Farmer Level Yeild Prediction Based On Credibility Theory

by

Leiguang Chen

A research paper
presented to the University of Waterloo
in partial fulfillment of the
requirement for the degree of
Master of Mathematics
in
Computational Mathematics

Supervisor: Prof. Kens Seng Tan

Waterloo, Ontario, Canada, 2015

I hereby declare that I am the sole author of this report. This is a true copy of the report, including any required final revisions, as accepted by my examiners.

I understand that my report may be made electronically available to the public.

Abstract

In America, farmer level yield prediction is required for fund allocation which is an important decision for crop reinsurance companies. One important property of farmer level yield is that the information of individual risk is limited while the information of collective even sub collective risk is extensive. The credibility theory which resides on both individual risk and collective risk is widely used because of the heterogeneous of policy holders [3].

This paper provides a comparison between three credibility measures and a statistic model for forecasting farmer level yield. The demonstration is operated on a proprietary and detailed data set representing a large portion of the actual farmer level experience in the U.S. crop reinsurance program. The results show that hierarchical credibility model which concentrates more on the differences between policy holders performs the best.

Acknowledgements

I would like to thank all the little people who made this possible.

Table of Contents

List of Tables

List of Figures

Chapter 1

Introduction

Crop industry is growing fast and highly profitable in America. Since 1995, the government required farmers to purchase multiple peril crop insurance protection (MPCIP) in order to participate in any subsidized farm programs offered by the government. The premium grows from 2 billion dollars to 12 billion dollars over the last two decades. Crop reinsurance companies purchase the insurance in order to transfer risk from insurance companies and receive premium from the insurance companies for taking on their risk. One major goal for reinsurance companies is to allocate the premium into different funds based on their risk levels.There are five steps to reach this goal. First we should predict the yield of each farmer. Secondly we use the correlation parameter disclosed by RMA (Risk Management Agency) and the predicted yield to simulate the harvest price of the next year. After the simulation, next step is to calculate the loss ratio of each farmer. The last step is to allocate the funds based on the loss ratios. Among these five steps, the most important one is to predict the farmer level yield.

There is a proliferation of literature on crop yield prediction. Most of previous studies can be classified into parametric studies and nonparametric studies. In parametric studies, Just and Weninger discussed the normality of the distribution of farmer level yield [8]. After that, many different distributions have been employed by many researchers for prediction purpose, for instance, normal distribution [2], gamma distribution [6], beta distribution [10] and other distributions. Gallagher also tested the hypothesis that National Average Corn yields were skewed with a relatively high chance of occasional low yields [6]. In nonparametric and semi-parametric studies, Goodwin and Ker used different kernel methods to estimate county level crop yield distributions [7]. Another approach is to model the central tendency of distributions with a stochastic trend model and allowing for non-normality errors [9]. Woodard and Sherrick also evaluated the performance of parametric

models and nonparametric models in different frameworks [11].There are also many studies about the relationship between the climate condition and yield prediction [4] [5].

However, the data used in the studies mentioned above are either county level yield which has more than 40 successive data points or farmer level yield which is prepared for research with more than 20 data points. For a reinsurance company, the farmer level yield data is limited. Usually there are only 5 to 10 discontinuous years of data because of the rotation. Then all the studies above are not applicable in this circumstance. To make the farmer level yield prediction for fund allocation possible, some new methods should be developed.

In this project, a new definition MOD is introduced to make the discontinuous data comparable. We define MOD as the ratio of farmer level yield to the corresponding county level yield. A specific distribution is assumed to fit this new variable. Since the number of data points for each farmer is quite limited, credibility theory models are used to adjust the mean of the distribution. Then a comparison is made to check whether there is an improvement using the credibility theory models. By this means, we convert a farmer level yield prediction into two problems. First predict county level yield data based on better information. Secondly estimate MOD distribution for each farmer. The final step is to multiple MOD with predicted county yield to get the prediction of farmer level yield for next year.

The rest of the paper is organized as follows: Chapter 2 provides some background knowledge on the assumptions of this project and credibility theory. Chapter 3 describes Bülmann-Straub model and hierarchical credibility model. Chapter 4 shows the numerical results of different approaches and compare their performances. Chapter 5 summarizes the paper with further discussion.

Chapter 2

Data Analysis and Credibility Theory

In this chapter, a new definition which helps to predict the farmer level yield is introduced. Firstly, we make a discussion about the trends of county yield historical data and the relationship between county level and farmer level yield data. Since the data have two dimensions: time and space. Also the fact that we only have 5 to 10 data points for each farmer makes it impossible for prediction. Therefore, a connection should be set up to make all the data points become comparable. Secondly, an introduction of credibility theory which can justify the estimation of data is provided .

2.1 Data Analysis

2.1.1 County level and Farm level Historical Yield Data

Both county level and farmer level historical yield data of policy holders in America is provided for this project and we use corn as the main crop for analysis.The data set provided contain a large portion of the actual farm level experience in the U.S. crop insurance program. When it comes to county level data, there are records of average yields from 1970 to 2013 for thousands of counties. Because of the technique improvement and other influences, there is an obvious increasing trend in county level yield data. The graph Figure 2.1 can make a clear explanation. The ADF test shown in Table 2.1 also indicates that there exist unit root in the county level yield data.

As shown above, there is an obvious increasing trend in the county level yield data. There must exist an increasing trend in farmer level yield since county level yield data is

Table 2.1: Example of ADF Test of County-level Yield Data

Results	County1 S1	County67 S21	County51 S27	County131 S39
Dickey Fuller	-2.4542	-1.8499	-0.293	-2.0087
P-value	0.394	0.6332	0.1663	0.5703
Unit Root	Exist	Exist	Exist	Exist

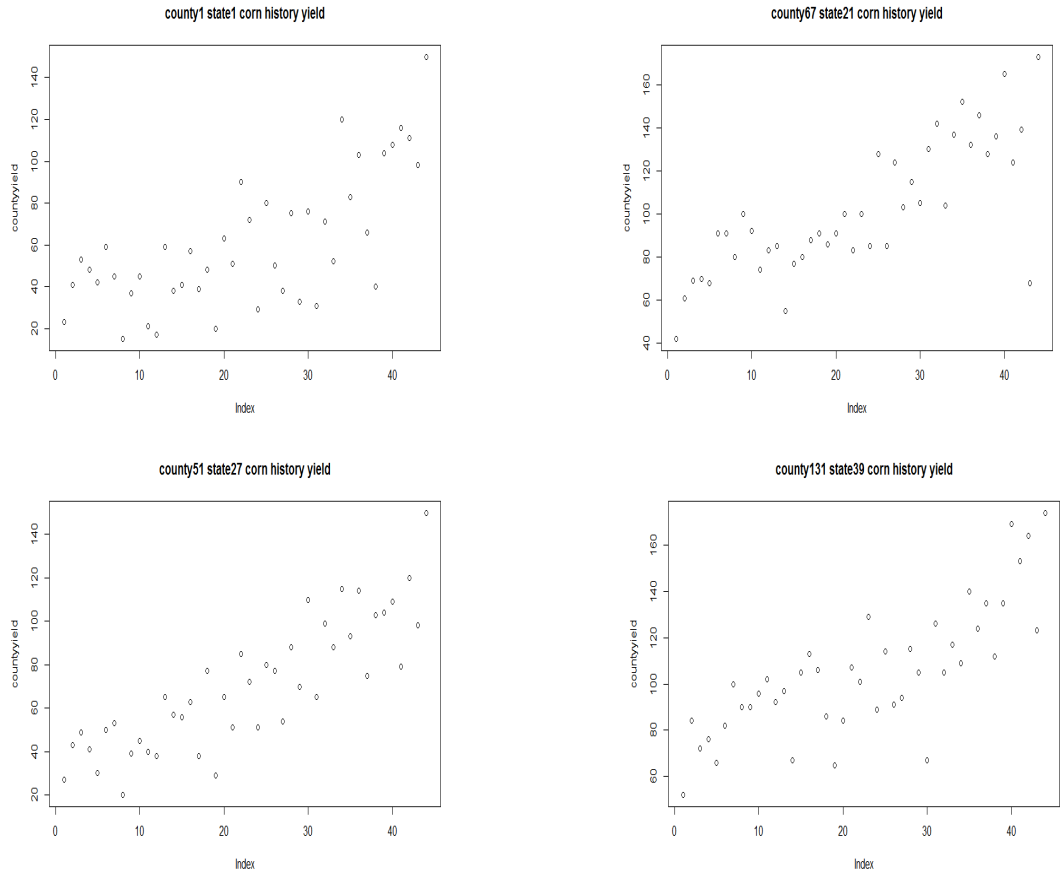


Figure 2.1: Example of Increasing Trend of Count-level Yield Data

Table 2.2: Example of Farmer-level Yield

Farmer	County Code	State Code	Year	Annual Yield
1	67	21	1996	102
1	67	21	1997	103
1	67	21	1998	127
1	67	21	2001	100
1	67	21	2007	152
1	67	21	2008	140
1	67	21	2012	87
2	51	27	2004	130
2	51	27	2005	196
2	51	27	2006	123
2	51	27	2007	168
2	51	27	2008	128
2	51	27	2009	32
2	51	27	2010	105
2	51	27	2011	50
2	51	27	2012	137

just the grouped averages of farmer level yield data. Then there are two dimensions of data of every farmer: time and space. Data structure can also be showed in Table 2.2

2.1.2 Assumptions of The Distribution of MOD

From Table 2.2, we can see that the data for every farmer is quite limited and discontinuous. Also even in the same year, farmers in different areas are not comparable. Because of these properties, a new definition called MOD is developed.

Definition 2.1.1 *MOD*

A MOD is a ratio between the farmer annual yield of a certain year and the corresponding county yield of that specific year.

$$MOD = \frac{farmer\ level\ yield\ of producer\ i\ in\ year\ t}{county\ level\ yield\ which\ contain\ producer\ i\ in\ year\ t} \tag{2.1}$$

After transferring farmer level yield to MOD, the increasing trend because of time disappears. Since a MOD is a ratio between two kinds of yields in the same area, it makes MOD comparable among the whole country. MOD can also be considered as a feature of farmers. A farmer with MOD always greater than 1 has a talent on farming. A farmer, on the contrary, with MOD always smaller than 1 may not perform as well as his neighbours.

Table 2.3: Example of Producer's MOD

Farmer	County Code	State Code	Year	MOD
1	67	21	1996	0.953
1	67	21	1997	0.837
1	67	21	1998	0.984
1	67	21	2001	0.636
1	67	21	2007	0.944
1	67	21	2008	0.833
1	67	21	2012	1.705
2	51	27	2004	1.015
2	51	27	2005	0.955
2	51	27	2006	0.129
2	51	27	2007	1.033
2	51	27	2008	0.994
2	51	27	2009	0.242
2	51	27	2010	0.75
2	51	27	2011	0.370
2	51	27	2012	0.792

From Table 2.3, it can be shown clearly that MOD are non-negative numbers around 1. Since the MOD could be a measure of performance of farmers, one simplest assumption of MOD is truncated normal distribution. Considering the existence of disasters which may cause the farmer even the whole county have very few production in a specific year, we assume that there is a probability the MOD equaling to 0. Then the distribution of a MOD is a mixed distribution with exponential and truncated normal. It could be considered as zero case (exponential distribution) and nonzero case (truncated normal distribution). Based on the study of GC, the mean of exponential is 0.04. The mean and variance of truncated normal should be estimated depending on the MOD of each farmer. Also if a farmer has MOD always smaller than 1 or has the history of zero yield, the probability to be the zero case (exponential distribution) will increase. If a farmer has MOD always greater than 1 or has no history of zero yield, the probability to be the

zero case (exponential distribution) will decrease. To estimate the probability of zero or nonzero case, logistic regression is employed. Because this project doesn't concentrate on the estimation of logistic regression, the discussion of it is escaped.

2.2 Credibility Theory

The basic assumption of insurance is that all the policy holders are identical independent distributed. Then the MOD of each farmer can be considered as realizations of a random variable from a particular distribution. But in practice, "there are no homogeneous risk classes in insurance" [3]. There is a number of characteristics would perhaps be useful in determining the quality of a farmer's yield. For example, the quality of the farm land, the amount of time a farmer spend in farming and the climate condition of a certain area will all influence the production of a farmer. This is why we need to treat the policy holders are not identical to find the fair distribution of every farm.

2.2.1 Assumptions of Simple Credibility Theory

Let random variable X_i are, conditional on $\Theta = \theta$, independent with the same distribution function F_θ with the conditional moments

- $\Theta = \theta$ is a random variable which represents the property of farmers. The random variable $X_i (i = 1, 2, \ldots, n)$ are conditional independent with the same θ which is a realization of Θ. X_i follows the distribution function U_θ. In this project, it is assumed that all the observations are in the same class.

$$\mu(\theta) = E[X_j|\Theta = \theta],$$

$$\sigma^2(\theta) = Var[X_j|\Theta = \theta].$$

- Θ is a random variable with distribution $\Phi(\theta)$.

From this model, it can be shown that

$$P^{ind} = \mu(\Theta) = E[X_{n+1}|\Theta],$$
$$P^{coll} = \mu_0 = \int_\Theta \mu(\theta) d\Phi(\theta).$$

2.2.2 Derivation of Credibility Estimators

The derivations from page 8 to page 9 are from book [3]:
Let $\widehat{\mu(\Theta)}$ denote the best estimator within the class. Then by definition, $\widehat{\mu(\Theta)}$ has to be the form

$$\widehat{\mu(\Theta)} = \widehat{a_0} + \sum_{i=1}^{n} \widehat{a_i} X_i$$

where the real coefficients $\widehat{a_0}, \widehat{a_2}, \ldots, \widehat{a_n}$ need to solve

$$E[(\mu(\Theta) - \widehat{a_0} - \sum_{i=1}^{n} \widehat{a_i} X_i)^2] = \min_{a_0, a_1, \ldots, a_n \in R} E[(\mu(\Theta) - a_0 - \sum_{i=1}^{n} \widehat{a_i} X_i)^2]$$

Since the probability distribution of $X_1, X_2, \ldots, X_n$ is invariant under permutations of X_j and $\widehat{\mu(\Theta)}$ is uniquely defined it must hold that

$$\widehat{a_n} = \widehat{a_n} = \ldots = \widehat{a_n},$$

i.e. the estimator $\mu(\Theta)$ has the form

$$\widehat{a_n} = \widehat{a_n} = \ldots = \widehat{a_n} + \hat{b}\overline{X},$$

where

$$\overline{X} = \frac{1}{n} \sum_{i=1}^{n} X_i$$

and where $\hat{a}$ and $\hat{b}$ are the solutions of the minimizing problem

$$E[(\mu(\Theta) - \widehat{a_0} - \sum_{i=1}^{n} \widehat{a_i} X_i)^2] = \min_{a_0, a_1, \ldots, a_n \in R} E[(\mu(\Theta) - a_0 - \sum_{i=1}^{n} \widehat{a_i} X_i)^2].$$

Taking partial derivatives with respect to a, resp.b, we get

$$E[\mu(\Theta) - a - b\overline{X}] = 0,$$

$$Cov(\overline{X}, \mu(\Theta)) - bVar(\overline{X}) = 0.$$

Based on the conditional independent model assumptions, it comes to

$$Cov(\overline{X}, \mu(\Theta)) = Var(\mu(\Theta)) = \tau^2,$$

$$Var(X) = \frac{E[\sigma^2(\Theta)]}{n} + Var(\mu(\Theta)) = \frac{\sigma^2}{n} + \tau^2.$$

from which it can be derived

$$b = \frac{\tau^2}{\tau^2 + \frac{\sigma^2}{n}} = \frac{n}{n + \frac{\sigma^2}{\tau^2}},$$

$$a = (1 - b)\mu_0.$$

From the derivation above, the credibility estimator for the simplest credibility model is give by

$$\widehat{\mu(\Theta)} = \alpha \widehat{X} + (1 - \alpha)\mu_0, \tag{2.2}$$

where

$$\begin{aligned} \mu_0 &= E[\mu(\Theta)], \\ \alpha &= \frac{n}{n + \frac{\sigma^2}{\tau^2}}. \end{aligned} \tag{2.3}$$

The quotient $\kappa = \frac{\sigma^2}{\tau^2}$ is called the credibility coefficient [1], which can also be written as $\kappa = (\frac{\sigma}{\mu_0})^2(\frac{\tau}{\mu_0})^{-2}$. Note that $\frac{\tau}{\mu_0}$ is the coefficient of variance of $\mu(\Theta)$, which is a good measure of the heterogeneity of the farmers. There are several circumstances that may cause the increment of α.

- the number of years n increases,
- the heterogeneity of the farmers (as measured by the coefficient of variation $\frac{\tau}{\mu_0}$) increases,
- the within risk variability (as measured by $\frac{\sigma}{\mu_0}$) decreases.

Chapter 3

Methodology

While the assumptions mentioned in the simplest model are always violated in the real world. On the basis of certain characteristics, the farmers have been grouped into various risk classes which means considering all the farmers are from the same distribution is not appropriate. Also, the size of each classes is different which also makes the analysis more difficult. To make the simplest model more suitable for the real world, two new credibility methods is developed. Both Bühlmann-Straub model and Hierarchical credibility model class the farmers into different levels and pay more efforts to distinguish the differences between farmers.

3.1 Bühlmann-Straub Model

"Bühlmann-Straub Model is developed by Bühlmann and Straub in 1970. It is still by far the most used and most important credibility model for insurance practice" [3].

3.1.1 Model Assumptions

Since there are differences between each farmers, it is appropriate to class the observations by farmers. Let X_{ij} represents the jth observations of MOD of farmer i, and ω_{ij} represent the associated volume measures of the jth observation of farmer i. Since there are only one observation in each year, ω_{ij} could also be considered as 1 for every farmer. The assumptions of Bühlmann-Straub model are as follows:

- The farmer i is characterized by an individual risk profile θ_i, which is itself the realization of a random variable Θ_i. Thus, conditionally given Θ_i, the $\{X_{ij} : j = 1, 2, \ldots, n\}$ are independent with

$$E[X_{ij}|\Theta_i] = \mu(\Theta_i)$$

$$Var[X_{ij}|\Theta_i] = \frac{\sigma^2(\Theta_i)}{\omega_{ij}}$$

- The pairs $(\Theta_1, X_1), (\Theta_2, X_2), \ldots$ are independent, and $\Theta_1, \Theta_2, \ldots$ are independent and identically distributed.

- The "true" individual claims ratio $\mu(\Theta_i)$ is constant over time.

The observation years n may also vary between farmers. This could be formally expressed by setting $\omega_{ij} = 0$ for non-observed years. From the model assumption it can be shown that the Bühlmann-Straub model is a two-urn model. From the first urn, we draw farmer profile Θ_i ,which determines the "content" of the second urn. In the second step, a random variable X_{ij} is drawn from the second urn. In this project, we assume all the farmers are in the same class, while the observations are categorised by farmers. This is modelled by the fact that the risk profiles Θ_i are all drawn from the same urn.

The notations are as follows:

- $\mu(\Theta_i)$: MOD of each farmer
- $\sigma^2(\Theta_i)$:variance within each farmer
- μ_0: collective MOD
- σ^2: average variance within each farmer
- τ^2: variance between MOD of farmers

3.1.2 Model Derivation

The derivations from page 13 to page 14 are from [3].
Our goal is to estimate for farmer i its MOD $\mu(\Theta_i)$. The available data is $D = \{X_i : 1, 2, \ldots, I\}$, where $X_i = (X_{i1}, X_{i2}, \ldots, X_{in},)'$ is the observation vector of farmer i. It is easy to see that the credibility estimator for $\mu(\Theta_i)$ only depends on the observations from ith risk. Let

$$\widehat{\mu(\Theta_i)} = a_{i0} + \sum_j a_{ij} X_{ij}$$

be the credibility estimator based on X_i. Because of the independence of the farmers, it follows that for $k \neq i$ and all l

$$Cov(\widehat{\mu(\Theta_i)}, X_{kl}) = Cov(\mu(\Theta_i), X_{kl}) = 0.$$

Let

$$X_i = \sum_j \frac{\omega_{ij}}{\omega_{i.}} X_{ij}.$$

For X_i it has

$$E[X_i|\Theta_i] = \mu(\Theta_i),$$

$$Var[X_i|\Theta_i] = \sum_j \frac{\omega_{ij}}{\omega_{i.}}^2 Var[X_{ij}|\Theta_i] = \frac{\sigma^2(\Theta_i)}{\omega_{i.}}$$

Now the credibility estimator can be derived based on X_i and then it can be shown that this is also the credibility estimators based on all data. Referred to [3], the credibility estimator must be of the form as following:

$$\widehat{\mu(\Theta_i)} = \alpha_i X_i + (1 - \alpha_i)\mu_0,$$
$$Cov(\widehat{\mu(\Theta_i)}, X_i) = \alpha_i Cov(X_i, X_i) = Cov(\mu(\Theta_i), X_i).$$

From the fact that

$$
\begin{aligned}
Var[X_i] &= E[Var[X_i|\Theta_i]] + Var[E[X_i|\Theta_i]] \\
&= \frac{\sigma^2}{\omega_{i\cdot}} + \tau^2,
\end{aligned}
$$

$$
\begin{aligned}
Cov(\mu(\Theta_i), X_i) &= E[Cov(\mu(\Theta_i), X_i|\Theta_i)] + Cov(\mu(\Theta_i), E[X_i|\Theta_i]) \\
&= 0 + Var[\mu(\Theta_i)] \\
&= \tau^2,
\end{aligned}
$$

it follows that

$$
\alpha_i = \frac{\tau^2}{\frac{\sigma^2}{\omega_{i\cdot}} + \tau^2} = \frac{\omega_{i\cdot}}{\frac{\sigma^2}{\tau^2}} \tag{3.1}
$$

3.1.3 Credibility Estimators of Bühlmann-Straub Model

From the discussion above, we can conclude that the credibility estimators for Bühlmann-Straub Model is as follows [3]:

Bühlmann-Straub Credibility Estimators

$$
\widehat{\mu(\Theta_i)} = \alpha_i X_i + (1-\alpha_i)\mu_0 = \mu_0 + \alpha_i(X_i - \mu_0)
$$

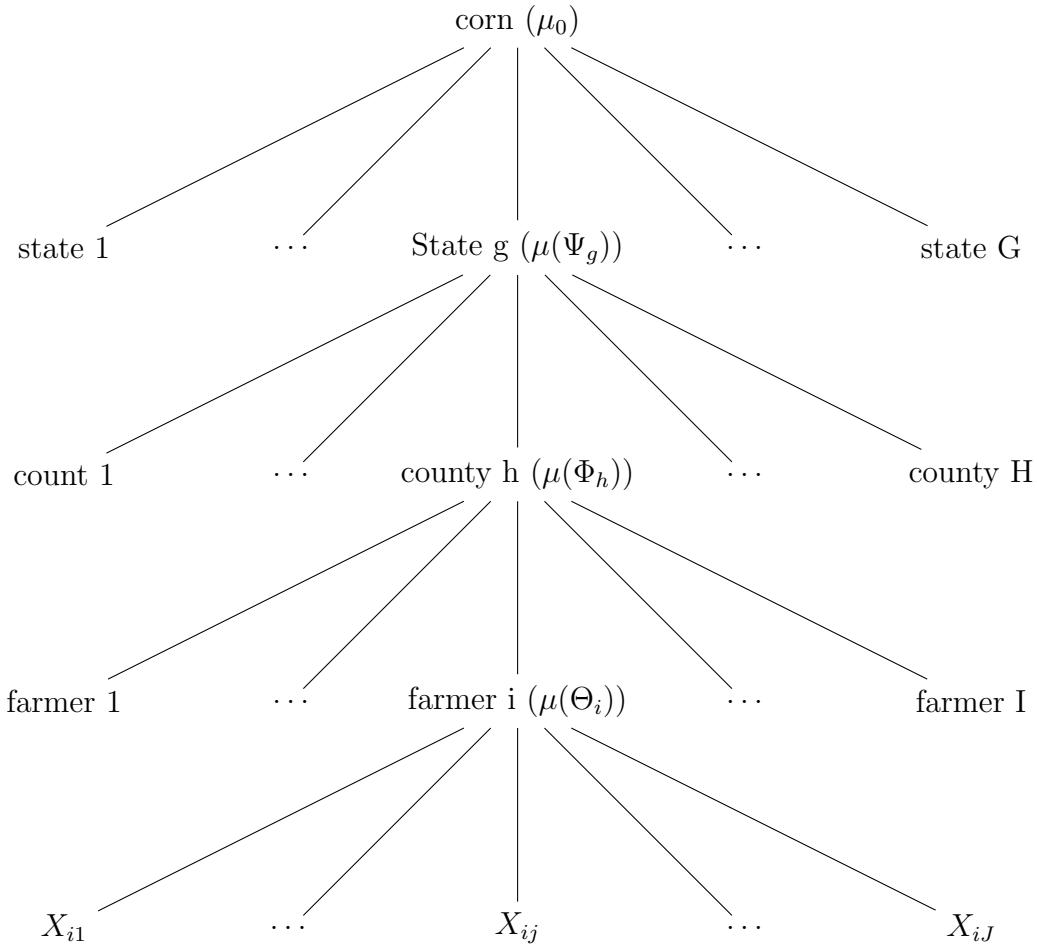
$$
where X_i = \sum_j \frac{\omega_{ij}}{\omega_{i\cdot}} X_{ij}
$$

$$
\omega_{i\cdot} = \sum_j \omega_{ij}
$$

$$
\alpha_i = \frac{\omega i\cdot}{\omega_i + \frac{\sigma^2}{\tau^2}} = \frac{\omega_{i\cdot}}{\omega_{i\cdot} + \kappa}
$$

3.2 Hierarchical Credibility Model

In practice, there are always hierarchical structures. Just as this project, the farmers are classified automatically according to their counties, counties are grouped together into states, states into crop types which together make the total of the crop industry. This is why hierarchical credibility model is introduced into this project. The structure of the model can be visualized as follows:



Tree Structure of Hierarchical Credibility Model

The notation are as follows [3]:

- $\Phi(\Psi_g)$ is the set of Φ's, that stem from Ψ_g
- $D(\Phi_h)$ is the set of observations X_{ij}, that stem from Φ_h

3.2.1 Model Assumption

- The random variables $\Psi_g(g = 1, 2, \ldots, |G|)$ are independent and indentically distributed $(i.i.d)$ with density $r_3(\psi)$
- Given Ψ_g the random variables $\Phi_h \in \Phi(\Psi_g)$ are $i.i.d$ with the conditional density $r_2(\psi|\Psi_g)$
- Given Φ_h the random variables $\Theta_i \in \Theta(\Phi_h)$ are $i.i.d$ with the conditional density $r_1(\theta|\Phi_h)$
- Given Θ_i the observations $X_{ij} \in D(\Theta_i)$ are conditionally independent with densities $r_0(x|\Theta_i, \omega_{ij})$, for which

$$E[X_{ij}|\Theta_i] = \mu(\Theta_i)$$

$$Var[X_{ij}|\Theta_i] = \frac{\sigma^2(\Theta_i)}{\omega_{ij}}$$

where ω_{ij} are the known weights

There is an obvious similarity between hierarchical credibility model and Bühlmann-Straub model. The later one also has a tree structure only with fever levels. Hierarchical models of higher orders are therefore nothing more than generalizations of the Bühlmann-Straub model to an increased number of levels.

3.2.2 Further Discussion On Model Assumption

Our goal is to find the credibility estimators $\widehat{\mu(\Theta_i)}$ for every farmer $\mu(\Theta_i)$ for $i = 1, 2, \ldots, I$. It is necessary to find the credibility estimators $\widehat{\mu(\Phi_h)}$, $h = 1, 2, \ldots, H$ and $\widehat{\mu(\Psi_g)}$, $g = 1, 2, \ldots, G$ and also $\widehat{\mu}_0$. The following discussions from page 17 to 18 are provided by [3].

We use the following notation:

$$\mu_0 := E[X_{ij}], (collective\ MOD)$$

$$\mu(\Psi_g) = E[X_{ij}|\Psi_g], \quad where\ X_{ij} \in D(\Psi_g)$$

$$\mu(\Phi_h) = E[X_{ij}|\Phi_h], \quad where\ X_{ij} \in D(\Phi_h)$$

$$\mu(\Theta_i) = E[X_{ij}|\Theta_i], \quad where\ X_{ij} \in D(\Theta_i)$$

It holds that

$$\mu_0 = E[\mu(\Psi_g)]$$

$$\mu(\Psi_g) = E[\mu(\Phi_h)|\Psi_g], \quad where\ \Phi_h \in \Phi(\Psi_g)$$

$$\mu(\Phi_h) = E[\mu(\Theta_i)|\Phi_h], \quad where\ \Theta_i \in \Theta(_h)$$

It follows easily from the model assumptions and properties of the conditional expectation, as is illustrated in the following for $\mu(\Phi_h)$:

$$\begin{aligned}\mu(\Phi_h) = E[X_{ij}|\Phi_h] &= E[E[X_{ij}|\Theta_i, \Phi_h]|\Phi_h] \\ &= E[E[X_{ij}|\Theta_i]|\Phi_h] \\ &= E[\mu(\Theta_i|\Phi_h)]\end{aligned}$$

Next it comes to the structural parameters of the hierarchical credibility model. These are the priori expected value and variance.

$$at \;\; level \;\; 0 \quad \sigma^2 = E[\sigma^2(\Theta_i)]$$

$$at \;\; level \;\; 1 \quad \tau_1^2 = E[Var[\mu(\Theta_i)|\Phi_h]] = E[\tau_1^2(\Phi_h)]$$

$$at \;\; level \;\; 2 \quad \tau_2^2 = E[Var[\mu(\Phi_h)|\Psi_g]] = E[\tau_2^2(\Psi_g)]$$

$$at \;\; level \;\; 3 \quad \tau_3^2 = Var[\mu(\Psi_g)]$$

It follows directely from the properties of the conditional expectation that

$$\sigma^2 = E[\omega_{ij}(X_{ij} - \mu(\Theta_i)))^2]$$

$$\tau_1^2 = E[(\mu(\Theta_i) - \mu(\Phi_h))^2]$$

$$\tau_2^2 = E[((\Phi_h) - \mu(\Psi_g))^2]$$

$$\tau_3^2 = E[(\mu(\Psi_g) - \mu_0)^2]$$

It is easy to show that the following equations hold for the unconditional variance

$$Var[\mu(\Theta_i)] = \tau_1^2 + \tau_2^2 + \tau_3^2$$

$$Var[\mu(\Phi_h)] = \tau_2^2 + \tau_3^2.$$

3.2.3 Model Derivation

It is discussed in [3] that all the credibility estimators can be understood as linear combinations of the "tree father" μ_0 and all "descendent data" X_{ij}. It suggest that it must be the form

$$\widehat{\mu(\Theta_i)}' = \alpha_i^{(1)} B_i^{(1)} + (1 - \alpha_i^{(1)})\mu(\Phi_h)$$

where $B_i^{(1)}$ are the compressed data from $D(\Theta_i)$ and $\alpha_i^{(1)}$ are suitable credibility weights.

Then we get

$$\widehat{\mu(\Theta_i)} = \alpha_i^{(1)} B_i^{(1)} + (1 - \alpha_i^{(1)})\widehat{\mu(\Phi_h)} \tag{3.2}$$

In order to determine $\widehat{\mu(\Phi_h)}$, the same calculation could be repeated into the higher level.

We get the following:

$$\begin{aligned} \widehat{\mu(\Phi_h)} &= \alpha_h^{(2)} B_h^{(2)} + (1 - \alpha_h^{(2)})\widehat{\mu(\Psi_g)} \\ \widehat{\mu(\Psi_g)} &= \alpha_g^{(3)} B_g^{(3)} + (1 - \alpha_g^{(3)})\mu_0 \end{aligned} \tag{3.3}$$

Based on the lemma provided by [3], we can finally prove the results as follows:

$$\widehat{\mu(\Phi_h)}' = \alpha_h^{(2)} B_h^{(2)} + (1 - \alpha_h^{(2)})\mu(\Psi_g)$$

$$where \quad \alpha_h^{(2)} = \frac{\tau_2^2}{\tau_2^2 + E[(\mu(\Phi_h) - B_h^{(2)})^2]}$$

$$B_g^{(3)} = \sum_{h \in H_g} \frac{\alpha_h^{(2)}}{\omega_g^{(3)}} B_h^{(2)}$$

$$where \quad H_g = \{h : \Phi_h \in \Phi(\Psi_g)\}$$

$$\omega_g^{(3)} = \sum_{h \in H_g} \alpha_h^{(2)}$$

$$E[(B_g^{(3)} - \mu(\Psi_g))^2] = \frac{\tau_2^2}{\omega_g^{(3)}}$$

3.2.4 Credibility Estimator of Hierarchical Model

From the discussion above, we can conclude that the credibility estimator for hierarchical model is as follows [3]:

Hierarchical Credibility Estimators

$$\widehat{\mu(\Psi_g)} = \alpha_g^{(3)} B_g^{(3)} + (1 - \alpha_g^{(3)})\mu_0$$
$$\widehat{\mu(\Phi_h)} = \alpha_h^{(2)} B_h^{(2)} + (1 - \alpha_h^{(2)})\widehat{\mu(\Psi_g)}, \quad \Phi_g \in \Phi(\Psi_g)$$
$$\widehat{\mu(\Theta_i)} = \alpha_i^{(1)} B_i^{(1)} + (1 - \alpha_i^{(1)})\widehat{\mu(\Phi_h)}, \quad \Theta_i \in \Theta(\Phi_h)$$

The parameters used above can be estimated as follows [3]:

Parameters Estimation

$$B_i^{(1)} = \sum_j \frac{\omega_{ij}}{\omega_{i.}} X_i j$$

$$where\ \omega_{i.} = \sum_j\ \omega_{ij}$$

$$\alpha_i^{(1)} = \frac{\omega_{i.}}{\omega_{i.} + \frac{\sigma^2}{\tau_1^2}}$$

$$B_h^{(2)} = \sum_{i \in I_h} \frac{\alpha_i^{(1)}}{\omega_h^{(2)}} B_i^{(1)}$$

$$where\ I_h = \{i : \Theta_i \in \Theta(\Phi_h)\},\ \omega_h^{(2)} = \sum_{i \in I_h} \alpha_i^{(1)}$$

$$\alpha_h^{(2)} = \frac{\omega_h^{(2)}}{\omega_h^{(2)} + \frac{\tau_1^2}{\tau_2^2}}$$

$$B_g^{(3)} = \sum_{h \in H_g} \frac{\overset{(2)}{h}}{\omega_g^{(3)}} B_h^{(2)}$$

$$where\ H_g = \{h : \Phi_h \in \Phi(\Psi_g)\},\ \omega_g^{(3)} = \sum_{h \in H_g} \alpha_h^{(2)}$$

$$\alpha_g^{(3)} = \frac{\omega_g^{(3)}}{\omega_g^{(3)} + \frac{\tau_2^2}{\tau_3^2}}$$

$$\widehat{\mu_0} = \sum_g \frac{\alpha_g^{(3)}}{\omega^{(4)}} B_g^{(3)}$$

$$where\ \omega^{(4)} = \sum_g \alpha_g^{(3)}$$

Chapter 4

Numerical Results

In this chapter, numerical results of the two credibility models discussed in Chapter 3 will be presented. These two methods are applied to the data set which is provided. The original data set contains five files which include the information of policy holders. These files contain information collected by insurance companies approved by RMA. To collect together the information we need, we should select records from all the files. Table 4.1 shows the procedures of data selection.

4.1 Data Preprocessing And Cleaning

Table 4.1 can show the procedures of data selection.

As shown in Table 4.1 and Table 2.3, every farm land has a specific key which will match its owner, location, yield history and other features. After specify a piece of farm land, MOD observations can be calculated based on the information within 44 years. The distribution of all the MOD is shown in figure 4.1

From Figure4.1, we can also see that most of the MOD are located around 1 which means the performances of the most farmers are about average. But there is still a considerable amount MOD located far from 1. The largest most is about 30 which is not possible. After analyze the data deeply, we find some methods to clean up the extreme large MOD. The rules of cleaning are based on the property of data, which are as follows:

- Choose MOD only from *yield type code* A (A means the farmer has the historical records of this type of crop within recent 5 years).

Table 4.1: Data Selection Form

Data set	Connection Key	Feature
P10 records	AIP Policy Producer Key Policy Number	Local State Code
P14 records	AIP Policy Producer Key AIP Insurance In Force Key	Local County Code Commodity Code
P15 records	AIP Policy Producer Key AIP Insurance In Force Key AIP Yield Key	Basic Unit Number Type Code Practice Code
P15A records	AIP Policy Producer AIP Insurance In Force Key AIP Yield Key	AIP Yield History Key Yield Type Code Annual Yield Annual Production Yield Acreage
County History Yield	Local State Code Local County Code AIP Yield History Key	County Crop Yield

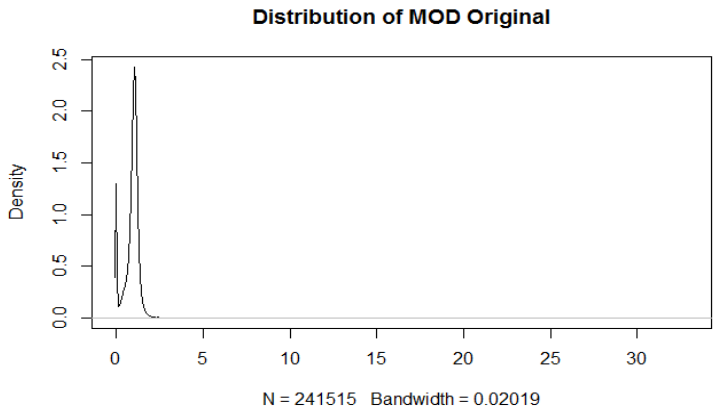


Figure 4.1: Distribution of Original MOD

- Delete the MOD if the corresponding *yield acreage* is 0 but *annual yield* is huge
- Delete the MOD if the corresponding *annual production* is 0 but *annual yield* is huge

Figure 4.2 shows the density distribution of MOD after cleaning up

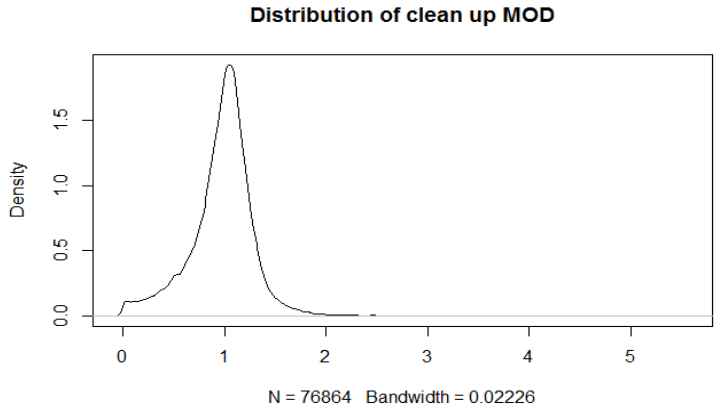


Figure 4.2: Distribution of Cleaned Up MOD

There are 0.8719976% of MOD greater than 2 and 0.1398661% of MOD greater than 3 which means the data is qualified for credibility calculation.

4.2 Results Comparison

Figure 4.3 shows the density distribution of $\mu(\Theta_i)$ calculated from different credibility models. We can see that the hierarchical model is more spread since it concentrates more on the differences between groups. The Bühlmann-Straub model is more centralized since it considers that all the farmers are from the same distribution. The original model provided by previous study is in the middle of other two models.

Because of the limited data size and discontinuity, likelihood function is used to compare the performance of these three models. The model which get the highest average realization value of likelihood function is the best. It is shown that the hierarchical model is the best with a highest likelihood score. So we can make conclusion that, in farmer level yield prediction, it is better to use hierarchical model to adjust the mean of MOD. After the

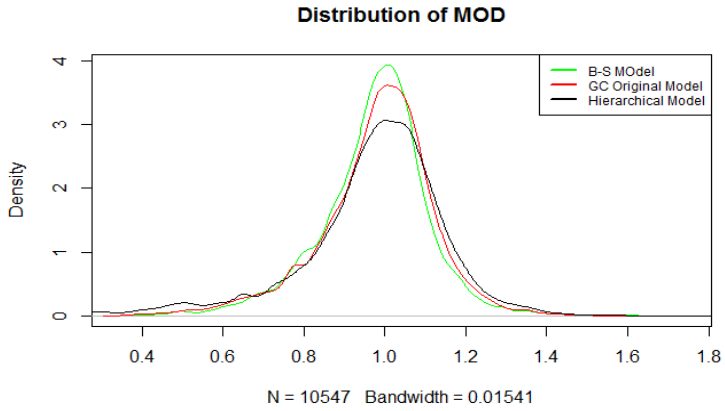


Figure 4.3: Distribution of Clean Up MOD

adjustment, we can use the adjusted mean to fit in truncated normal distribution for further study. The results is shown in table 4.2.

Table 4.2: Comparison of Model Performance

Method	Hierarchical	Bühlmann-Straub	Previous Study
likelihood score	1.573519	1.46594	1.491558

Chapter 5

Conclusion

This paper presents an application of credibility models solving real world reinsurance problems. Since the limitation of the data in the real world, a different approach for farmer level yield prediction is used. Two main discussed models have different assumptions which may influence the results of the estimation. Compared three models we already have, the hierarchical model performs the best. Which means the location of farm land do impact the production of crops.
Possible extension includes fitting the MOD with other distributions see whether mixed truncated normal and exponential is the most suitable distribution and finding a way to speed up the calculation when hierarchical model is used.

References

[1] Arthur L Bailey. A generalized theory of credibility. In *Proceedings of the Casualty Actuarial Society*, volume 32, pages 13–20, 1945.

[2] Ralph R Botts and James N Boles. Use of normal-curve theory in crop insurance ratemaking. *Journal of Farm Economics*, pages 733–740, 1958.

[3] Hans Bühlmann and Alois Gisler. *A course in credibility theory and its applications*. Springer Science & Business Media, 2006.

[4] AJ Challinor, JM Slingo, TR Wheeler, and FJ Doblas-Reyes. Probabilistic simulations of crop yield over western india using the demeter seasonal hindcast ensembles. *Tellus A*, 57(3):498–512, 2005.

[5] Sulochana Gadgil, PR Seshagiri Rao, and K Narahari Rao. Use of climate information for farm-level decision making: rainfed groundnut in southern india. *Agricultural Systems*, 74(3):431–457, 2002.

[6] Paul Gallagher. Us soybean yields: estimation and forecasting with nonsymmetric disturbances. *American Journal of Agricultural Economics*, 69(4):796–803, 1987.

[7] Barry K Goodwin and Alan P Ker. Nonparametric estimation of crop yield distributions: implications for rating group-risk crop insurance contracts. *American Journal of Agricultural Economics*, 80(1):139–153, 1998.

[8] Richard E Just and Quinn Weninger. Are crop yields normally distributed? *American Journal of Agricultural Economics*, 81(2):287–304, 1999.

[9] Charles B Moss and J Scott Shonkwiler. Estimating yield distributions with a stochastic trend and nonnormal errors. *American Journal of Agricultural Economics*, 75(4):1056–1062, 1993.

[10] Carl H Nelson and Paul V Preckel. The conditional beta distribution as a stochastic production function. *American Journal of Agricultural Economics*, 71(2):370–378, 1989.

[11] Joshua D Woodard and Bruce J Sherrick. Estimation of mixture models using cross-validation optimization: implications for crop yield distribution modeling. *American Journal of Agricultural Economics*, 93(4):968–982, 2011.