity is relatively small. In this case, the extraneous data are not important. As mentioned before, the problem with scarce data is that the estimators are relatively inefficient. However, as we have more individual data, our estimators will have less variance. The advantage of adding extraneous information for potential efficiency gains may not be as prominent as in small sample. In addition, because the set of densities are very dissimilar in this case, adding extraneous information will introduce

Figure 4.1: Relationship Between Sample Size and Similarity Among Experimental Units



Note:

Quadrant I: The importance of extraneous data depends on the trade-off between the increase in bias and decrease in variance.

Quadrant II: Extraneous data are important. Estimators do better because variance decreases more than bias increases.

Quadrant III: The importance of extraneous data depends on the trade-off between the increase in bias and decrease in variance.

Quadrant IV: Extraneous data are unimportant. Estimators do worse because bias increases more than variance decreases.

large biases into the estimation process. Therefore, it is less likely that any decreases in variance resulting from using extraneous data will be sufficient to offset the increase in bias. Estimators using extraneous data will do worse in this case.

In quadrant I, whether the estimators will do better or worse depends on the trade-off between the additional biases relative to the decrease in variance resulting from the incorporation of extraneous information. In this situation, because the set of densities are relatively similar, bias increase will be small. But at the same time, the decrease in variance is also small because of the relatively large sample size. Overall, if the bias increases more than the variance decreases, then the estimators will do worse and vice versa.

In quadrant III, there is again the trade-off between increased bias and decreased variance. But in this case, bias increase will be large due to the relatively large dissimilarity among the experimental units. The decrease in variance is also large because of the small sample size. Overall, if the variance decreases more than the bias increases, the estimators will do better and vice versa.

4.3 Designs of Other Scenarios

4.3.1 The Relationship among the Twelve Rating Methodologies

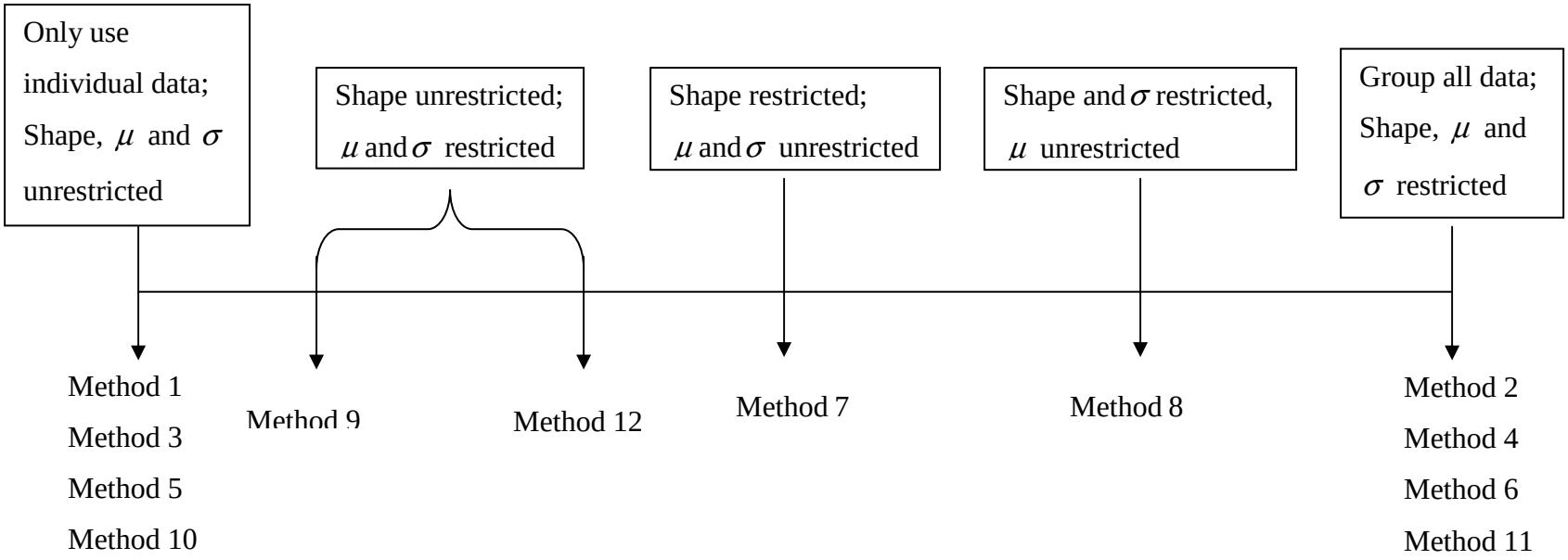
In Chapter 3, we discussed twelve rating methodologies for estimating the actuarially fair premium rate of the first experimental unit. The notations for these methodologies are contained in Table 4.1. We will follow these notations in all of our discussions of results.

Table 4.1: Notations of Rating Methodologies

Method 1	Empirical Rate for Experimental Unit 1 (See Section 3.3.1)
Method 2	Empirical Rate for All Experimental Unit (See Section 3.3.2)
Method 3	Assuming Normal Distribution for Experimental Unit 1 (See Section 3.2.1)
Method 4	Assuming Normal Distribution for All Experimental Units (See Section 3.2.2)
Method 5	Kernel Density Estimation for Experimental Unit 1 (See Section 3.3.3)
Method 6	Kernel Density Estimation for All Experimental Units (See Section 3.3.4)
Method 7	Kernel Density Estimation of All Experimental Units with Transformation of Both Location and Scale Parameters of Experimental Unit 1 (See Section 3.3.5)
Method 8	Kernel Density Estimation of All Experimental Units with Transformation of Location Parameter of Experimental Unit 1 (See Section 3.3.6)
Method 9	Bayesian Nonparametric Kernel Density Estimation (See Section 3.3.7)
Method 10	Assuming Beta Distribution for Experimental Unit 1 (See Section 3.2.3)
Method 11	Assuming Beta Distribution for All Experimental Units (See Section 3.2.4)
Method 12	Estimation of Similar Densities (See Section 3.3.8)

The relationship among the twelve methodologies is summarized in Figure 4.2. In fact, they represent a continuation in the way that extraneous data are utilized in the estimation process. Because we are trying to estimate the density for the first experimental unit, if the shape, location and scale parameters are based on individual data, we would consider they are *unrestricted*, in the sense that the characteristics of the individual data are preserved. However, if they are based on the pooled data, which consists of the data from all experimental units, we consider them to be *restricted*. Restriction on the location parameter is the strongest because it has the largest effect on the estimation of premium rate, followed by restrictions on the scale parameter and then by the shape.

Figure 4.2: Relationship Among the Twelve Rating Methodologies



Note:

***Unrestricted*:** preserving the characteristics of the individual data

***Restricted*:** preserving the characteristics of the pooled data

At one extreme located at the left side of the figure, only individual data are used (methods 1, 3, 5 and 10) and thus the characteristics of the individual data are completely preserved. Located at the far right side, data from other experimental units are used in the estimation of the density for the first experimental unit without any adjustments for individual effect (methods 2, 4, 6 and 11). That is, the shape, location and scale parameters of the density are derived entirely from the pooled data. There are four methodologies that fall between these two extremes. They use the pooled data with respect to either changing the shape, location or scale parameters of the density. Depending on how they preserve the characteristics of the density of the individual data, these four methodologies also come in order. Both methods 9 and 12 use the individual data to correct the shape of the density while restricting the location and scale parameters to the pooled data. Method 9 is closer to the left-hand side because it uses the densities derived in method 5 as the base, while method 12 starts with the densities derived in method 6 and then multiplies an individual correction function to adjust for individual effect. Method 7 and 8 come before method 9 and 12. Both of them preserve the shape of the density of the pooled data, but method 7 uses the location and scale parameters of the first experimental unit to transform the pooled data, while method 8 only uses the location parameter in the transformation. In this sense, method 7 uses more individual information and thus is less restrictive than method 8.

We can see that these twelve methodologies use the pooled data in different degrees, from the lowest degree on the left-hand side to the highest degree on the right-hand side. Therefore, varying the relevancy or heterogeneity of the pooled data should

affect the performances of the different estimators in the reverse order. That is, methods to the right are affected most and methods to the left are least affected.

4.3.2 Variations from the Base Scenario

The base scenario is the starting point of our simulation study. Based on the discussion in the previous sections, we can consider three distinct variations from the base scenario to design other scenarios.

4.3.2.1 Variation in the Quality of Extraneous Data

The first variation is in the quality of extraneous data. Essentially we have two types of estimators. The first type of estimator only utilizes individual data, i.e., only the data from the first experimental unit is used in estimating the premium rate for it. The second type of estimator uses both individual data and extraneous data, i.e., information in other experimental units is "borrowed" to recover the premium rate for the first experimental unit. When the sample size is small, the extraneous data may assist in improving the efficiency in estimating the premium rate for the first experimental unit, depending on how much and how relevant the information contained in other experimental units is. Therefore, by varying the information contained in other experimental units and varying the similarities among the experimental units, we can design scenarios that should affect the performances of estimators that use the extraneous data. This kind of variation includes changes in the number of experimental units, in the spatial correlation and in the heterogeneity among experimental units.

(1) Changing the number of experimental units. The number of experimental units in the simulation is an important factor when we use extraneous data. The more experimental units, the more potential information are added in the estimation and vice versa. Therefore, we design scenarios 2, 3 and 4 to see the effects of increasing (twice) and decreasing (once) the number of experimental units.

(2) Changing the spatial correlation. By increasing the spatial correlation, the information contained in contiguous experimental units is decreased. By decreasing the spatial correlation, the information contained in contiguous experimental units is increased. Therefore, scenarios 5 and 6 are designed to check the effect of changing the spatial correlation.

(3) Changing the heterogeneity of experimental units. As mentioned before, the similarities among the experimental units play an important role on the performances of the estimators that use extraneous data. Whether these estimators will perform better or worse depend on how identical the densities of the experimental units are. If the experimental units are, in fact, very dissimilar, incorporating extraneous data could lead to efficiency losses. Therefore, scenarios 7 and 8 are constructed to examine the effect of changing the degree of similarities among the experimental units. By making the distributions of the yields among the experimental units more heterogeneous, the information contained in neighboring units is less relevant or less similar. Conversely, by making the distributions of the yields less heterogeneous, the information contained in the neighboring units is more relevant. We vary the

heterogeneity by increasing or decreasing the ranges of the uniform distribution for the mean and standard deviation of yields.

4.3.2.2 Variation in the Temporal Correlation

The second variation is the temporal correlation. Because we assume the temporal correlation just occurs in individual experimental unit, there is no serial correlation among different experimental units. Although all estimators will be affected by this kind of variation, estimators that use only the individual data can be checked to see the net effect of changing the quantification of the temporal correlation. As such, scenarios 9 and 10 are designed.

4.3.2.3 Variation in the Severity Levels

The third variation is the quantification of the severity levels of the risks. It will have influence on performances of all estimators because it changes the shape of the density of the yields. If the quantification of the severity levels is higher, that means, when the risk occurs, it will cause more losses to the insured. The probability that catfish producers will incur a loss is higher. Therefore, the mass in the lower tail—below the guarantee—of the yield density will be greater. But if the quantification is lower, then the lower tail will be thinner. As such, scenarios 11 and 12 are designed to check these effects.

The variations of other scenarios from the base scenarios are summarized in Table 4.2 and the detailed structures of designs for all scenarios are contained in Appendix C.

Table 4.2: Design of Other Scenarios

Scenario	How It Differs from the Base Scenario
1	This is the base scenario
2	Decrease the number of experimental units (from 30 to 10)
3	Increase the number of experimental units (from 30 to 70)
4	Increase the number of experimental units (from 30 to 100)
5	Increase spatial correlation
6	Decrease spatial correlation
7	Increase heterogeneity
8	Decrease heterogeneity
9	Increase temporal correlation
10	Decrease temporal correlation
11	Increase severity levels
12	Decrease severity levels

Table 4.3: The True Premium Rate in Each Scenario

Scenario	True Rate
1	13.15%
2	13.15%
3	13.15%
4	13.15%
5	13.15%
6	13.15%
7	14.99%
8	11.72%
9	13.15%
10	13.15%
11	19.96%
12	6.92%

4.4 Simulation Results

This section discusses the MSE results for each scenario and the performances of the twelve methodologies are compared both horizontally (within scenarios) and vertically (across scenarios, the first scenario being the baseline). For a complete list of results, readers are directed to Appendix D.

The true rate in each scenario is reported in Table 4.3. Note that the true rates in scenarios 2, 3, 4, 5, 6, 9 and 10 are the same as the true rate in scenario 1 because variations in these scenarios do not change the distribution of the yields of experimental unit 1. Therefore, the true rate should remain unchanged. But in scenarios 7, 8, 11 and 12, the distribution of the yields of experimental unit 1 is changed. The true rates under these scenarios are expected to change.

4.4.1 Comparison of Simulation Results within Scenarios

Although the performances of the twelve methodologies may be different under different scenarios, they do show some consistent patterns. Appendix E contains a graphic comparison of the performances of the twelve methodologies in each scenario. In most of the scenarios, methods that use only individual data—methods 1, 3, 5 and 10—have larger MSEs than their counterparts—methods 2, 4, 6 and 11, which use extraneous data. This is expected because the sample size in our simulation is from 2 to 50. In addition, the data generating process might make the densities of the experimental units quite similar. Therefore, adding external information from other experimental units in estimating the premium rate for experimental unit 1 results in large efficiency gains.

In all of the scenarios, methods 2, 6, 11 seem to have relatively small MSEs at all levels of data. Recall that these methods use extraneous data without any adjustments for individual effect. The fact that these estimators perform better than other estimators overall, again, suggests the set of the densities of all experimental units are very similar so that grouping the data will assist in improving the accuracy of the estimators.

For most of the estimators, when the sample size increases, MSEs decrease. But for method 8, this is not always the case. In some of the scenarios, as the sample size increases from 40 to 50, the MSEs of method 8 increases. An extreme case is in scenario 12, where the MSEs decrease until the sample size reaches 10 and then increase steadily afterwards. This fact suggests that method 8 may be very sensitive to the similarities among the experimental units. Recall that method 8 is the kernel density estimation of all experimental units with transformation of location parameter of experimental unit 1. Hence the shape and the scale parameter are restricted by the pooled data. The more restrictions, the higher the potential biases. In addition, in larger samples, we have less variance of our estimators. The decrease in variance resulting from pooling the data may be less than that in smaller samples. Therefore, as the sample size gets larger, the bias introduced by incorporating extraneous data may outweigh the decrease in variance, resulting in larger MSEs. In spite of this, method 8 overall performs competitively relative to other methods. For sample size less than 5, methods 2, 6 and 11 have smaller MSEs than method 8. But from sample size larger than 5, method 8 begins to dominate other methods significantly.

4.4.2 Comparison of Base Scenario and Scenarios 2, 3 and 4

In scenario 2, we decrease the number of experimental units from 30 to 10. Because this change will affect the information contained in other experimental units, we would expect estimators that use extraneous data to be affected, while the performances of the estimators that only use individual data will remain almost unchanged. Simulation results show that, for most of the methods that use extraneous data, such as methods 2, 6, 8, 9 and 12, MSEs are larger compared to the base scenario. For methods 4 and 11, MSEs are first larger and then smaller. For methods that only use individual data, such as methods 1, 3, 5 and 10, MSEs don't change significantly.

Decreasing the number of experimental units has insignificant effect on methods 1, 3, 5 and 10 because this variation only affects the extraneous data. For methods 2, 4, 6, 7, 8 and 11, we will expect that, as the sample size gets larger, MSEs are first larger and then smaller than those in the base scenario. This is because these methods use extraneous data to improve efficiency. When the sample size is small, extraneous data are very important. If we decrease the number of experimental units, the information obtained from contiguous units is decreased and the efficiency gains will be decreased. Therefore, MSEs will be larger. Although the external information is less, the bias introduced also decreases because we are using less extraneous data. When the sample size is large and the decrease in variance resulting from using extraneous data is less, the bias introduced by incorporating external information will outweigh the decrease in variance. Using less external information will introduce less bias, which then results in smaller MSEs. Note that the turning point differs across each method. In our simulation,

results show that MSEs for method 4 becomes smaller when the sample size reaches 10, while methods 2, 6, 8 have larger MSEs for all levels of data, which suggests that the turning points for these methods may occur at larger sample sizes. For methods 9 and 12, decreasing the number of experimental units will result in larger MSEs for all levels of data. This is true because when the sample size is large, these estimators will converge to the individual estimators. Using extraneous data would not introduce large bias for these two methods at large samples.

In scenario 3 and 4, we increase the number of experimental units to 70 and 100, respectively. Again, we expect the changes in the performances of the estimators will follow the trend described in the previous paragraph. For methods 1, 3, 5 and 10 in these two scenarios, MSEs don't change much because increasing the number of experimental units will only affect the methods that use extraneous data. In scenario 3, MSEs of methods 4, 6 and 11 are smaller compared to the base scenario, while MSEs of methods 2 and 8 are first smaller and then larger than those in the base scenario. These results are consistent with the trend we mentioned in the previous paragraph, which shows that, for these methods, using more extraneous data will benefit more in small samples than large samples. But for methods 9 and 12, we can see some inconsistent results. MSEs of both estimators are larger than that in the base scenario, although they still converge in large samples. This may suggest that if we have a large number of experimental units, incorporating all this information will result in large biases, which may offset any decreases in variance, even in smaller samples. This trend is even more obvious in scenario 4, where the number of experimental units is ever greater. In scenario 4, most of

the methods that use extraneous data have larger MSEs than they do in the base scenario. Only methods 2, 6 and 8 behave normally in this scenario.

4.4.3 Comparison of Base Scenario and Scenarios 5 and 6

In scenario 5, we increase the spatial correlations among the experimental units, while we decrease the spatial correlations in scenario 6. The effects on the estimators mimic the effects of changing the number of experimental units. By increasing the spatial correlations, the information contained in contiguous experimental units is decreased, just as decreasing the number of experimental units. Conversely, by decreasing the spatial correlations, the information contained in contiguous experimental units is increased, just like the effect of increasing the number of experimental units. Simulation results show that, again, for methods 1, 3, 5 and 10 in these two scenarios, MSEs do not change much because changing the spatial correlation only affect the data contained in other experimental units. Thus methods that don't use extraneous data will not be affected. Results in scenario 5 show that methods 7, 8 and 9 have larger MSEs than they do in the base scenario, which is expected because of the following reason. When the sample size is small, we need extra information in the estimation for potential efficiency gains. But if the information contained in other experimental units is less, which is the case in scenario 5, then MSEs will increase. Results in scenario 6 show that methods 7, 9 and 12 have smaller MSEs than in the base scenario, which is another consistent example. For other methods—those methods that using extraneous data—it seems that they all behave almost similarly in three scenarios. This may suggest that

changing the spatial correlations among the experimental units does not really have much effect on the performances of most of the estimators. This may be true because of the way we define the spatial correlations. Remember the spatial correlations among the experimental units decrease exponentially if the experimental units are farther away. Therefore, only the most contiguous experimental units will really have influences on experimental unit 1. Moreover, we use small number to quantify the levels of spatial correlations. For example, we assign 0.5 as the highest level of spatial correlation. This can also be reason for the insignificant effect of changing spatial correlations.

4.4.4 Comparison of Base Scenario and Scenarios 7 and 8

In scenario 7, we decrease the similarity among the experimental units by making their yield distributions more heterogeneous, while in scenario 8, we increase the similarity by making the distribution less heterogeneous. Again, this variation will have effects on the methods that use extraneous data. Simulation results show that, the performances of the methods that only use individual data—methods 1, 3, 5 and 10—don't change much compared to the base scenario. However, for methods that use extraneous data, things are different. In scenario 7, methods 2, 6, 7 and 9 have larger MSEs than in the base scenario. This is due to the lower similarity among the different experimental units. If the experimental units are less identical, when we use the extraneous data in the estimation, a larger bias will be introduced, which then results in larger MSEs. In addition, the increase of MSEs will go up at a faster rate in larger samples than in small samples. Recall that in a small sample, the extraneous data are

very important, which will decrease the variance of the estimator significantly. In that case, even though the experimental units are less similar, incorporating extraneous data will end up with only slightly higher MSEs. However, when the sample size is large, the decreases in variance are not large enough to offset the bias increases, MSEs will increase even more, which explains the larger difference in MSEs in larger samples than in smaller samples. Let's take an example. For methods 9 in scenario 7, the increase in MSE is 12% at sample size 3, 18% at sample size 7, 46% at sample size 30 and 49% at sample size 50, compared to the base scenario. Other methods such as methods 2, 7 and 9 also show the same pattern. In scenario 8, methods 4, 7, 9 and 12 have smaller MSEs than in the base scenario, which is a counter example to scenario 7 as expected.

4.4.5 Comparison of Base Scenario and Scenarios 9 and 10

In scenario 9, we make the temporal correlations among time periods stronger, while in scenario 10, we make the temporal correlations weaker. Methods 1, 3, 5 and 10 are checked to see the net effect of changing temporal correlation. Results show that, in scenario 9 these methods have higher MSEs than they do in the base scenario, while in scenario 10 they have smaller MSEs. This suggests that decreasing the temporal correlations might improve the performances of the estimators. This situation is very similar to the case when we change the spatial correlations among the experimental units. When we use individual data to estimate the density, the more information contained in the data, the better our estimates will be. If the temporal correlations are increased, the information contained in the data is decreased and the estimators will do worse.

Conversely, if the temporal correlations are decreased, the information contained in the data is increased and thus the estimators will do better.

4.4.6 Comparison of Base Scenario and Scenarios 11 and 12

In scenario 11, we make the severity levels stronger, i.e., the risks will cause more losses, while in scenario 12, the severity levels are made weaker. This kind of variation will affect performances of all estimators because it changes the shape of the yield distributions. The true rate under scenario 11 is significantly larger than in any other scenarios. This is expected because when we increase the quantification of severity, the associated loss also increases. Therefore, the probability that a loss occurs is greater, which results in a fatter lower tail—below the guarantee—of the yield density. Premium rates depend on the mass in the lower tail. The larger mass in the lower tail, the higher premium rates. In addition, MSEs of each estimator in this scenario are significantly higher than those in the base scenario. This is true because when the mass under the yield density below the guarantee is greater, the variance of the density at each support point in that area will be larger. Since premium rates also depend on the density estimates in that area, the estimators of premium rates will have larger errors. Conversely, the true rate in scenario 12 is much smaller than those in any other scenarios. All estimators perform significantly better than they do in the base scenario.

4.5 Summary

There exist some general results that may be ascertained from the simulation study.

- Overall, estimators that use extraneous data perform better than those that only use individual data. In most of the twelve scenarios, the performances of methods 1, 3 and 5, which are the simplest methods that utilize individual data, are poorest. This suggests that, only using the individual data may result in large inefficiency in the estimated premium rates. Therefore, when the sample size is small, it is appropriate to include external information.
- Using extraneous data does not necessarily improve the performances of the estimators. This is due to two reasons. First, when the sample size is large, the decrease in variance resulting from incorporating external information may be small relative to the increase in bias. Therefore, the estimators will perform worse. Second, extraneous data should be used only if one is reasonably confident that there is some form of similarities among the experimental units. If the similarities are too small, estimators will do worse.
- In all of the scenarios, methods 2, 6, 11 have relatively small MSEs at all levels of data. Method 8 also performs encouragingly, except in scenario 12. At sample size less than 5, methods 2, 6 and 11 perform better than method 8, while in larger sample, method 8 dominates all other methods with significantly smaller MSEs.

5. SUMMARY AND CONCLUSION

For the past two decades, aquaculture has been the fastest growing segment of agriculture in the United States. Given the importance of aquaculture, the Risk Management Agency of the United States Department of Agriculture has begun to investigate the feasibility of providing insurance tools for four aquaculture species, catfish, trout, salmon and baitfish, because these species have the largest economic values. This thesis focuses on catfish insurance.

Today, the catfish industry is the largest sector in the U. S. aquaculture industry in terms of production and sales values. As with other agriculture practices, catfish producers also face a variety of production hazards. The major perils include diseases, water related problems, off-flavor, bird predation etc., which significantly affect the profitability of the industry and hinder its further development. Due to the nature of catfish aquaculture production practices, the implementation of aquaculture insurance to the catfish industry will present a number of difficulties. Lack of historical yield data is one of the most fundamental obstacles in rating catfish insurance policies. If historical yield data were available, one could employ time series models to predict future yields and use various parametric and nonparametric approaches to construct the premium rates. However, catfish insurance, or aquaculture insurance is just a pilot program. While there might be data on country-level and state-level, there is no data on county-level or farm-level. Therefore, simulations are needed to generate possible yield data under relatively reasonable assumptions.

The contribution of this thesis is to conduct simulations to generate the yield data based on some possibly relevant data generating processes, and to evaluate the performances of various parametric and nonparametric approaches in determining premium rates. These data generating processes involve the various major risk factors associated with the catfish production that affect the yields negatively. This thesis considers twenty major risk factors. Four aquaculture specialists provided their opinions on the frequency and severity of each risk factor as well as their temporal and spatial correlations, which helped to build a very general structure to our simulations. Yield data are generated by modeling these four characteristics of each risk factor.

Twelve rating methodologies are considered to determine the actuarially fair premium rates based on the simulated data. These methodologies are distinct in two ways: parametric or nonparametric, and whether they use individual data or pooled data. The parametric methodologies include normal distribution and beta distribution. The nonparametric methodologies include empirical rates, kernel density estimation, Bayesian nonparametric kernel density estimation, and estimation of possibly similar densities. These methodologies are common in the crop insurance literature. The second distinction is motivated by the fact that data is scarce in reality. Incorporating extraneous yield data may provide large potential efficiency gains in the estimation. The twelve methodologies are considered in an attempt to minimize inefficiencies or inequities in the catfish insurance program. Recovering accurate premium rates is essential to an actuarially sound catfish insurance program. If the premium rates are overestimated or underestimated, program losses will increase due to adverse selection and moral hazard

problems. Of course, these losses cannot be eliminated. Even with abundant data, one still could not estimate the premium rates without any errors. However, the losses may be minimized by appropriate choice of rating methodologies. It is our hope that, through this simulation study, we can eliminate some inappropriate rating methodologies. In the quest of finding some appropriate methodologies for rating catfish insurance policies, mean squared error is used as the criterion.

In order to increase the applicability of this simulation study, we developed twelve scenarios to see how the methodologies will perform under different assumptions. The performances of the twelve methodologies are compared both horizontally (within each scenario) and vertically (across scenarios, the first scenario being the baseline). The findings from the simulations are: First, estimators that use extraneous data perform better than those that only use individual data when the sample size is small. Second, using extraneous data does not necessarily improve the performances of the estimators because it depends on the availability of individual data and the similarities among the experimental units. When the sample size is large, the decrease in variance resulting from incorporating external information may be small relative to the increase in bias. Therefore, the estimators will perform worse. Moreover, extraneous data should be used only if one is reasonably confident that there is some form of similarities among the experimental units. If the similarities are too small, estimators will do worse. Third, in most of the scenarios, methods 2, 6, 8 and 11 perform relatively better than other methods. Methods 1, 3 and 5 perform poorly in all scenarios, which may suggest that these methods are inappropriate in rating catfish insurance contract.

Our simulation study also has an empirical application in that it provides a general structure for simulating yield data when actual data is not available. Many agricultural products also face the problem of scarce data or no data. In that situation, we can follow the simulation process discussed in this thesis to generate yield data for estimation purposes.

There are mainly two future studies on this thesis. First, we might add more rating methodologies such as semiparametric approach in the estimation. Second, a formal survey will be carried out in the industry soon. Based on the results of the survey, we might need to refine the parameters or assumptions in our simulation in order to make it more realistic.

APPENDIX A: CHOLESKY DECOMPOSITION OF NON-POSITIVE-DEFINITE MATRIX

The Cholesky decomposition of the correlation matrix is needed for the sampling of correlated risk numbers. A matrix must be symmetric and positive-definite to be decomposed. Although our correlation matrix is symmetric by construction, sometimes it may not be positive-definite, depending on the values of the temporal and spatial correlation. For example, when T=6, N=4, ρ=0.25 and α=0.5, the correlation matrix is not positive-definite. An easy way to check the definiteness of a matrix is by looking at the eigen values of the matrix. If all the eigen values are positive, then the matrix is positive-definite. For the previous example, one of the eigen values is negative (-0.002), hence it is not positive-definite. When the correlation matrix is not positive- definite, we cannot carry out the Cholesky decomposition.

There are a number of ways to deal with the decomposition of a non-positive-definite matrix and we think the following approach is most appropriate and easy to implement in our situation. The key is to adjust the negative eigen values of the correlation matrix. Recall, for a symmetric positive definite matrix A, it can be written as:

$$A = P\Lambda P^T = (v_1 \cdots v_k)\begin{pmatrix} \lambda_1 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \lambda_k \end{pmatrix}(v_1 \cdots v_k)^T$$

where $\lambda_1, \cdots, \lambda_k$ are eigen values of A, and P is a matrix of independent eigen vectors of A. Our correlation matrix is constructed in the same way. First, we need to find out the eigen values and eigen vectors of the correlation matrix and construct P and Λ. In case any of the eigen values are negative, we will adjust them to a small positive value, 0.05. Other eigen values that are greater than 0.05 will remain unchanged. Hence the new set of eigen values are all positive numbers that are greater than or equal to 0.05. Suppose the diagonal matrix with the new eigen values on the diagonal is $\tilde{\Lambda}$, then the correlation matrix can be reconstructed using $\tilde{\Lambda}$ and the original P, i.e.,

$$\tilde{A} = P\tilde{\Lambda}P^T$$

The new correlation matrix is, by construction, symmetric and positive definite because $\tilde{\Lambda}$ is a diagonal matrix made up of positive numbers. We find that the difference between the new correlation matrix and the original one is very small after this adjustment. For instance, in the previous example, the maximal difference in elements of the two matrices is only 0.006.

APPENDIX B: NONPARAMETRIC TECHNIQUES

B.1 Kernel Density Estimation

For a set of independent observations ($y_1, y_2, ..., y_n$), the kernel density estimator at support point y is defined as:

$$\hat{f}(y) = \frac{1}{nh}\sum_{i=1}^{n} K(\frac{y - y_i}{h})$$

where *h* is the smoothing parameter or bandwidth and $K(\cdot)$ is the kernel function. The kernel estimator places bumps, or individual kernel at each observation and then sums over those bumps. That is, the density estimate at any support point is the sum of the height of the bumps, or kernels at that particular point.

The following assumptions are made on the smoothing parameter *h*:

$$(i)\lim_{n\to\infty} h = 0 \qquad (ii)\lim_{n\to\infty} nh = \infty$$

Assumption (ii) means that the smoothing parameter approaches zero at a slower rate than n^{-1}.

Mean Squared Error (MSE) is a commonly used error metric in density estimation because it captures both the variance and bias of the estimator. It measures the distance between the estimated function and the true function for a given point.

$$MSE = E(\hat{f}(y) - f(y))^2 = \text{var}(\hat{f}(y)) + bias(\hat{f}(y))^2$$

Using the Taylor's series expansion, it can be shown that

$$E\hat{f}(y) = f(y) + \frac{1}{2}h^2 f''(y)\int z^2 K(z)dz + o(h^2)$$

$$= f(y) + \frac{1}{2}h^2 \mu_2(K) f''(y) + o(h^2)$$

The bias is therefore

$$Bias = E\hat{f}(y) - f(y) = \frac{1}{2}h^2 \mu_2(K) f''(y) + o(h^2)$$

That means the kernel density estimator is biased. However, assumption (i) guarantees that the estimator is asymptotically unbiased. The variance of the estimator can be shown in the same way

$$\text{var } \hat{f}(y) = (nh)^{-1} R(K) f(y) + o(nh)^{-1}$$

where $R(K) = \int K(z)^2 dz$. As with the bias, the variance of the estimator goes to zero as the sample size goes to infinity because of assumption (ii). Therefore, the nonparametric kernel density estimator is consistent. Once the bias and variance are known, the MSE is given by:

$$MSE = (nh)^{-1} R(K) f(y) + \frac{1}{4} h^2 \mu_2(K) f''(y)^2 + o(h^4 + (nh)^{-1})$$

The kernel function $K(\cdot)$ determines the shape of the bumps and the smoothing parameter *h* determines their dispersion. Therefore, the choices of $K(\cdot)$ and *h* are very important. MISE is a logical criterion, which is the integration of MSE over the entire support.

$$\begin{aligned}
\text{MISE} &= \int_{-\infty}^{\infty} E[\hat{f}(y) - f(y)]^2 dy \\
&= \int_{-\infty}^{\infty} [E\hat{f}(y) - f(y)]^2 dy + \int_{-\infty}^{\infty} \text{var } \hat{f}(y) dy \\
&= \left\{ (nh)^{-1} R(K) f(y) + \frac{1}{4} h^2 \mu_2(K)^2 \int_{-\infty}^{\infty} f''(y)^2 \right\} \hat{f}(\cdot, h) + o(h^4 + (nh)^{-1})
\end{aligned}$$

Thus, MISE is the sum of the integrated squared bias and the integrated variance of the estimator $\hat{f}(y)$.

The first decision is how to choose a kernel function. Epanechnikov (1969) derived the optimal non-negative kernel function with respect to minimizing MISE of the estimated density. Subsequently, Rosenblatt (1971) showed that choice of a suboptimal kernel, such as the standard normal, results in only a moderate loss in the asymptotic MISE. Therefore, the standard normal kernel function is often chosen in practice. For our analysis, we use the standard normal kernel and evaluate the densities over a range of

minus 5 standard deviations from the mean and up to the guarantee of the insurance policy.

A variety of methods are available for choosing the bandwidth *h* such as cross-validation and Silverman's "rule-of-thumb". Cross validation usually involves repeatedly estimating the density with a single observation omitted and selecting the bandwidth that minimizes the MISE. Parzen (1962) showed that the optimal choice of *h* that minimizes the MISE is given by:

$$h_{opt} = k_2^{-(2/5)}[\int K(t)^2 dt]^{1/5}[\int f''(y)^2 dy]^{-(1/5)} n^{-(1/5)}$$

where $k_2 = \int t^2 K(t) dt$, f is the true density and f'' represents $\partial^2 f / \partial y^2$. However, one can see that the optimal *h* depends on the unknown density $f(y)$ being estimated. Silverman (1986) suggests a rule of thumb for choosing *h* in empirical applications given a normal kernel:

$$h_{opt} = 0.79 \times \min(\text{standard deviation}, \frac{\text{interquartile range}}{1.34}) \times n^{-(1/5)}$$

Silverman found this decision rule to be robust against the level of skewness and the degree of bi-modalness that would be present in the yield data.

There is another decision about the smoothing parameter as to whether it is global or local. A global smoothing parameter gives equal weight to each data realization. But sometimes, this global parameter may undersmooth the detail in the tails of the distribution. This is problematic particularly for long-tailed densities such as the conditional yield densities. Because premium rates depend highly on the lower tail of the conditional yield density, a local smoothing parameter may be considered, and hence the *adaptive kernel* methods. The adaptive kernel estimator allows the smoothing parameter to vary with each realization. Therefore, a vector of smoothing parameters with dimension equal to the data is used instead of just a single smoothing parameter. The smoothing parameter would be inversely related to the denseness of the data so that the tail realizations will not be undersmoothed. For more elaboration of the adaptive kernel method, the readers are directed to Ker and Coble (1998), Ker and Goodwin (2000).

B.2 Bayesian Nonparametric Kernel Density Estimation

Recall the following hierarchical model:

$$\hat{f}_{ij} \mid \mu_{ij} \sim Normal(\mu_{ij}, \sigma_{ij})$$

$$\mu_{ij} \sim Normal(\mu_j, \tau_j^{\,2})$$

where $\mu_{ij} = f_{ij} + \beta_{ij}$, f_{ij} is the unknown density value for unit *i* at support point y_j, β_{ij} is the bias for unit *i* at support point y_j, $\sigma_{ij}^{\,2}$ is the variance of the kernel density estimate for unit i at support point y_j, μ_j is the mean value of the densities across units at support point y_j, and $\tau_j^{\,2}$ is the variance of the densities across units at support point y_j. The posterior estimate for the hierarchical is:

$$\tilde{f}_{ij} = \hat{f}_{ij}\left(\frac{\tau_j^{\,2}}{\tau_j^{\,2} + \sigma_{ij}^{\,2}}\right) + \mu_j\left(\frac{\sigma_{ij}^{\,2}}{\tau_j^{\,2} + \sigma_{ij}^{\,2}}\right)$$

where the unknowns ($\mu_j, \tau_j^{\,2}, \sigma_{ij}^{\,2}$) must be estimated. $\sigma_{ij}^{\,2}$ is estimated by the asymptotic variance:

$$\sigma_{ij}^2 = \text{Var}(\hat{f}_{ij}) = (nh)^{-1} f_{ij} R(K)$$

where $R(K) = \int K^2 dt$. If *K* is the probability density function of the standard normal distribution, then:

$$R(K) = \int \left(\frac{1}{\sqrt{2\pi}} e^{-\frac{t^2}{2}}\right)^2 dt = \int \frac{1}{2\pi} e^{-t^2} dt$$

Using a change of variable technique, let $t = \frac{u}{\sqrt{2}}, then:$

$$R(K) = \int \frac{1}{2\pi} e^{-\frac{u^2}{2}} \frac{du}{\sqrt{2}} = \frac{1}{2\sqrt{2}\pi} \int e^{-\frac{u^2}{2}} du = \frac{1}{2\sqrt{2}\pi} \sqrt{2\pi} = \frac{1}{2\sqrt{\pi}}$$

Therefore,

$$\sigma_{ij}^2 = f_{ij} / (2\sqrt{\pi} nh)$$

where $f(y_{ij})$ is the true density at support point y_j for experimental unit *i*. An estimate of σ_{ij}^2 would be:

$$\hat{\sigma}_{ij}^2 = \hat{f}_{ij} / (2\sqrt{\pi} nh)$$

An estimate of the mean and variance across units is obtained using the method of moment estimators: $\hat{\mu}_j = \frac{1}{10}\sum_{i=1}^{10} \hat{f}_{ij}$ and $\hat{\tau}^2 = \hat{s}_j{}^2 - \frac{1}{10}\sum_{i=1}^{10} \sigma_{ij}{}^2$ where

$\hat{s}_j{}^2 = \frac{1}{(10-1)}\sum_{i=1}^{10} (f_{ij}{}^2 - \hat{\mu}_j)^2$. The following shows the proof of these formulas.

Lemma: $E[\hat{s}_j^2] = (\sum_{i=1}^{Q} \sigma_{ij}^2 / Q) + \tau_j^2$ where $\hat{s}_j^2 = \frac{1}{Q-1}\sum_{i=1}^{Q} (\hat{f}_{ij} - \hat{\mu}_j)^2$ and

$\hat{\mu}_j = \frac{1}{Q}\sum_{i=1}^{Q} \hat{f}_{ij}$.

If $\hat{s}_j^2 = \frac{1}{Q-1}\sum_{i=1}^{Q} (\hat{f}_{ij} - \hat{\mu}_j)^2$, then

$$E[\hat{s}_j^2] = \frac{1}{Q-1} E[\sum_{i=1}^{Q} \hat{f}_{ij}^2 - Q\hat{\mu}_j^2] = \frac{1}{Q-1}[\sum_{i=1}^{Q} \left[E[\hat{f}_{ij}^2]\right] - QE[\hat{\mu}_j^2]].$$

Since $\hat{f}_{ij} \mid \mu_{ij} \sim Normal(\mu_{ij}, \sigma_{ij})$ where $\mu_{ij} \sim Normal(\mu_j, \tau_j{}^2)$ then $\hat{f}_{ij}^2 \sim N(\mu_j, \sigma_{ij}^2 + \tau_j^2)$ and $E[\hat{f}_{ij}^2] = \sigma_{ij}^2 + \tau_j^2 - \mu_j^2$. Similarly, under independence across counties, then:

$$\sum_{i=1}^{Q} \hat{f}_{ij} \sim N(Q\mu_j, \sum_{i=1}^{Q} \sigma_{ij}^2 + Q\tau_j^2)$$

$$\hat{\mu}_j = \sum_{i=1}^{Q} (\hat{f}_{ij} / Q) \sim N(\mu_j, (\sum_{i=1}^{Q} \sigma_{ij}^2 / Q^2) + (\tau_j^2 / Q))$$

$$E[\hat{\mu}_j^2] = (\sum_{i=1}^{Q} \sigma_{ij}^2 / Q^2) + (\tau_j^2 / Q) - \mu_j^2$$

Hence,

$$E[\hat{s}_j^2] = \frac{1}{Q-1}\left[\sum_{i=1}^{Q} \left[E[\hat{f}_{ij}^2]\right] - QE[\hat{\mu}_j^2]\right]$$

$$= \frac{1}{Q-1}\left[\sum_{i=1}^{Q} \sigma_{ij}^2 + Q\tau_j^2 - Q\mu_j^2 - \frac{\sum_{i=1}^{Q} \sigma_{ij}^2}{Q} - \tau_j^2 + Q\mu_j^2\right]$$

$$= \frac{1}{Q-1}\left[\frac{\sum_{i=1}^{Q}\sigma_{ij}^2}{Q}(Q-1)+\tau_j^2(Q-1)\right]$$

$$= \frac{\sum_{i=1}^{Q}\sigma_{ij}^2}{Q}+\tau_j^2$$

Thus, $\tau_j{}^2$ is estimated by $\hat{\tau}^2 = \hat{s}_j{}^2 - \frac{1}{Q}\sum_{i=1}^{Q}\sigma_{ij}{}^2$. Note, if $\hat{\sigma}_{ij}^2 = \hat{\sigma}_j^2 \ \forall\, i$ =1, … , Q, then the common estimator $\hat{\tau}_j^2 = \hat{s}_j^2 - \hat{\sigma}_j^2$ would result. Therefore, the empirical Bayes nonparametric kernel density estimator at support point y_j for experimental unit *i* is:

$$\tilde{f}_{ij} = \hat{f}_{ij}\left(\frac{\hat{\tau}_j{}^2}{\hat{\tau}_j{}^2+\hat{\sigma}_{ij}{}^2}\right)+\hat{\mu}_j\left(\frac{\hat{\sigma}_{ij}{}^2}{\hat{\tau}_j{}^2+\hat{\sigma}_{ij}{}^2}\right)$$

APPENDIX C: THE STRUCTURE OF SIMULATION DESIGN

Table C.1: Base scenario

Number of experimental units	*N*=30			
Heterogeneity of yields	Mean		Standard Deviation	
	U [50000, 100000]		*U* [10000, 25000]	
Characteristics of risk	High	Medium	Low	Negligible
Frequency	0.4-0.5	0.2-0.3	<0.1	<0.01
Severity	0.2-0.3	0.1-0.15	<0.05	<0.01
Temporal Correlation	0.25	0.1	0.01	0
Spatial Correlation	0.25	0.1	0.01	0

Table C.2: Scenario 2 (decrease the number of experimental units)

Number of experimental units	*N*=10			
Heterogeneity of yields	Mean		Standard Deviation	
	U [50000, 100000]		*U* [10000, 25000]	
Characteristics of risk	High	Medium	Low	Negligible
Frequency	0.4-0.5	0.2-0.3	<0.1	<0.01
Severity	0.2-0.3	0.1-0.15	<0.05	<0.01
Temporal Correlation	0.25	0.1	0.01	0
Spatial Correlation	0.25	0.1	0.01	0

Table C.3: Scenario 3 (increase the number of experimental units)

Number of experimental units	*N*=70			
Heterogeneity of yields	Mean		Standard Deviation	
	U [50000, 100000]		*U* [10000, 25000]	
Characteristics of risk	High	Medium	Low	Negligible
Frequency	0.4-0.5	0.2-0.3	<0.1	<0.01
Severity	0.2-0.3	0.1-0.15	<0.05	<0.01
Temporal Correlation	0.25	0.1	0.01	0
Spatial Correlation	0.25	0.1	0.01	0

Table C.4: Scenario 4 (increase the number of experimental units)

Number of experimental units	*N*=100			
Heterogeneity of yields	Mean		Standard Deviation	
	U [50000, 100000]		*U* [10000, 25000]	
Characteristics of risk	High	Medium	Low	Negligible
Frequency	0.4-0.5	0.2-0.3	<0.1	<0.01
Severity	0.2-0.3	0.1-0.15	<0.05	<0.01
Temporal Correlation	0.25	0.1	0.01	0
Spatial Correlation	0.25	0.1	0.01	0

Table C.5: Scenario 5 (increase spatial correlation)

Number of experimental units	*N*=30			
Heterogeneity of yields	Mean		Standard Deviation	
	U [50000, 100000]		*U* [10000, 25000]	
Characteristics of risk	High	Medium	Low	Negligible
Frequency	0.4-0.5	0.2-0.3	<0.1	<0.01
Severity	0.2-0.3	0.1-0.15	<0.05	<0.01
Temporal Correlation	0.25	0.1	0.01	0
Spatial Correlation	0. 5	0.25	0.1	0

Table C.6: Scenario 6 (decrease spatial correlation)

Number of experimental units	*N*=30			
Heterogeneity of yields	Mean		Standard Deviation	
	U [50000, 100000]		*U* [10000, 25000]	
Characteristics of risk	High	Medium	Low	Negligible
Frequency	0.4-0.5	0.2-0.3	<0.1	<0.01
Severity	0.2-0.3	0.1-0.15	<0.05	<0.01
Temporal Correlation	0.25	0.1	0.01	0
Spatial Correlation	0. 1	0.05	0.005	0

Table C.7: Scenario 7 (increase heterogeneity)

Number of experimental units	*N*=30			
Heterogeneity of yields	Mean		Standard Deviation	
	U [30000, 120000]		*U* [10000, 40000]	
Characteristics of risk	High	Medium	Low	Negligible
Frequency	0.4-0.5	0.2-0.3	<0.1	<0.01
Severity	0.2-0.3	0.1-0.15	<0.05	<0.01
Temporal Correlation	0.25	0.1	0.01	0
Spatial Correlation	0.25	0.1	0.01	0

Table C.8: Scenario 8 (decrease heterogeneity)

Number of experimental units	*N*=30			
Heterogeneity of yields	Mean		Standard Deviation	
	U [70000, 80000]		*U* [10000, 15000]	
Characteristics of risk	High	Medium	Low	Negligible
Frequency	0.4-0.5	0.2-0.3	<0.1	<0.01
Severity	0.2-0.3	0.1-0.15	<0.05	<0.01
Temporal Correlation	0.25	0.1	0.01	0
Spatial Correlation	0.25	0.1	0.01	0

Table C.9: Scenario 9 (increase temporal correlation)

Number of experimental units	*N*=30			
Heterogeneity of yields	Mean		Standard Deviation	
	U [50000, 100000]		*U* [10000, 25000]	
Characteristics of risk	High	Medium	Low	Negligible
Frequency	0.4-0.5	0.2-0.3	<0.1	<0.01
Severity	0.2-0.3	0.1-0.15	<0.05	<0.01
Temporal Correlation	0.5	0.25	0.1	0
Spatial Correlation	0.25	0.1	0.01	0

Table C.10: Scenario 10 (increase temporal correlation)

Number of experimental units	*N*=30			
Heterogeneity of yields	Mean		Standard Deviation	
	U [50000, 100000]		*U* [10000, 25000]	
Characteristics of risk	High	Medium	Low	Negligible
Frequency	0.4-0.5	0.2-0.3	<0.1	<0.01
Severity	0.2-0.3	0.1-0.15	<0.05	<0.01
Temporal Correlation	0.1	0.05	0.005	0
Spatial Correlation	0.25	0.1	0.01	0

Table C.11: Scenario 11 (increase severity)

Number of experimental units	*N*=30			
Heterogeneity of yields	Mean		Standard Deviation	
	U [50000, 100000]		*U* [10000, 25000]	
Characteristics of risk	High	Medium	Low	Negligible
Frequency	0.4-0.5	0.2-0.3	<0.1	<0.01
Severity	0.3-0.4	0.2-0.25	<0.1	<0.05
Temporal Correlation	0.25	0.1	0.01	0
Spatial Correlation	0.25	0.1	0.01	0

Table C.12: Scenario 12 (decrease severity)

Number of experimental units	*N*=30			
Heterogeneity of yields	Mean		Standard Deviation	
	U [50000, 100000]		*U* [10000, 25000]	
Characteristics of risk	High	Medium	Low	Negligible
Frequency	0.4-0.5	0.2-0.3	<0.1	<0.01
Severity	0.1-0.2	0.05-0.1	<0.01	0
Temporal Correlation	0.25	0.1	0.01	0
Spatial Correlation	0.25	0.1	0.01	0

APPENDIX D: SIMULATION RESULTS FOR EACH SCENARIO

Notation:

Method 1: Empirical Rate for Experimental Unit 1

Method 2: Empirical Rate for All Experimental Units

Method 3: Assume Normal for Experimental Unit 1

Method 4: Assume Normal for All Experimental Units

Method 5: Kernel Density Estimation for Experimental Unit 1

Method 6: Kernel Density Estimation for All Experimental Units

Method 7: Kernel Density Estimation for All Experimental Units with Transformation of Experimental Unit 1's Mean and Variance

Method 8: Kernel Density Estimation for All Experimental Units with Transformation of Experimental Unit 1's Mean

Method 9: Empirical Bayesian Nonparametric Density Estimation

Method 10: Assume Beta for Experimental Unit 1

Method 11: Assume Beta for All Experimental Units

Method 12: Estimation of Similar Densities

Table D.1: MSE × 1,000 For Scenario 1

Time	Method 1	Method 2	Method 3	Method 4	Method 5	Method 6	Method 7	Method 8	Method 9	Method 10	Method 11	Method 12
2	17.06	3.44	22.02	5.96	8.61	2.93	10.42	6.56	9.32	8.09	3.99	9.31
3	11.50	2.22	13.48	3.80	5.71	1.95	7.05	2.45	6.17	4.39	2.56	6.22
4	9.24	1.74	9.85	2.90	4.73	1.55	5.64	1.22	4.88	2.91	2.06	5.11
5	7.45	1.39	7.63	2.39	3.81	1.25	4.55	0.71	3.75	2.32	1.69	4.12
6	6.12	1.14	5.96	2.06	3.09	1.04	3.72	0.47	2.92	1.88	1.43	3.36
7	5.31	1.01	5.22	1.85	2.83	0.91	3.45	0.41	2.45	1.76	1.27	3.08
8	4.70	0.90	4.57	1.71	2.61	0.82	3.16	0.36	2.13	1.63	1.16	2.83
9	4.19	0.80	4.10	1.59	2.35	0.73	2.87	0.31	1.75	1.48	1.03	2.55
10	3.84	0.74	3.81	1.51	2.18	0.68	2.65	0.27	1.56	1.42	0.97	2.38
15	2.63	0.47	2.65	1.26	1.54	0.44	1.83	0.15	0.92	1.10	0.67	1.70
20	1.99	0.35	2.08	1.16	1.17	0.34	1.37	0.09	0.65	0.97	0.53	1.31
30	1.25	0.22	1.59	1.01	0.80	0.21	0.95	0.07	0.44	0.78	0.36	0.93
40	0.95	0.17	1.37	0.97	0.63	0.16	0.75	0.07	0.35	0.69	0.31	0.75

50	0.76	0.14	1.24	0.95	0.52	0.14	0.63	0.08	0.29	0.63	0.29	0.63

Table D.2: MSE $\times$ 1,000 For Scenario 2

Time	Method 1	Method 2	Method 3	Method 4	Method 5	Method 6	Method 7	Method 8	Method 9	Method 10	Method 11	Method 12
2	16.89	4.82	21.51	6.57	8.61	3.62	9.73	7.00	9.41	7.99	4.81	9.61
3	12.81	3.30	15.52	4.36	6.02	2.43	6.90	3.54	6.51	4.61	3.33	6.78
4	9.77	2.53	10.38	3.37	4.73	1.84	5.39	2.12	4.90	3.01	2.60	5.36
5	8.19	2.05	8.54	2.83	4.07	1.50	4.56	1.49	3.97	2.49	2.18	4.57
6	6.92	1.72	6.93	2.45	3.46	1.29	3.88	1.13	3.13	2.21	1.87	3.91
7	6.06	1.58	6.03	2.17	3.16	1.18	3.53	0.96	2.68	2.00	1.71	3.54
8	5.23	1.40	5.16	1.94	2.72	1.04	3.07	0.82	2.17	1.72	1.52	3.08
9	4.59	1.26	4.57	1.74	2.51	0.92	2.82	0.72	1.86	1.59	1.34	2.84
10	4.26	1.13	4.18	1.58	2.30	0.82	2.57	0.63	1.66	1.52	1.21	2.60
15	2.85	0.79	2.80	1.20	1.55	0.55	1.73	0.42	0.91	1.12	0.83	1.77
20	2.19	0.60	2.22	1.01	1.20	0.41	1.31	0.29	0.64	0.98	0.61	1.38
30	1.44	0.44	1.66	0.83	0.88	0.28	0.93	0.18	0.47	0.86	0.42	1.04
40	1.14	0.36	1.45	0.74	0.71	0.22	0.73	0.13	0.37	0.77	0.32	0.85
50	0.87	0.30	1.26	0.70	0.56	0.18	0.58	0.10	0.30	0.65	0.26	0.70

Table D.3: MSE $\times$ 1,000 For Scenario 3

Time	Method 1	Method 2	Method 3	Method 4	Method 5	Method 6	Method 7	Method 8	Method 9	Method 10	Method 11	Method 12
2	17.68	3.07	22.51	5.57	8.95	2.73	11.49	6.70	9.61	8.36	3.77	9.46
3	12.05	1.97	13.87	3.40	6.39	1.74	8.00	2.32	6.82	4.71	2.38	6.73
4	9.33	1.45	9.98	2.56	4.93	1.29	5.90	1.04	5.13	2.97	1.79	5.21
5	7.76	1.11	7.92	2.08	3.98	1.02	4.66	0.53	4.00	2.56	1.44	4.22
6	6.76	0.97	6.58	1.84	3.43	0.89	4.02	0.35	3.28	2.20	1.29	3.64
7	5.69	0.85	5.41	1.65	2.91	0.78	3.47	0.27	2.64	1.89	1.13	3.10
8	5.05	0.73	4.84	1.52	2.57	0.67	3.12	0.22	2.18	1.67	0.98	2.75
9	4.46	0.66	4.26	1.44	2.24	0.60	2.77	0.19	1.79	1.44	0.90	2.42
10	4.09	0.60	3.88	1.35	2.06	0.54	2.55	0.16	1.57	1.30	0.82	2.22
15	2.79	0.41	2.67	1.12	1.45	0.36	1.79	0.08	0.90	1.02	0.58	1.58
20	2.11	0.32	2.10	1.01	1.14	0.28	1.35	0.06	0.64	0.94	0.46	1.25
30	1.47	0.22	1.59	0.88	0.85	0.19	0.97	0.06	0.45	0.84	0.32	0.96
40	1.10	0.17	1.34	0.82	0.66	0.14	0.75	0.07	0.35	0.71	0.25	0.78
50	0.94	0.15	1.24	0.79	0.58	0.12	0.64	0.08	0.30	0.67	0.22	0.70

Table D.4: MSE $\times$ 1,000 For Scenario 4

Time	Method 1	Method 2	Method 3	Method 4	Method 5	Method 6	Method 7	Method 8	Method 9	Method 10	Method 11	Method 12
2	16.65	2.59	20.60	6.43	9.17	2.51	12.02	7.55	9.85	8.16	3.42	9.84
3	11.67	1.80	13.40	4.36	6.39	1.79	8.09	2.76	7.01	4.57	2.46	6.87
4	9.23	1.35	9.68	3.36	5.20	1.37	6.32	1.27	5.56	3.11	1.95	5.58
5	7.41	1.06	7.51	2.89	4.19	1.11	5.03	0.68	4.37	2.47	1.63	4.51
6	6.23	0.86	6.09	2.57	3.49	0.91	4.21	0.42	3.48	2.16	1.38	3.79
7	5.35	0.76	5.20	2.41	3.06	0.82	3.74	0.32	2.92	1.92	1.29	3.34
8	4.86	0.68	4.66	2.26	2.83	0.74	3.49	0.28	2.61	1.79	1.19	3.11
9	4.31	0.62	4.19	2.17	2.54	0.68	3.19	0.23	2.28	1.60	1.13	2.80
10	3.97	0.58	3.87	2.12	2.31	0.65	2.90	0.20	2.03	1.44	1.09	2.55
15	2.69	0.41	2.80	1.84	1.65	0.47	2.09	0.09	1.30	1.14	0.87	1.88
20	2.06	0.33	2.35	1.73	1.32	0.39	1.68	0.05	0.95	0.98	0.78	1.55
30	1.41	0.25	1.85	1.62	0.96	0.30	1.20	0.03	0.67	0.86	0.67	1.16
40	1.11	0.20	1.65	1.54	0.80	0.24	1.01	0.02	0.56	0.80	0.60	0.99
50	0.94	0.17	1.57	1.49	0.72	0.21	0.92	0.02	0.52	0.75	0.55	0.91

Table D.5: MSE × 1,000 For Scenario 5

Time	Method 1	Method 2	Method 3	Method 4	Method 5	Method 6	Method 7	Method 8	Method 9	Method 10	Method 11	Method 12
2	16.07	3.42	20.02	6.04	8.57	2.93	10.49	6.70	9.24	7.77	3.83	9.31
3	10.99	2.29	12.27	3.74	6.13	1.97	7.44	2.69	6.55	4.50	2.57	6.69
4	8.79	1.72	9.08	2.87	4.92	1.52	5.81	1.35	5.10	2.96	2.00	5.33
5	7.38	1.39	7.58	2.43	4.13	1.26	4.88	0.80	4.06	2.54	1.70	4.46
6	6.31	1.16	6.07	2.14	3.55	1.05	4.21	0.52	3.34	2.26	1.45	3.85
7	5.42	0.98	5.24	1.93	3.07	0.90	3.67	0.40	2.75	1.95	1.25	3.37
8	4.86	0.87	4.74	1.80	2.76	0.81	3.33	0.34	2.33	1.81	1.14	3.03
9	4.42	0.78	4.36	1.68	2.49	0.73	3.04	0.30	1.99	1.67	1.03	2.74
10	4.03	0.71	3.98	1.56	2.33	0.67	2.85	0.27	1.78	1.52	0.94	2.57
15	2.80	0.46	2.98	1.30	1.75	0.44	2.17	0.15	1.12	1.22	0.65	1.94
20	2.10	0.35	2.43	1.14	1.40	0.33	1.72	0.10	0.82	1.06	0.51	1.58
30	1.37	0.24	1.80	1.02	0.93	0.23	1.16	0.06	0.54	0.89	0.38	1.09
40	0.98	0.17	1.49	0.99	0.69	0.17	0.85	0.06	0.40	0.72	0.32	0.83
50	0.78	0.13	1.33	0.95	0.57	0.13	0.71	0.06	0.34	0.64	0.27	0.70

Table D.6: MSE $\times$ 1,000 For Scenario 6

Time	Method 1	Method 2	Method 3	Method 4	Method 5	Method 6	Method 7	Method 8	Method 9	Method 10	Method 11	Method 12
2	17.06	3.45	22.37	6.13	8.19	3.02	10.13	6.30	8.95	7.79	3.90	8.93
3	11.42	2.40	13.58	3.93	5.70	2.11	6.99	2.60	6.24	4.28	2.70	6.21
4	8.69	1.73	9.50	2.95	4.41	1.55	5.29	1.24	4.62	2.82	2.02	4.80
5	6.94	1.34	7.43	2.38	3.61	1.21	4.30	0.75	3.59	2.31	1.63	3.93
6	6.01	1.14	6.21	2.08	3.11	1.04	3.72	0.52	2.92	2.03	1.42	3.40
7	5.21	0.97	5.31	1.84	2.69	0.89	3.24	0.40	2.37	1.67	1.22	2.94
8	4.46	0.86	4.57	1.69	2.34	0.79	2.84	0.34	1.92	1.36	1.10	2.57
9	4.02	0.77	4.04	1.54	2.12	0.70	2.62	0.30	1.59	1.21	0.98	2.34
10	3.56	0.68	3.53	1.49	1.90	0.63	2.36	0.25	1.36	1.07	0.89	2.10
15	2.52	0.46	2.66	1.24	1.45	0.43	1.79	0.13	0.85	0.95	0.64	1.61
20	1.96	0.35	2.20	1.14	1.15	0.33	1.40	0.09	0.62	0.89	0.52	1.29
30	1.30	0.23	1.64	1.02	0.81	0.22	0.99	0.07	0.44	0.77	0.38	0.94
40	1.01	0.17	1.42	0.96	0.65	0.16	0.79	0.08	0.36	0.69	0.31	0.77
50	0.80	0.13	1.26	0.93	0.53	0.12	0.64	0.09	0.30	0.61	0.26	0.64

Table D.7: MSE × 1,000 For Scenario 7

Time	Method 1	Method 2	Method 3	Method 4	Method 5	Method 6	Method 7	Method 8	Method 9	Method 10	Method 11	Method 12
2	19.17	5.90	23.18	5.58	10.05	3.40	11.83	3.95	10.98	9.65	4.09	10.55
3	12.39	4.24	14.47	2.77	6.01	2.47	7.41	1.05	6.99	5.47	2.58	6.38
4	9.70	3.71	10.48	1.83	4.68	2.03	5.83	0.55	5.47	3.49	2.06	4.99
5	7.66	3.23	8.02	1.35	3.64	1.70	4.64	0.44	4.26	2.60	1.68	3.91
6	6.17	2.95	6.22	1.05	2.97	1.48	3.87	0.42	3.46	2.04	1.42	3.22
7	5.36	2.84	5.36	0.90	2.71	1.31	3.62	0.37	2.99	1.77	1.26	2.94
8	4.80	2.71	4.68	0.79	2.50	1.21	3.33	0.35	2.69	1.60	1.13	2.71
9	4.26	2.59	4.07	0.69	2.21	1.10	2.99	0.32	2.30	1.47	1.00	2.41
10	3.90	2.52	3.75	0.63	2.07	1.04	2.81	0.32	2.12	1.33	0.95	2.26
15	2.58	2.13	2.45	0.41	1.46	0.77	2.01	0.27	1.47	1.03	0.64	1.60
20	1.91	1.97	1.85	0.30	1.13	0.63	1.58	0.23	1.13	0.89	0.48	1.25
30	1.21	1.85	1.32	0.19	0.82	0.46	1.23	0.19	0.82	0.76	0.31	0.94
40	0.93	1.77	1.07	0.14	0.66	0.39	1.04	0.17	0.67	0.67	0.25	0.78
50	0.76	1.71	0.93	0.12	0.57	0.36	0.94	0.16	0.57	0.62	0.23	0.67

Table D.8: MSE $\times$ 1,000 For Scenario 8

Time	Method 1	Method 2	Method 3	Method 4	Method 5	Method 6	Method 7	Method 8	Method 9	Method 10	Method 11	Method 12
2	15.66	2.94	21.77	4.97	7.20	2.60	9.03	7.80	7.67	7.00	4.07	7.69
3	10.85	1.93	13.26	3.37	5.05	1.70	6.24	3.87	5.19	3.67	2.67	5.30
4	8.73	1.55	9.60	2.63	4.26	1.35	4.86	2.22	4.13	2.47	2.16	4.34
5	7.20	1.28	7.38	2.18	3.47	1.12	3.81	1.43	3.14	2.02	1.78	3.51
6	6.00	1.06	5.76	1.87	2.82	0.92	3.01	0.98	2.37	1.76	1.50	2.83
7	5.19	0.95	4.98	1.65	2.56	0.82	2.70	0.80	1.94	1.65	1.33	2.55
8	4.58	0.86	4.33	1.50	2.35	0.74	2.43	0.67	1.65	1.50	1.20	2.33
9	4.08	0.76	3.89	1.36	2.13	0.66	2.18	0.58	1.33	1.38	1.06	2.10
10	3.77	0.72	3.59	1.28	1.97	0.62	1.99	0.52	1.16	1.35	1.00	1.93
15	2.68	0.49	2.47	1.01	1.41	0.42	1.28	0.30	0.64	1.08	0.67	1.33
20	2.10	0.39	1.91	0.89	1.08	0.34	0.88	0.18	0.43	0.92	0.52	0.99
30	1.39	0.29	1.39	0.73	0.74	0.26	0.54	0.10	0.30	0.81	0.34	0.65
40	1.10	0.24	1.17	0.68	0.59	0.22	0.40	0.09	0.26	0.71	0.27	0.51
50	0.92	0.22	1.03	0.66	0.50	0.20	0.31	0.09	0.23	0.61	0.25	0.41

Table D.9: MSE $\times$ 10,000 For Scenario 9

Time	Method 1	Method 2	Method 3	Method 4	Method 5	Method 6	Method 7	Method 8	Method 9	Method 10	Method 11	Method 12
2	16.20	3.55	19.50	6.56	9.02	3.12	10.92	7.29	9.56	8.07	4.11	9.71
3	11.27	2.44	12.37	4.27	6.40	2.17	7.79	3.18	6.82	4.56	2.77	6.95
4	9.25	1.81	10.06	3.35	5.14	1.64	6.19	1.74	5.29	3.16	2.16	5.56
5	7.82	1.46	8.18	2.81	4.31	1.35	5.17	1.10	4.24	2.70	1.81	4.67
6	6.73	1.19	6.87	2.48	3.63	1.11	4.39	0.74	3.39	2.26	1.52	3.93
7	5.95	1.06	5.94	2.25	3.27	0.99	3.96	0.60	2.90	2.09	1.37	3.54
8	5.29	0.95	5.29	2.07	2.99	0.89	3.62	0.50	2.54	1.96	1.26	3.24
9	4.69	0.86	4.73	1.96	2.65	0.81	3.19	0.42	2.15	1.75	1.16	2.86
10	4.32	0.76	4.27	1.87	2.39	0.73	2.91	0.36	1.86	1.54	1.06	2.59
15	2.82	0.48	2.88	1.53	1.65	0.48	2.03	0.19	1.13	1.18	0.73	1.84
20	2.19	0.36	2.36	1.34	1.34	0.37	1.65	0.12	0.82	1.06	0.58	1.51
30	1.51	0.24	1.85	1.19	0.96	0.25	1.17	0.07	0.56	0.90	0.44	1.12
40	1.12	0.18	1.57	1.11	0.75	0.19	0.93	0.06	0.44	0.79	0.36	0.91
50	0.90	0.14	1.42	1.06	0.63	0.16	0.79	0.05	0.38	0.70	0.32	0.79

Table D.10: MSE × 1,000 For Scenario 10

Time	Method 1	Method 2	Method 3	Method 4	Method 5	Method 6	Method 7	Method 8	Method 9	Method 10	Method 11	Method 12
2	16.36	3.44	20.28	5.77	8.32	2.95	10.30	6.39	9.00	7.69	3.77	9.00
3	11.57	2.25	13.43	3.59	5.94	1.95	7.31	2.40	6.32	4.42	2.51	6.44
4	8.99	1.61	9.42	2.67	4.59	1.43	5.46	1.19	4.74	2.81	1.91	4.96
5	7.19	1.28	7.43	2.22	3.68	1.16	4.43	0.70	3.62	2.33	1.56	4.01
6	5.78	1.03	5.74	1.92	3.05	0.95	3.66	0.48	2.82	1.97	1.30	3.33
7	4.99	0.86	4.80	1.75	2.56	0.80	3.13	0.37	2.21	1.60	1.12	2.81
8	4.41	0.77	4.31	1.63	2.33	0.72	2.85	0.32	1.90	1.45	1.02	2.54
9	4.00	0.70	3.93	1.50	2.15	0.64	2.65	0.27	1.66	1.36	0.92	2.35
10	3.59	0.62	3.51	1.40	1.97	0.57	2.43	0.24	1.43	1.30	0.82	2.16
15	2.43	0.43	2.53	1.18	1.41	0.40	1.76	0.13	0.90	0.99	0.59	1.58
20	1.88	0.31	2.05	1.06	1.10	0.29	1.35	0.09	0.61	0.86	0.46	1.25
30	1.28	0.20	1.55	0.95	0.77	0.19	0.91	0.08	0.40	0.77	0.33	0.89
40	1.00	0.16	1.35	0.90	0.62	0.14	0.71	0.09	0.33	0.68	0.27	0.73
50	0.78	0.12	1.22	0.88	0.50	0.11	0.60	0.10	0.28	0.61	0.23	0.61

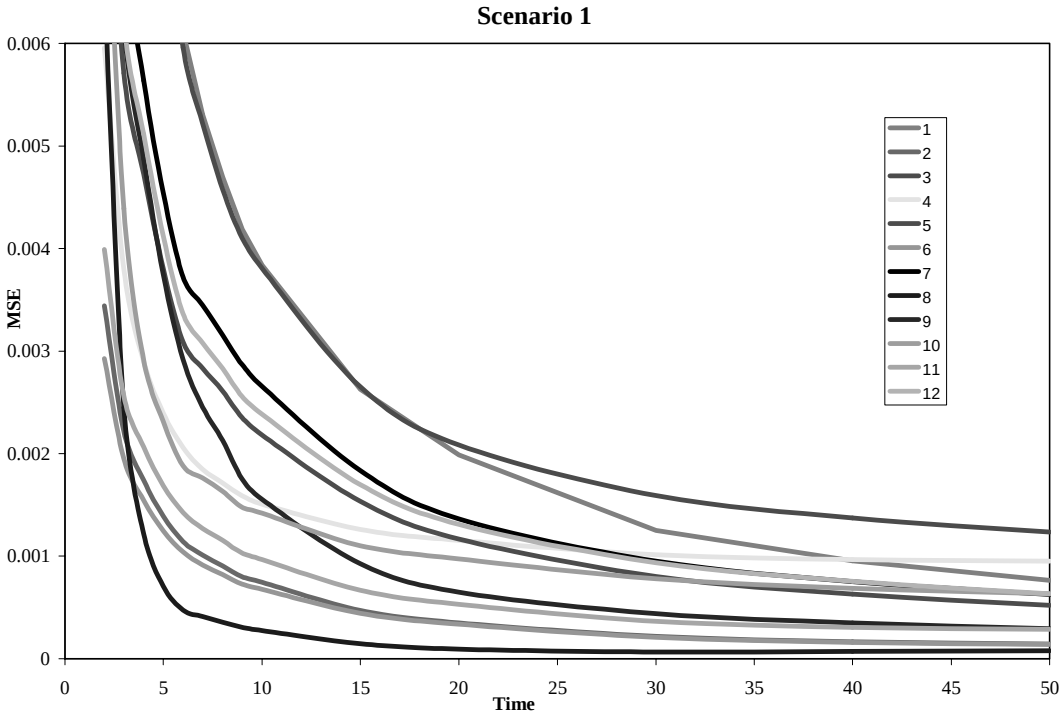
Table D.11: MSE × 1,000 For Scenario 11

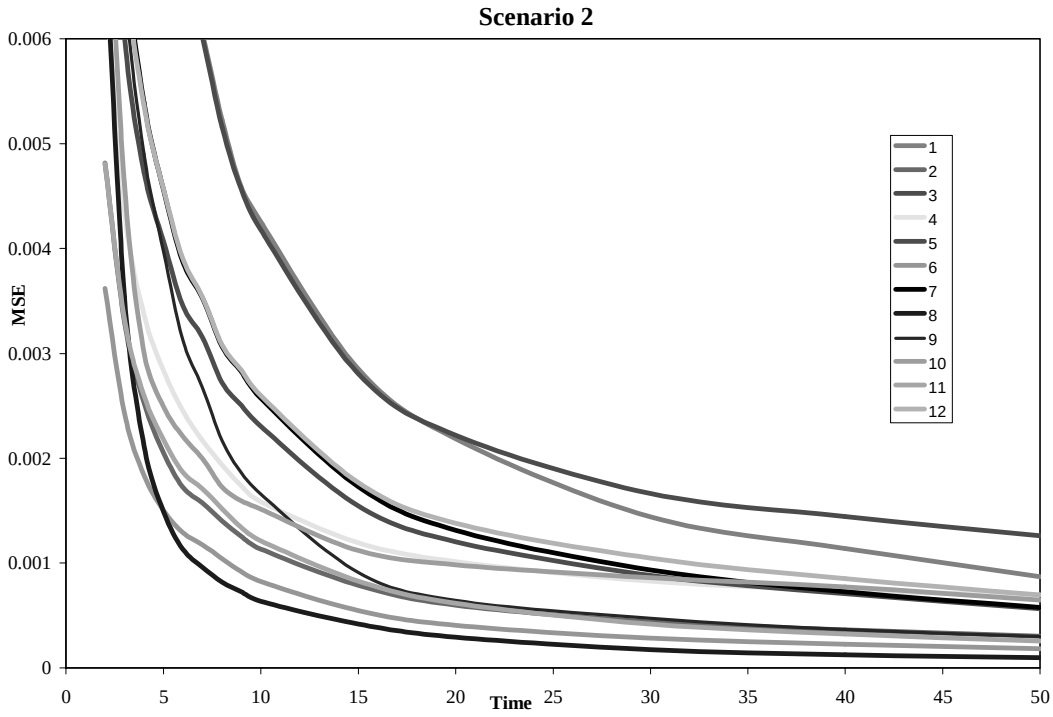
Time	Method 1	Method 2	Method 3	Method 4	Method 5	Method 6	Method 7	Method 8	Method 9	Method 10	Method 11	Method 12
2	33.56	4.79	42.66	12.17	22.10	6.27	21.32	15.68	24.38	19.28	7.41	22.60
3	23.81	2.93	29.41	7.62	15.96	5.23	15.55	7.50	18.51	13.64	4.61	16.31
4	19.30	2.21	22.92	5.80	13.74	4.55	13.80	5.97	15.85	10.77	3.63	13.99
5	15.87	1.78	17.48	4.80	11.86	4.07	12.04	5.53	13.43	8.58	2.95	12.11
6	12.94	1.47	13.64	4.17	10.41	3.74	10.60	5.41	11.58	6.83	2.49	10.60
7	11.55	1.27	11.91	3.77	9.75	3.39	10.00	5.35	10.37	5.90	2.19	9.93
8	10.37	1.11	10.48	3.52	9.32	3.18	9.47	5.24	9.66	5.22	1.97	9.48
9	9.40	0.97	9.40	3.30	8.73	2.98	8.85	5.10	8.80	4.43	1.74	8.87
10	8.68	0.89	8.73	3.13	8.34	2.85	8.36	5.01	8.33	3.97	1.62	8.44
15	5.98	0.57	6.09	2.72	6.74	2.39	6.37	4.56	6.68	2.30	1.07	6.77
20	4.48	0.42	4.80	2.55	5.85	2.12	5.16	4.27	5.83	1.54	0.80	5.85
30	2.86	0.26	3.70	2.31	4.87	1.73	3.84	3.97	4.88	1.00	0.50	4.83
40	2.19	0.19	3.23	2.24	4.28	1.54	3.14	3.80	4.34	0.76	0.38	4.24
50	1.76	0.16	2.92	2.22	3.86	1.42	2.72	3.75	3.95	0.63	0.33	3.80

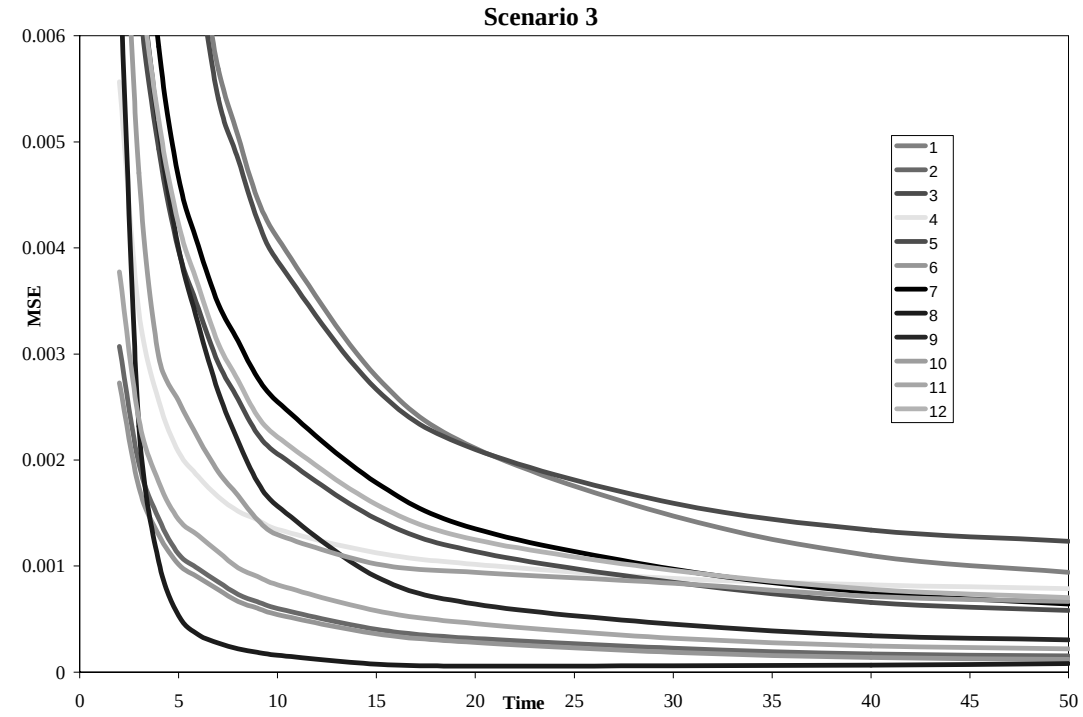
Table D.12: MSE $\times$ 1,000 For Scenario 12

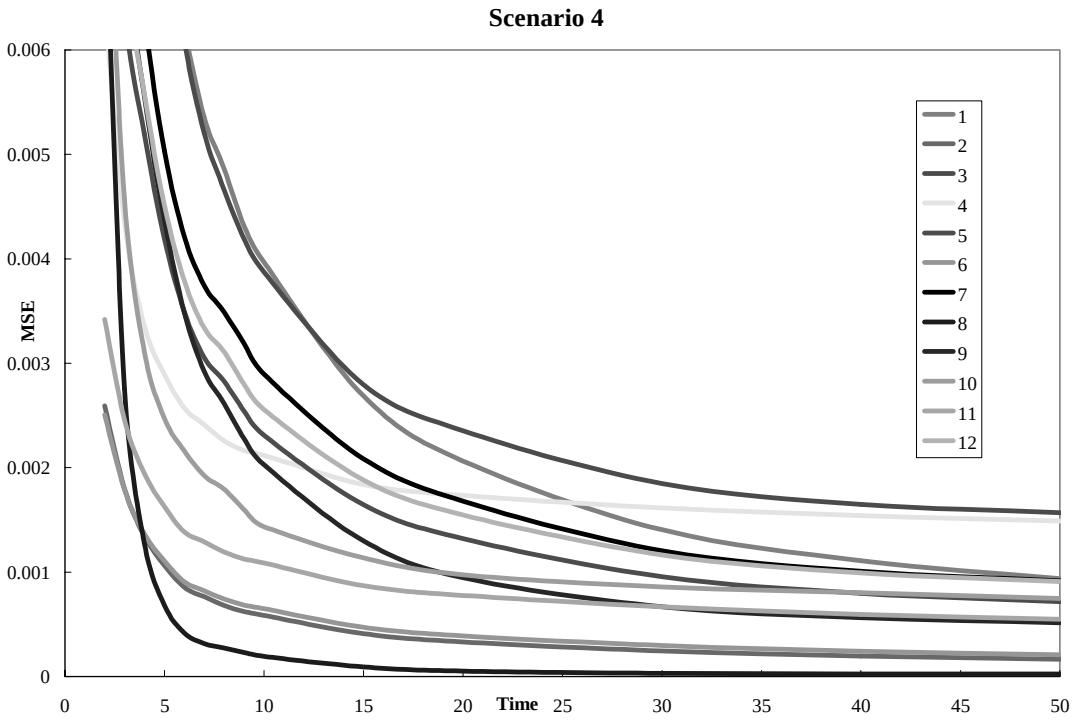
Time	Method 1	Method 2	Method 3	Method 4	Method 5	Method 6	Method 7	Method 8	Method 9	Method 10	Method 11	Method 12
2	5.67	2.06	7.98	1.89	3.30	2.02	4.51	2.40	2.40	1.92	1.96	2.74
3	4.08	1.31	4.51	1.19	2.75	1.30	3.29	1.35	1.57	0.76	1.22	1.96
4	3.28	1.05	3.20	0.89	2.37	1.07	2.54	1.16	1.25	0.75	1.00	1.64
5	2.60	0.83	2.44	0.71	2.05	0.86	2.05	1.07	0.99	1.03	0.79	1.36
6	2.19	0.70	1.93	0.59	1.79	0.73	1.68	1.03	0.82	1.21	0.66	1.13
7	1.88	0.64	1.64	0.53	1.61	0.67	1.50	1.01	0.74	1.32	0.60	1.03
8	1.65	0.58	1.41	0.48	1.46	0.61	1.32	1.00	0.67	1.42	0.54	0.93
9	1.44	0.52	1.24	0.43	1.30	0.55	1.17	0.99	0.56	1.46	0.48	0.82
10	1.31	0.49	1.15	0.40	1.20	0.52	1.08	1.02	0.51	1.47	0.45	0.76
15	0.88	0.33	0.76	0.30	0.86	0.35	0.74	1.09	0.36	1.37	0.28	0.53
20	0.66	0.26	0.59	0.25	0.69	0.28	0.56	1.15	0.29	1.20	0.21	0.40
30	0.42	0.19	0.43	0.19	0.47	0.20	0.39	1.21	0.22	0.84	0.14	0.28
40	0.32	0.15	0.36	0.17	0.37	0.16	0.31	1.27	0.18	0.60	0.10	0.22
50	0.26	0.14	0.31	0.17	0.30	0.15	0.26	1.31	0.16	0.43	0.09	0.18

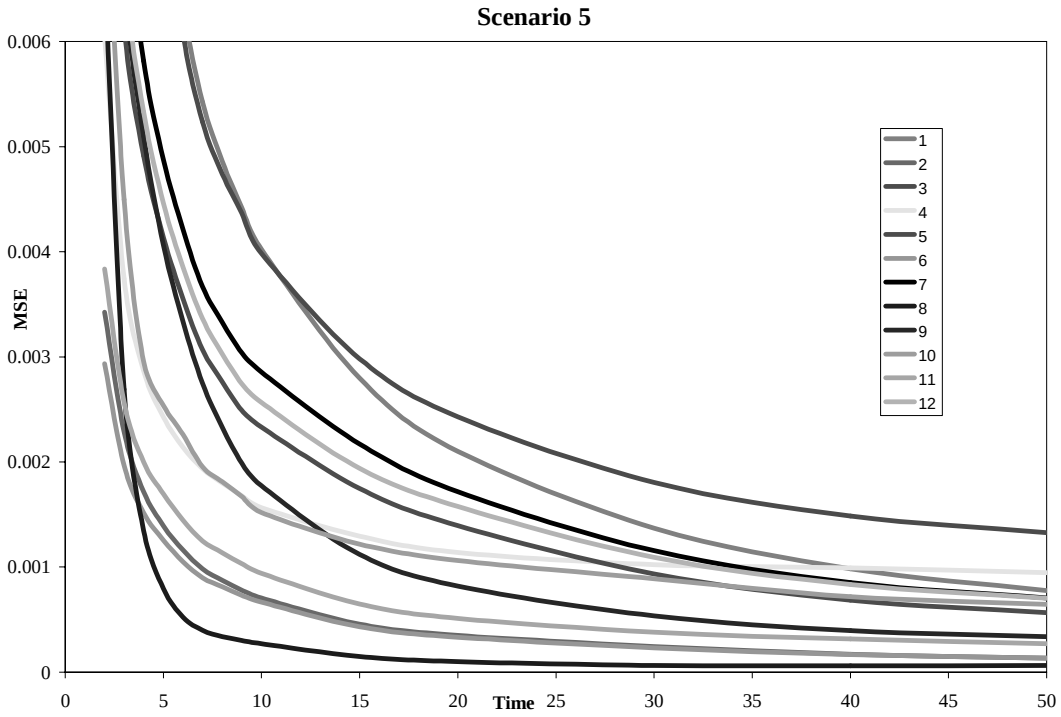
APPENDIX E: COMPARISON OF DIFFERENT METHODS FOR EACH SCENARIO

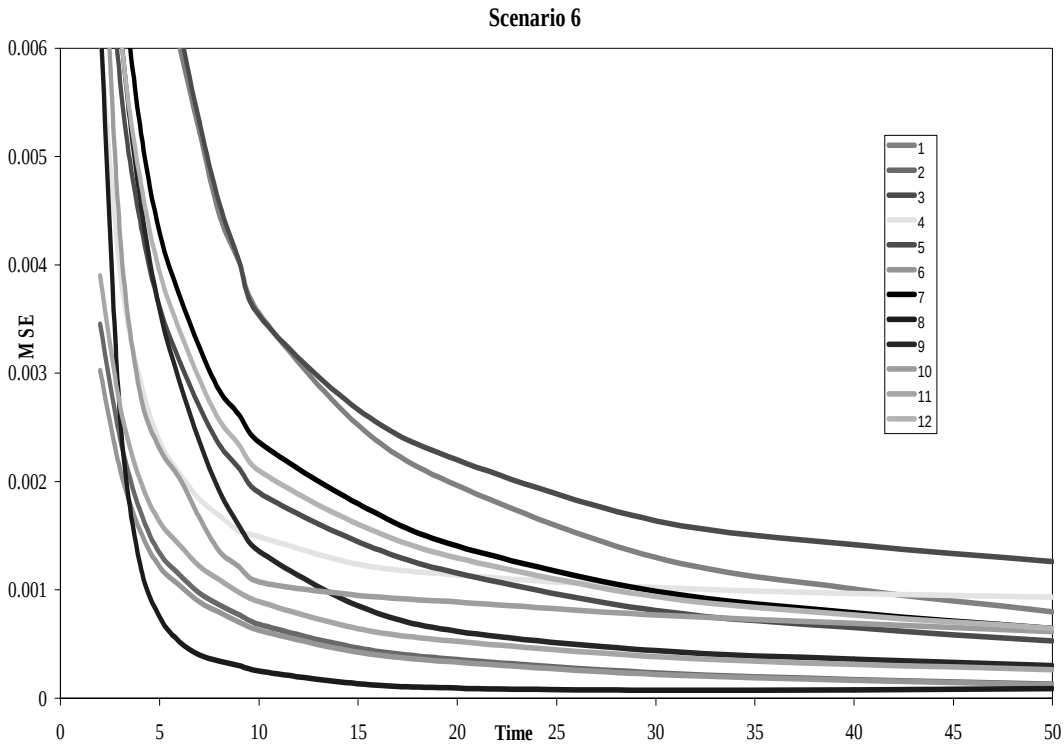


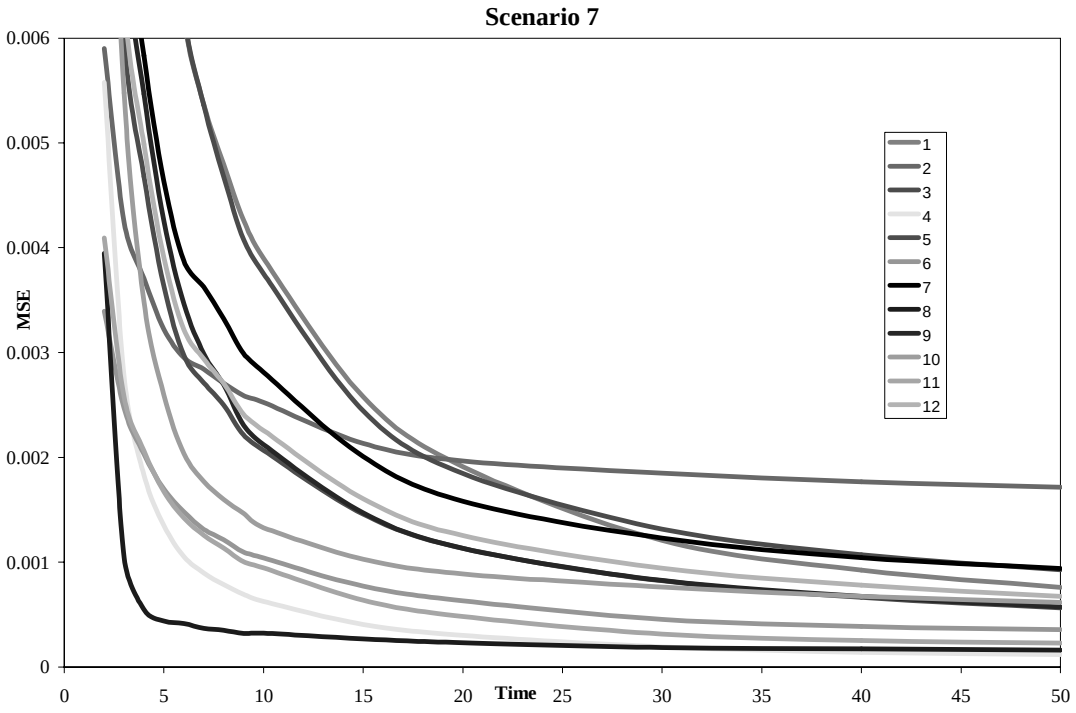


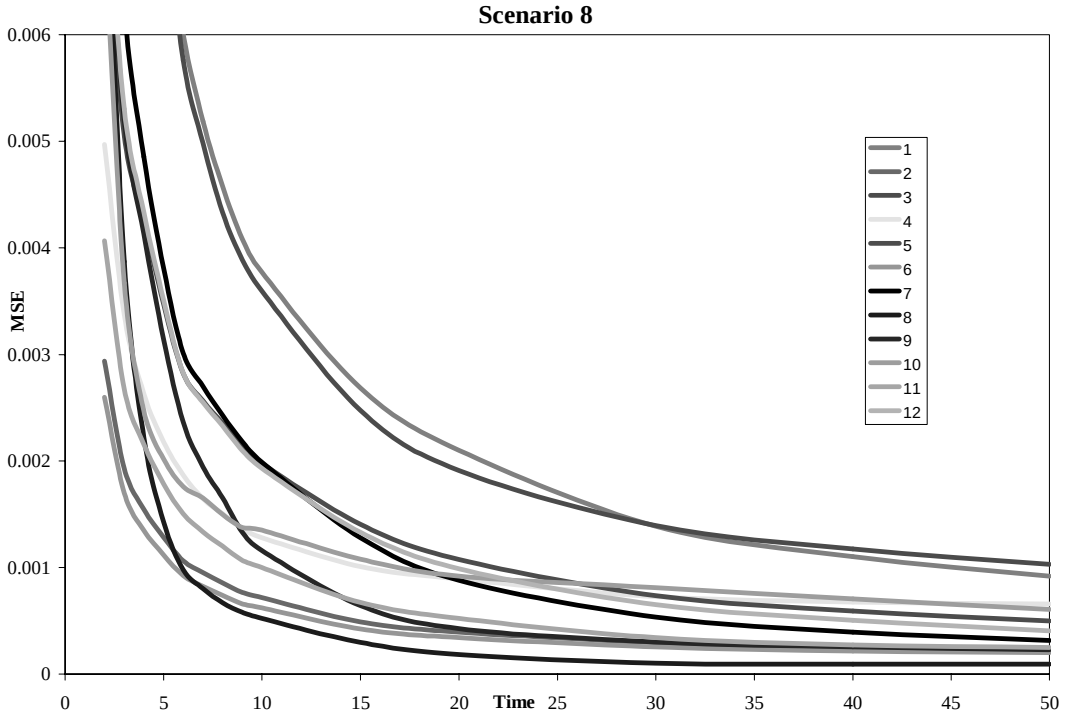

Scenario 3
0.006
0.005
0.004
0.003
0.002
0.001
0
MSE
0
5
10
15
20
Time
25
30
35
40
45
50
1
2
3
4
5
6
7
8
9
10
11
12

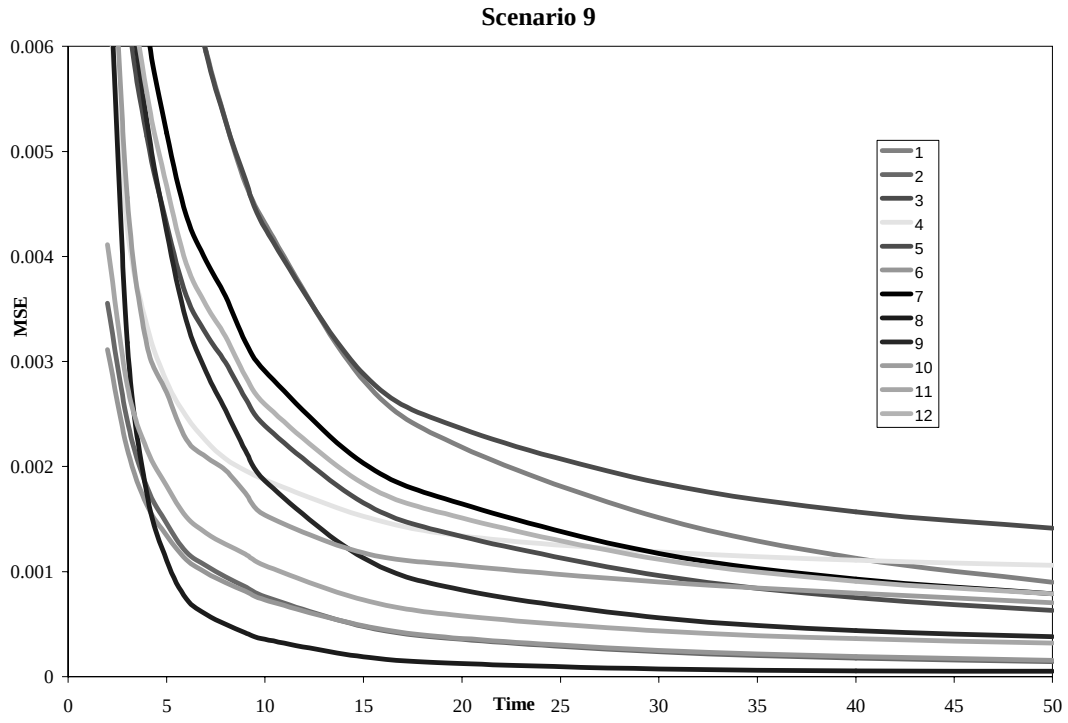

Scenario 4
0.006
0.005
0.004
0.003
0.002
0.001
0
MSE
0
5
10
15
20
Time
25
30
35
40
45
50
1
2
3
4
5
6
7
8
9
10
11
12

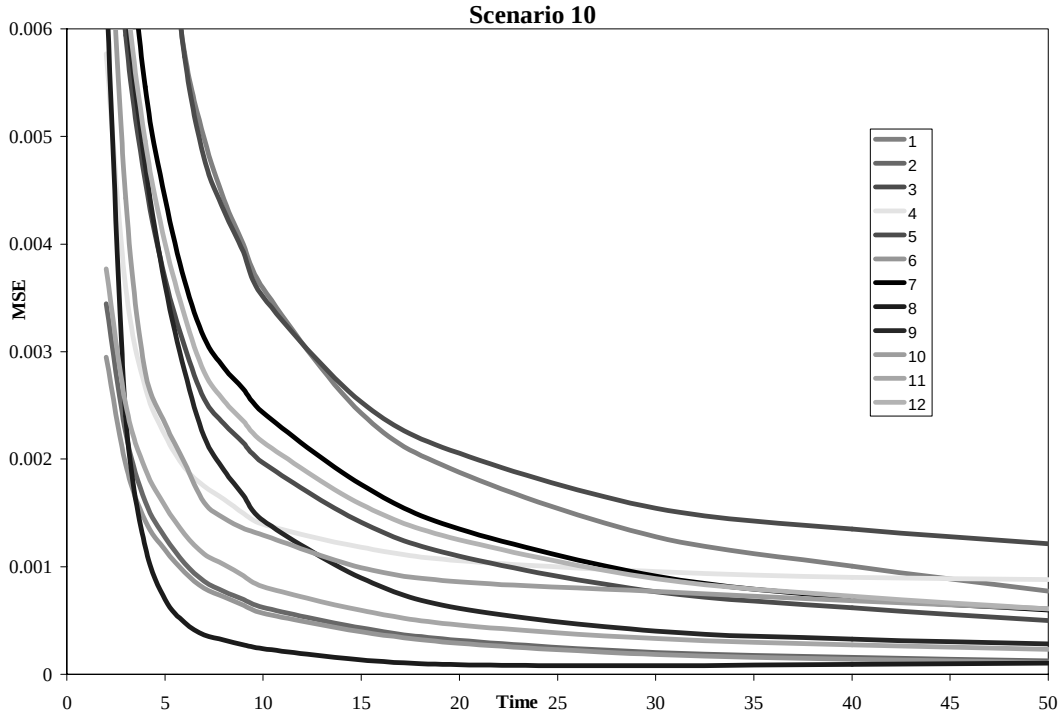

Scenario 5
0.006
0.005
0.004
0.003
0.002
0.001
0
MSE
0
5
10
15
20
Time
25
30
35
40
45
50
1
2
3
4
5
6
7
8
9
10
11
12

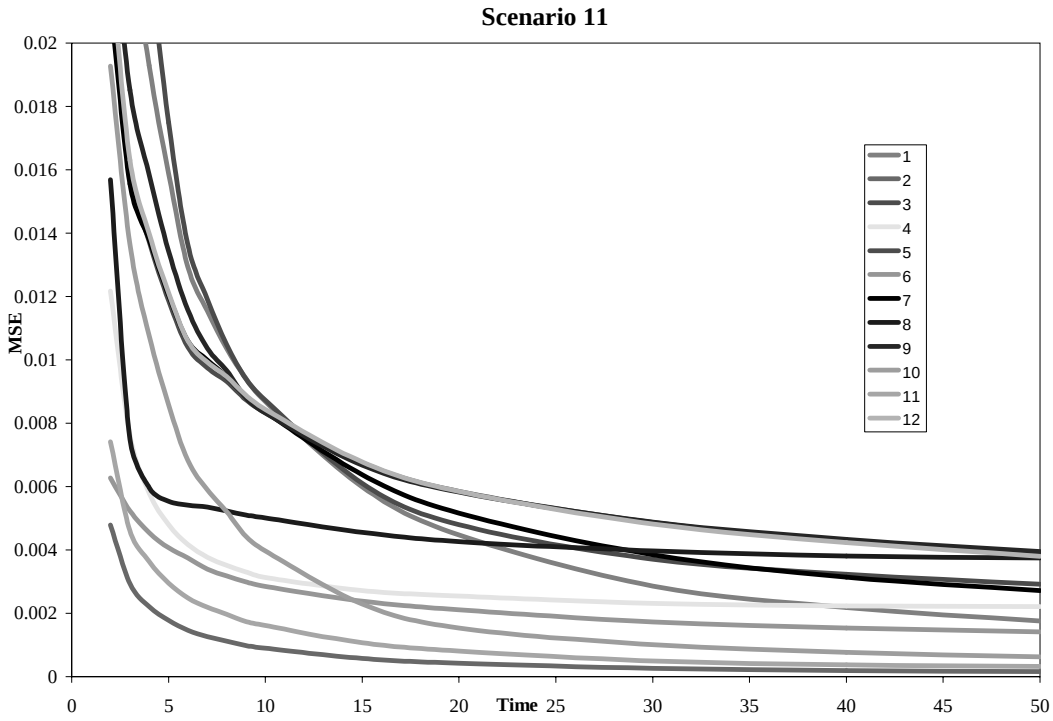

Scenario 6
0.006
0.005
0.004
0.003
0.002
0.001
0
MSE
0
5
10
15
20
Time
25
30
35
40
45
50
1
2
3
4
5
6
7
8
9
10
11
12

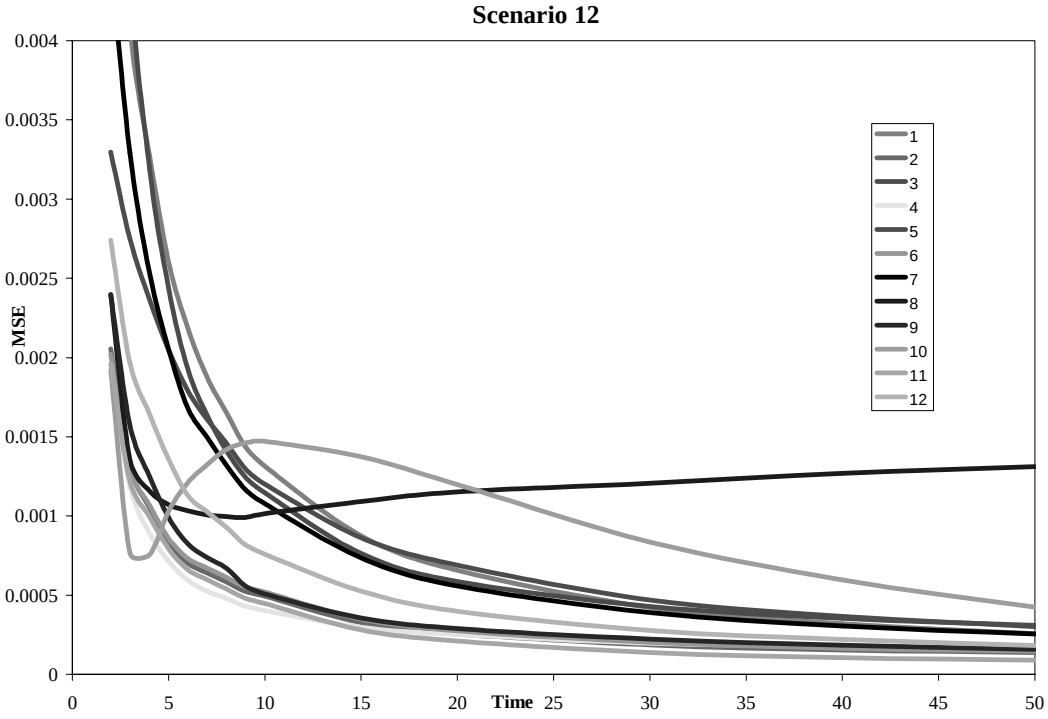

Scenario 7
MSE
Time
0.006
0.005
0.004
0.003
0.002
0.001
0
0
5
10
15
20
25
30
35
40
45
50
1
2
3
4
5
6
7
8
9
10
11
12


Scenario 8
MSE
Time
0.006
0.005
0.004
0.003
0.002
0.001
0
0
5
10
15
20
25
30
35
40
45
50
1
2
3
4
5
6
7
8
9
10
11
12


Scenario 9
0.006
0.005
0.004
0.003
0.002
0.001
0
MSE
0
5
10
15
20
Time
25
30
35
40
45
50
1
2
3
4
5
6
7
8
9
10
11
12


Scenario 10
0.006
0.005
0.004
0.003
0.002
0.001
0
MSE
0
5
10
15
20
Time
25
30
35
40
45
50
1
2
3
4
5
6
7
8
9
10
11
12


Scenario 11
MSE
Time
0.02
0.018
0.016
0.014
0.012
0.01
0.008
0.006
0.004
0.002
0
0
5
10
15
20
25
30
35
40
45
50
1
2
3
4
5
6
7
8
9
10
11
12


Scenario 12
MSE
Time
0.004
0.0035
0.003
0.0025
0.002
0.0015
0.001
0.0005
0
0
5
10
15
20
25
30
35
40
45
50
1
2
3
4
5
6
7
8
9
10
11
12

APPENDIX F: ADDITIONAL SAS-IML CODE FOR YIELD SIMULATOR IN THE BASE SCENARIO

```
/* scenario 1: method 1*/
proc iml symsize=95000000;
time=50;  /* number of time periods */
space=30;  /* number of experimental units */
sim=1000;  /* number of simulations */
yield=j(time,space,0);
truerate=j(sim,1,0);

lambda=0.75;  /* coverage level */

truerate=0.1315;

rate1=j(49,sim,0);

/* parameter values of severity */
sevhigub=0.3; sevhiglb=0.2; sevmedub=0.15; sevmedlb=0.1; sevlow=0.05;
sevneg=0.01;

/* parameter values of temporal correlation */
temhig=0.25; temmed=0.1; temlow=0.01; temneg=0.000001;

/* parameter values of spatial correlation */
spahig=0.25; spamed=0.1; spalow=0.01; spaneg=0.000001;

total=time*space;
iden=i(time);

eee=i(total);

/* randomize the mean and variance of yield distributions*/
mean=j(space,1,0);
var=j(space,1,0);
do j=1 to space;
   mean[j]=50000+50000*ranuni(1234); /* mean is between 50000 and
                                        100000 */
   var[j]=100000000+525000000*ranuni(1234); /* variance is between
                                               100000000 and 625000000*/
end;

/* risk 1: enteric septicemia of catfish */
p1=0.2+0.1*ranuni(1234);  /* frequency */
temrho=temmed*(2/3)+temlow*(1/3);    /* temporal correlation */
sparho=spalow*(2/3)+spaneg*(1/3);    /* spatial correlation */
omega=j(space,space,0);
do i=1 to space;
   do j=1 to space;
   omega[i,j]=sparho**abs(i-j);
   end;
```

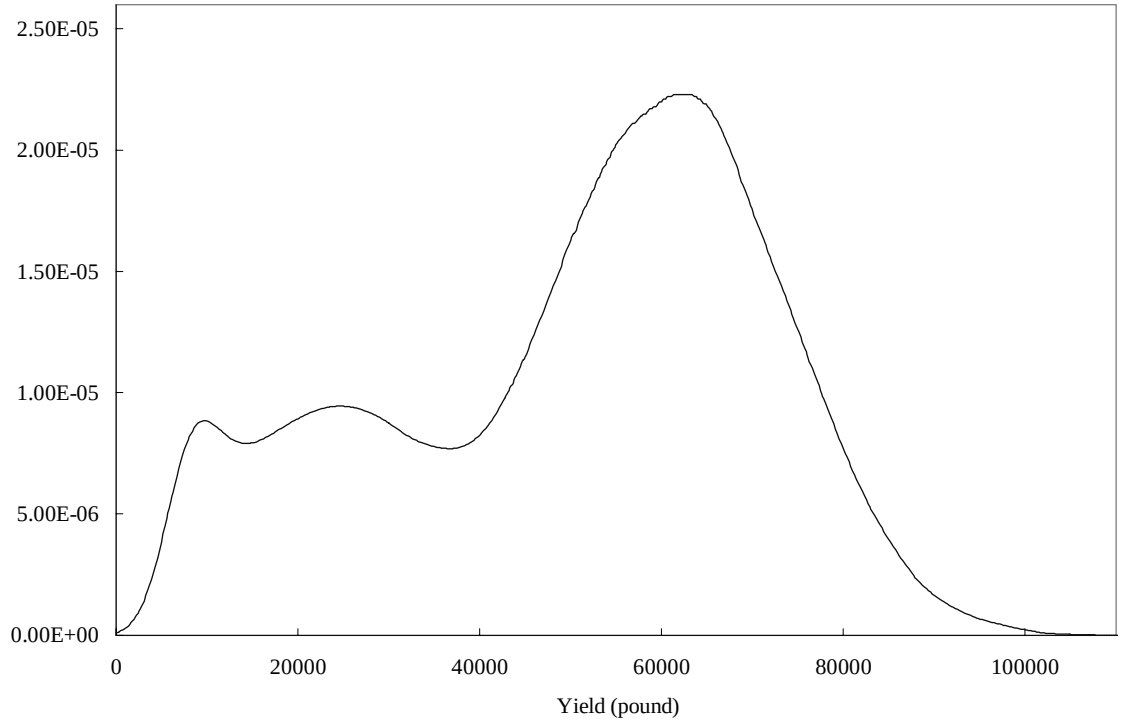
```
end;
varcov=iden@omega;
upper1=root(varcov);

/*...modeling the remaining risk factors */

/* do simulations */
do s=1 to sim;

temp=j(total,1,0);
do i=1 to total;
   temp[i]=rannor(1234);
end;

/* generate mutivariate normal risk numbers for risk factor 1 */
sample=upper1`*temp;
mvn=shape(sample,time,space);  /* reshape the vector sample into a
                                  timeXspace matrice */
d1=cdf('normal',mvn,0,1);   /* take cdf to get U[0,1] number */

/*...do the same step for the remaining risk factors */

do i=1 to time;
   do j=1 to space;

      if d1[i,j]<p1
      then k1=ranuni(4321)*(sevmedub-sevmedlb)+(1-sevmedub);
      else k1=1;

/*   do the same step for the remaining risk factors */

yield[i,j]=(rannor(1234)*sqrt(var[j])+mean[j])*k1*k2*k3*k4*k5*k6*k7*k8*
k9*k10*k11*k12*k13*k14*k15*k16*k17*k18*k19*k20;

   end;
end;

do t=1 to 49;     /* # of time periods we are going to consider */
   c=0;   /* time indicator */
   if t<10 then c=1;
   if t=14 then c=1;
   if t=19 then c=1;
   if t=29 then c=1;
   if t=39 then c=1;
   if t=49 then c=1;

if c=1 then do;

y=yield[1:(1+t),];
ones=j(t+1,1,1);
avg=y[+,]/(t+1);
std=sqrt(((y-ones*avg)##2)[+,]/(t+1-1));
guarantee=lambda*avg[1];
```

```
/* method 1. empirical rate for farm 1 */
rate1[t,s]=sum((guarantee-y[,1])<>0)/(t+1)/guarantee;

end;   /* end of time indicator */

end;   /* end of loop over different time periods */

end;   /* end of simulations */

mse1=((rate1-truerate)##2)[,+]/sim;   /* mse for method 1 */

print mse1;
```

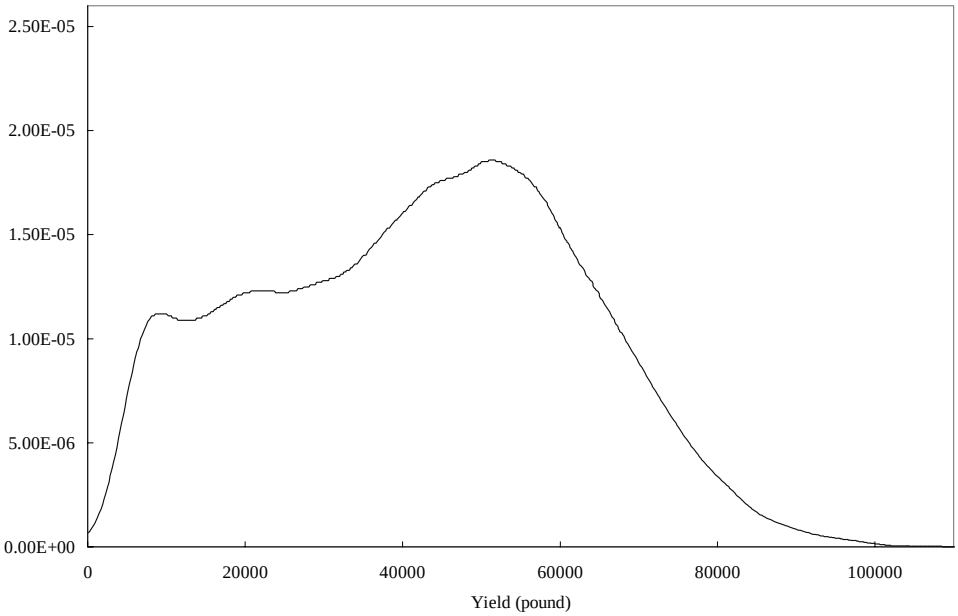
APPENDIX G: KERNEL DENSITY ESTIMATES OF YIELDS

Note: The yield distributions in scenarios 2, 3, 4, 5, 6, 9 and 10 are the same as in scenario 1.

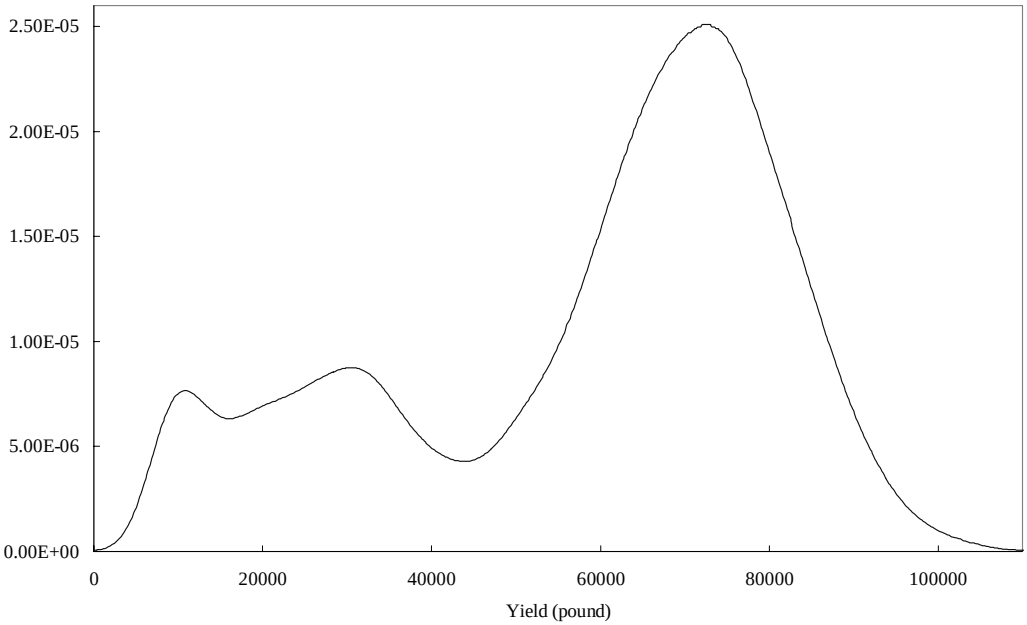
Scenario 1: Kernel Density Estimates of Yields



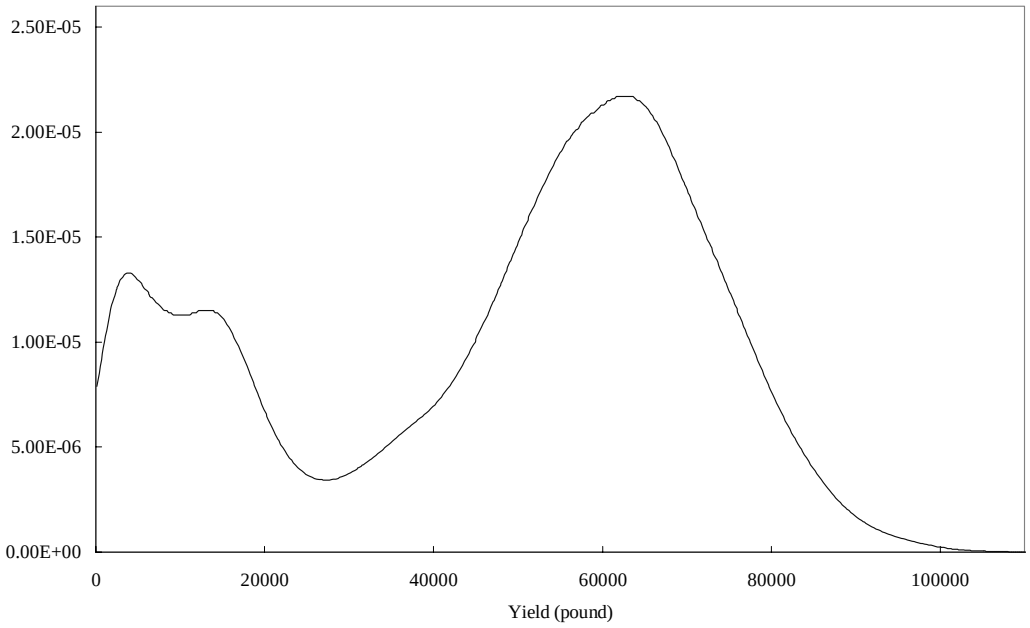
Scenario 7: Kernel Density Estimates of Yields



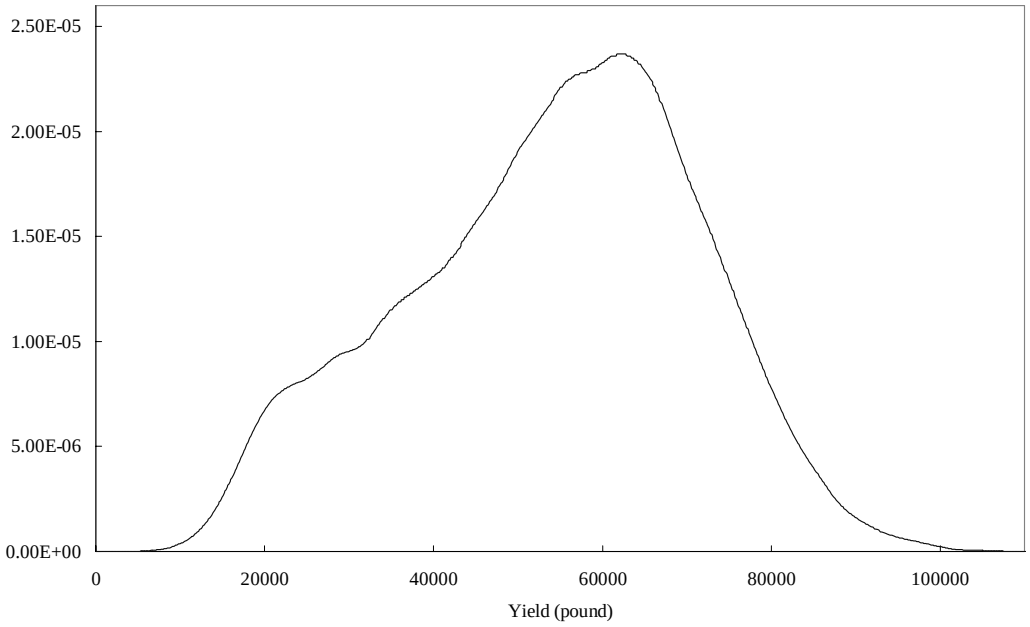
Scenario 8: Kernel Density Estimates of Yields



Scenario 11: Kernel Density Estimates of Yields



Scenario 12: Kernel Density Estimates of Yields



References:

1. 2002 Risk Management for Aquaculture Workshop Report. Memphis, Tennessee.
2. "Aquaculture Industry Report." Animal and Plant Health Inspection Service.
3. "Catfish 2003." USDA, National Animal Health Monitoring System (NAHMS).
4. "Catfish Health Research 2003 Annual Report." USDA, Agricultural Research Service.
5. Economic Impact and Management of Bird Predation at Aquaculture Facilities in the Southeastern U. S., 2004, National Wildlife Research Center Project.
6. "ESC and Vaccination Practices." Catfish 2003 Info Sheet, USDA National Animal Health Monitoring System.
7. 2003 National Risk Management for Aquaculture Workshop Report. Louisville, Kentucky.
8. "Trematodes on U.S. Catfish Operations." Catfish 2003 Info Sheet, USDA, National Animal Health Monitoring System.
9. Avery, J. “The U. S. Catfish Industry: Scope, Cultural Practices, and Associated Risks. ” Invited Presentation, 2002, Memphis, Tennessee.
10. Beem, M. "Aquaculture: Realities and Potentials When Getting Started." *Southern Reginal Aquaculture Center (SRAC) Publication* No. 441(1998).
11. Boyd, C. E., et al. "Environmental Assessment of Channel Catfish Ictalurus punctatus Farming in Alabama." *World Aquaculture Society* 31, no. 4(2000): 511-544.
12. Brunson, M. W. "Brown Blood Disease." *Mississippi State University Extension Service Information Sheet 1390*.
13. Brunson, M. W., C. G. Lutz, and R. M. Durborow. "Algae Blooms in Commercial Fish Production Ponds." *Southern Reginal Aquaculture Center (SRAC) Publication* No. 466(1994).
14. Burtle, G. J. ""No Blood" Disease: Feed Related Anemia in Catfish." Animal & Dairy Science Department 1997 Annual Report.
15. Camus, A. C. "Channel Catfish Virus Disease." *Southern Reginal Aquaculture Center (SRAC) Publication* No. 4702(2004).

16. Carlberg, J. M., and J. C. Van Olst. "U. S. Aquaculture: Current Status and Future Directions." *Aquaculture Magazine* 27, no. 4(2001).

17. Dean, S., T. Hanson, and S. Murray. "Economic Impact of the Mississippi Farm-Raised Catfish Industry at the Year 2003." Mississippi State University Extension Service Publication 2317.

18. Durborow, R. M., A. J. Mitchell, and M. D. Crosby. "Ich (White Spot Disease)." *Southern Reginal Aquaculture Center (SRAC) Publication* No. 476(1998).

19. Durborow, R. M., et al. "Columnaris Disease: A Bacterial Infection Caused by Flavobacterium columnare." *Southern Reginal Aquaculture Center (SRAC) Publication* No. 479(1998).

20. Durborow, R. M., D. J. Wise, and J. S. Terhune. "Saprolegniasis (Winter Fungus) and Branchiomycosis of Commercially Cultured Channel Catfish." *Southern Reginal Aquaculture Center (SRAC) Publication* No. 4700(2003).

21. Engle, C. "The Evolution of Farm Management, Production Efficiencies, and Current Challenges to Catfish Production in the United States." *Aquaculture Economics and Management* 7, no. (1/2)(2003): 67-84.

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

23. Greene, W. H. *Econometric Analysis*. 4th ed. Upper Saddle River, New Jersy: Prentice-Hall, Inc., 2000.

24. Hatch, U., and H. Kinnucan. *Aquaculture: Models and Economics*. Boulder Oxford: Westview Press, Inc., 1993.

25. Hawke, J. P., et al. "ESC - Enteric Septicemia of Catfish." *Southern Regional Aquaculture Center (SRAC) Publication* No. 477(1998)

26. Jensen, J., and J. Crews. "Channel Catfish Production in Ponds." Alabama Cooperative Extension System Publication ANR-195 (1997).

27. Kazmierczak, J., R. F., and P. Soto. "Stochastic Economic Variables and Their Effect on Net Returns to Channel Catfish Production." *Aquaculture Economics and Management* 5, no. 1/2(2001): 15-36.

28. Ker, A. P. “Nonparametric Estimation of Possibly Similar Densities.” Working paper, Department of Agricultural and Resource Economics, University of Arizona, 2002.

29. Ker, A. P. "Rating and Yield Predicting Procedures for the Group Risk Federal Crop Insurance Program: A Nonparametric Approach." PhD dissertation, North Carolina State University, March 1996.

30. Ker, A. P., and A. T. Ergun “Empirical Bayes Nonparametric Kernel Density Estimation.” Working paper, Department of Agricultural & Resource Economics, University of Arizona, 2003.

31. Ker, A. P., and K. Coble. "Modeling Conditional Yield Densities." *American Journal of Agricultural Economics* 85, no. 2(2003): 291-304.

32. Ker, A. P., and K. H. Coble. "On Choosing a Base Coverage Level for Multiple Peril Crop Insurance Contracts." *Journal of Agricultural and Resource Economics* 23, no. 2(1998): 427-444.

33. Ker, A. P., and B. K. Goodwin. "Nonparametric Estimation of Crop Insurance Rates Revisited." *American Journal of Agricultural Economics* 83(2000): 463-478.

34. Ker, A. P., and P. McGowan. "Weather-Based Adverse Selection and the U. S. Crop Insurance Program: The Private Insurance Company Perspective." *Journal of Agricultural and Resource Economics* 25, no. 2(2000): 386-410.

35. Khoo, L., and W. Holmes (2003) Visceral Toxicosis Of Catfish: An Emerging Disease.

36. Lewis, G., and J. L. Shelton. "Channel Catfish Production." Aquaculture Technical Series, Cooperative Extension Service, The University of Georgia College of Agricultural and Environmental Sciences, 1994.

37. Littauer, G. A., et al. "Control of Bird Predation at Aquaculture Facilities: Strategies and Cost Estimates." Southern Regional Aquaculture Center Publication N0. 402.

38. Mitchell, A. J., R. M. Durborow, and M. D. Crosby. "Proliferative Gill Disease (Hamburger Gill Diesease)." *Southern Reginal Aquaculture Center (SRAC) Publication* No. 475(1998).

39. Morris, J. E. "Pond Culture of Channel Catfish in the North Central Region." *North Central Regional Extension Fact Sheet Series* No. 106(1993).

40. Robinson, E. H., and J. L. Avery. "Farm-Raised Catfish." Fact Sheet 005. Thad Cochran National Warmwater Aquaculture Center, 2002.

41. Sam, A. G. “Nonparametric Regression of Possibly Similar Curves” MS thesis, Department of Agricultural and Resource Economics, University of Arizona, 2002

42. Shaik, S. "Aquaculture Insurance: Issues, Policy Design, and Potential Research." White Paper. National Risk Management Feasibility Program for Aquaculture.

43. Terhune, J. S., et al. "Infestations of the Trematode Bolbophorus sp. in Channel Catfish." *Southern Reginal Aquaculture Center (SRAC) Publication* No. 1801(2003).

44. The Catfish Institute Website: www.catfishinstitute.com (accessed in July 2004).

45. Tucker, C. S. "Characterization and Management of Effluents from Aquaculture Ponds in the Southeastern United States." *Southern Reginal Aquaculture Center (SRAC) Publication* No. 470(1999).

46. Tucker, C. S., and J. A. Hargreaves. "Effluents From Channel Catfish Aquaculture Ponds." Mississippi State University.

47. Tucker, C., J. Avery, C. Engle, and A. Goodwin. “Industry Profile: Pond-Raised Channel Catfish.” A review developed for the National Risk Management Feasibility Program for Aquaculture, Department of Agricultural Economics, Mississippi State University, May 2004.

48. Wellborn, T. L. "Channel Catfish Life History and Biology." *Southern Reginal Aquaculture Center (SRAC) Publication* No. 180(1988).

49. Whitis, G. N. "Watershed Fish Production Ponds Guide to Site Selection and Construction." *Southern Reginal Aquaculture Center (SRAC) Publication* No. 102(2002).

50. Wise, D., P. R. Waterstrat, and R. Hopper. "Winter Kill in Channel Catfish." *Mississippi State University Extension Service Information Sheet 1392*.