A Relational Data-Matching Model for Enhancing Individual Loss Experience: the Case of Crop Insurance

Lysa Porth* Ken Seng Tan† Wenjun Zhu‡

Abstract

The focus of this paper is on predictive analytics regarding data scarcity and credibility, which are major difficulties facing the agricultural insurance sector often due to limited loss experience data for those infrequent, but, extreme weather events. A new relational data-matching model is presented to predict individual farmer yields in the absence of farm-level data. The relational model defines a similarity measure based on an Euclidean distance metric that considers weather information, farm size, county size and the coefficient of variation of yield, to search for the most "similar" region in a different country to borrow individual loss experience data that is otherwise not available. Detailed farm-level and county-level corn yield data in Canada and the U.S. are used to empirically evaluate the proposed relational model. Compared to the benchmark model, the relational model developed in this paper has lower prediction error with smaller variation, and is also able to recover the actual premium rate more accurately. While the example in this paper uses historical data from Canada and the U.S. to illustrate the approach, the model may be extended to other countries, including developing countries, where farm-level data may not be available. This provides

*E-mail: lysa.porth@ad.umanitoba.ca. Warren Centre for Actuarial Studies and Research and Department of Agribusiness and Agricultural Economics, University of Manitoba, Canada.

†E-mail: lysa.porth@ad.umanitoba.ca. Department of Statistics and Actuarial Science, University of Waterloo.

‡E-mail: wjzhu@ntu.edu.sg. Nanyang Business School, Nanyang Technological University.

a new approach for insurers, reinsurers, and governments to enhance risk modelling, pricing, and the development of new insurance programs.

Keywords: Predictive modeling; Crop Insurance; Relational model; Aggregation bias; Data scarcity; Euclidean distance; Ratemaking.

1 Introduction

The U.S. Federal Crop Insurance Program (FCIP) has provided Multiple Peril Crop Insurance policies since its establishment in 1938. In 1980 the Federal Crop Insurance Act was established, and several changes were made to help improve the crop insurance program to make it more commercially widespread and sustainable over the long-run. One of the significant changes to the program was that coverage and rates were based on individual expected yields (i.e., the Actual Production History - APH program). This helped to avoid problems such as adverse selection (Ker and Coble, 2003), and the result has been a more commercially viable crop insurance program with substantially higher participation rates across producers. However, lack of data has been the biggest impediment preventing insurers and reinsurers from moving forward with more comprehensive offerings, leading to a gap between risk appetites and exposures being faced (Marsh & McClennan, 2014; Stelmakowich, 2015). In the absence of sufficient data, the industry may face significant challenges in building credible models, assessing the true risk, and confidently providing coverage.

This paper seeks to address this issue through focusing on predictive analytics regarding data scarcity and credibility, which are major difficulties facing the agricultural insurance sector often due to more limited loss experience data for those infrequent, but, extreme events. This lack of data is particularly relevant when considering individual loss events, rather than aggregate loss events. A new relational data-matching model is presented to predict individual loss events (i.e. individual farmer yields) in the absence of individual loss data (i.e. farm-level data). Central to crop insurance programs worldwide is the ability

to determine actuarially fair and sustainable premium rates. However, data scarcity and credibility, particularly related to the lack of farm-level yield data, are by far, the most fundamental difficulty facing the crop insurance sector (Porth et al., 2014).

The agricultural insurance sector is a high growth, niche market (Brockett et al., 2005; Mahul, 2000). In recent years, the agricultural insurance sector has developed significantly with direct global agricultural insurance premiums increasing from US$8 billion in 2005 to almost US$30 billion in 2013 (Schneider and Roth, 2013). The share of emerging market premiums is only about 22 percent of the total global portfolio (Swiss Re, 2013), leaving substantial growth opportunities in future. Comparatively, North America accounts for 55% of premiums written, with the U.S. operating the largest crop insurance program in the world. The total premiums of the U.S. crop insurance program were US$11.1 billion in 2012, with an insured value of US$117 billion. As agricultural insurance programs continue to grow in terms of scale and scope, actuarial foundations have become more important in order to ensure that programs are efficient and actuarially sound. As a result, the agricultural insurance sector is an important market, yet, few actuaries are involved and specialized in this area.

Data Scarcity and Credibility in Crop Insurance

There are many contributing factors to data scarcity and credibility in crop insurance. In most regions, there is only crop growing season per year, therefore, there is only one yield (or loss) observation per year. Further, due to crop rotation and other market forces farmers do not grow the same crop each year, and this leaves fewer observations or an inconsistent time series at the farm-level. In addition to limited data, there can also be concerns over the credibility of data. This can be attributed to data from the past that may not be representative of the current environment, due to such factors as changes in technology, biotechnology, farming practices, coverage levels, and other program details (Okhrin et al., 2013; Porth et al., 2014; Wang and Zhang, 2003; Woodard et al., 2012). The result is a need

to balance using as much of the time series as possible so that those infrequent, but, extreme weather events are considered, versus the concern that older data is not representative and, therefore, should be discarded. In order to help balance this incongruity, actuaries attempt to "restate" historical yield or loss experience to bring it on level with the current environment so that as many of the older observations can be used as possible.

In countries with developed crop insurance programs, data is often more substantial and of higher quality. In practice, this means that there is often as many as 40-45 years of historical yield or loss data at the aggregate level. However, at the farm-level data is still quite limited with approximately 5-10 years of historical data for a particular crop type, and most often there are gaps in the time series. In countries with developing crop insurance programs, the data challenges can be even more pronounced. This means that farm-level data may be limited to only a few years, or in some cases there are possibly no historical farm-level yield records. As a result, insufficient farm-level yield data can be one of the major challenges to the successful implementation of crop insurance programs (Porth et al., 2014). As an alternative to farm-level yield data, aggregate yield experience at the county level may be available. Since county-level data typically represents several producers' yield experience together in a region, there are often less challenges regarding missing data and insufficient time series.

Aggregation Bias

The concern with using aggregated county-level data in place of farm-level data, however, is that aggregating data may lead to the possible cancellation of idiosyncratic risk, and hence there may be smaller total risk in the aggregated data (Gerlt et al., 2014; Turvey, 2012). The result may be an underestimation of risk in the aggregate county-level data relative to the farm-level data. In the literature, this occurrence is often referred to as "aggregation bias" and is known to decrease the correlation between county-level and farm-level yield. For example, Claassen and Just (2009) estimate that farm-level yield variation is understated by

50% when using county-level averages, and 61% of systemic variation and 42% of random variation is lost when yields are aggregated to the county-level.

A Comparison of Individual Yield Insurance and Area Yield Insurance

The two main types of crop insurance programs considered in this paper include individual farm-level insurance products and area-based insurance products. In general, individual farm-level yield insurance is more common than area-based yield insurance. Under this program an indemnity is paid when the actual yield on the farm falls below the producer's individual long-term average historical yield. The payment is calculated as the shortfall in yield multiplied by a pre-determined price guarantee. In Canada, this type of insurance product is referred to as Multi-Peril Crop Insurance (MPCI) and is offered under the Agri-Insurance program (Pai et al., 2014), and in the U.S. it is referred to as Yield Protection (YP), formally called Actual Production History (APH) insurance. An alternative to farm-level yield insurance is area-yield insurance, where indemnities are based on shortfalls in the county average yield, rather than the individual farmer yield. The payment is calculated as the difference between the long-term average historical county yield and the actual average county yield multiplied by a pre-determined price guarantee. In Canada there are currently only a few area-yield products offered, however, in the U.S. there is a more widely used program called Area Revenue Protection (ARP), formally called the Group Risk Program (GRP).

Area-yield insurance was first offered in the U.S. in 1993, and has been encouraged over individual farm-level insurance, because area-yield products may help to reduce program losses and contribute to a more sustainable crop insurance program in the longer-term. County-level yield data serves as the foundation for the program. County-level data are much more plentiful compared to farm-level data, therefore, premiums may be able to be determined more accurately. Area-yield products are also less susceptible to issues of information asymmetry, including moral hazard and adverse selection, since an individual

farmer is not influential enough to impact the county average yields. Other cited benefits of area-yield products are that they have comparably lower transaction costs because individual farm-level yields do not need to be established and verified, and there is no need for on-farm loss adjustment. The main concern with area-yield insurance, however, is basis risk. This refers to the imperfect correlation between the county-level yields and farm-level yields, and this basis risk may make area-yield insurance unattractive to farmers. In 2005, area-based products accounted for only 9% of the liability of the Federal Crop Insurance Program (FCIP) in the U.S., and in Canada they would similarly account for a very small portion of products.

A New Relational Data-Matching Model to Link Aggregate Level and Individual Level Yield Data

Recognizing that farm-level yield data are likely sparse and may not be representative over time, some research has been conducted to link county-level data and farm-level data in order to enhance the data and provide a proxy for estimating the individual farm-level loss distribution (Coble and Barnett, 2008; Deng et al., 2007; Miranda, 1991). For instance, Miranda (1991) relates individual farm level yields to the county-level yields via a linear regression model. This framework, which strongly resembles the celebrated Capital Asset Pricing Model (CAPM), produces a "beta" coefficient that measures the sensitivity of the farm-level yields relative to the county-level yields. The decomposition of the farm-level yield variations into a systematic component and an idiosyncratic component in turns facilitates the analysis of risk reduction between individual-yield insurance and area-yield insurance. Ramaswami and Roe (2004) further establish that this model is valid if systemic and individual risks are additive in individual yields, and if the aggregation is such that the law of large numbers holds.

Extending Miranda (1991), Coble and Barnett (2008) investigate the systemic risk and idiosyncratic risk in farm-level data by allowing the farm-specific beta to vary over an assumed

normal distribution. Deng et al. (2007) assume that yields at both farm and county levels have a multiplicative relationship, and provide a simple way of generating additional pseudo farm-level yields from the county-level data. Cooper et al. (2012) generate a farm-level yield distribution that explicitly incorporates within county yield heterogeneity while accounting for systemic risk and other spatial or intertemporal correlations among farms within the county. They show that individualizing premiums for a farm results in substantial mispricing of crop insurance premiums because they do not adequately capture farm yield variability and yield correlations between farms.

To date, the methods that have been proposed focus on enhancing farm-level yield prediction using aggregate data at the county-level within the same region and same country. These methods assume that at least some farm-level historical crop yield data exists in the same country to serve as the underlying basis of prediction. However, in the situation where there are no farm-level data available, the methods developed in previous research cannot be applied. Therefore, this paper contributes to the literature by developing a new relational model to predict farm-level crop yield distributions in the absence of farm-level yield data, and this model may be extended to other areas of the Property & Casualty insurance sector beyond crop insurance.

The data-matching relational model will help to address the scenario where we are interested in predicting the individual loss experience, yet, only more aggregate level loss data is available. This may be the case in countries with extremely developed crop insurance programs, such as in Canada and the U.S., as well as in countries with new or developing crop insurance programs. More specifically, in the crop insurance example discussed in this paper, we are interested in the farm-level yield distribution in country A, which has relevant data, but, only at the county level. However, we know that there exists a country B where detailed yield experience are available at both farm and county levels. To estimate country A's individual farm-level yield distribution, our proposed solution is to develop a two-stage

relational model that enables us to "borrow" information from country B. For this reason, we refer to country A as our "base" country with missing farm-level data, and country B as our "reference" country with detailed farm and county yield data.

For our relational model to produce a credible individual farm-level yield distribution in the base country, two outstanding issues arise. First, we must determine which country in the reference country we should borrow the data from. Second, we must determine how the borrowed county and farm-level data from the reference country can be used to estimate the farm-level yield distribution in the base country.

Intuitively it is desirable to choose a reference county that is as similar as possible to the base county. By similarity we mean that agricultural production in both counties are alike, in the past and in the future. Since agricultural production of a farm/county depends on a number of variables, such as climate conditions, size of a farm/county, and so on, one plausible way of quantifying the similarity is via these variables. Therefore, in the first stage of the data-matching relational model we formally use a similarity measure based on an Euclidean metric to select the optimal reference county. Our adopted Euclidean distance measure incorporates weather information over the target growing season, average farm size, county size and county-level yield volatility for quantifying the similarity of two the counties. The optimal reference county corresponds to the county that has the smallest Euclidean distance. In the second stage of the relational model the data sets from the base country and reference country are linked. This is achieved through equating ratio and modeling the dependence structure between the farm-level and county-level yield losses in the reference country and applying the dependence structures to the base country.

To the best of our knowledge, this is the first paper to address the issue of farm-level yield prediction in the absence of farm-level data within the country. Therefore, it is important to validate the proposed data-matching relational model developed in this research. A unique and comprehensive data set is used, which is further described in the next section. The data

set includes both farm-level and county-level yield data from the U.S. and Canada, as well as corresponding weather data for the same regions. The U.S. and Canada are chosen only for an example, as the rich data set allows the opportunity to back-test and compare the out-of-sample prediction errors of the proposed model. While Canada faces less challenges in terms of data scarcity and credibility, compared to countries with developing crop insurance programs, there remains several situations in which this is an issue. For example, when new crop varieties are introduced in a particular region data scarcity and credibility can be a major impediment to the timely implementation of crop insurance coverage. Therefore, it is assumed that only the county-level and farm-level yield data in the U.S. (reference country) and the county-level yield data in Canada (base country) are used. The available farm-level yields in Canada are not used in the model construction, and are reserved for the validation of the proposed data-matching relational model approach. In addition, the method outlined by Deng et al. (2007) is generalized to predict the farm-level yields in the base country, providing additional benchmarking against the proposed relational model. Further, the model is validated by several robustness checks, as well as a MPIC pricing exercise. The empirical results demonstrate that the data-matching relational model developed in this paper is useful in predicting the farm-level yield distribution in the base country in the absence of farm-level yields, through borrowing information from a reference country that has more detailed information.

The results of this research may be useful for government as well as insurers and reinsurers looking to improve farm-level yield prediction in order to set coverage levels and determine premium rates more accurately. The data-matching relational model proposed in this paper may also help to overcome aggregation bias, which may lead to misleading and erroneous pricing results. Further, the relational model may be useful for developing more flexible crop insurance programs that are capable of responding to changing farming conditions/practices more quickly, where lack of data would normally prevent the development of a new insurance

products or enhanced coverage levels. An example may include a scenario where a certain crop variety is grown in a new region where it has not been grown before, due to biotechnology advancements, etc. Another example may include countries with relatively new and developing crop insurance programs, where yield data may not be available or may not be reliable. The relational model proposed in this study may, therefore, be useful in these situations in order to enhance the farm-level yield information through borrowing more detailed information from a reference country. The result may be that insurance coverage in these situations may be available to producers much sooner compared to current practice, and in the case of countries with new or developing crop insurance programs this method may be useful to help ensure that there is sufficient insurance and reinsurance capacity that may otherwise be lacking. This is because it is often necessary to collect sufficient yield or loss experience for as many as 5 - 10 years in order to develop new insurance products and compute the actuarially fair premium rate.

The remainder of this paper proceeds as follows. Section 2 describes the data sets that will be used in this paper. The proposed relational model is presented in two sections, with Section 3 focusing on Stage 1, which is the optimal selection of the county in the reference country, and Section 4 dealing with Stage 2, which links the two data sets. Section 5 generalizes the method of Deng et al. (2007) for benchmarking so that it can be used to predict the farm-level yields even though the farm data is not available in the base county. The validation and robustness of the models are analyzed in Section 6, and the MPCI pricing example is contained in Section 7. Section 8 concludes the paper.

2 Data Description

A unique and comprehensive data set that includes actual farm-level and county-level corn production experience from the U.S. and Canada is utilized in this research. The Canadian data was obtained from Manitoba Agricultural Services Corporation (MASC), a provincial

government insurance company in the province of Manitoba, and the U.S. data was obtained from Guy Carpenter & Company, a reinsurance intermediary. Given that both data sets include farm-level records, this portion of the data sets are confidential. The U.S. and Canada were selected for the empirical example because they both have relatively long and credible data sets available at both the farm and county levels. This provides an opportunity to evaluate the relational model developed in this research by comparing the error of the predicted farm-level yields to the actual farm-level yields. In this empirical example, Canada is the base country for which we are interested in predicting the farm-level yield distribution in the absence of farm-level data, and the U.S. is the reference country for which there is more detailed information that we borrow from. An overview of each of the datasets is presented next.

Canadian Data Set

The base country data set considered in this study is for Canada, and the focus is on the province of Manitoba. The data is provided by MASC, and contains farm-level and municipal-level (which is equivalent to county-level in the U.S. and hereafter is simply referred to as "county") corn yields per acre from 1,440 corn farms. The coverage period of the data is from 1996 to 2011. In addition, weather data is considered, which covers the period from 1975 to 2011. It is the Adjusted and Homogenized Canadian Climate (AHCC) data obtained from Environment Canada, including daily temperature (maximum, minimum, and average) from 24 weather stations, as well as daily precipitation from 30 weather stations.

U.S. Data Set:

The reference data set considered in this study is for the U.S., and the focus is on the entire country. The data set is provided by Guy Carpenter & Company, and represents a "sample company" that contains farm-level and county-level records from a number of the approved

insurance providers in the U.S. This "sample company" dataset is created to protect the proprietary nature of the data. The data set contains farm-level and county-level corn yields per acre from 5,817 corn farms. The coverage period of the data is from 1975 to 2013. In addition, weather data is considered, which also covers the same period. It includes daily average temperature and daily rainfall. The county-level yields and weather data are obtained from the National Agricultural Statistics Service (NASS), and are also available in an open source format from Ag-Analysis (Woodard, 2016a;b;c).

The geographical locations of the reference counties in the U.S. are displayed in Figure 1, and are marked with "+". The figure shows that the majority of counties are located in the central and north part of the country. It should be noted that the county-level data set is relatively large, however, the farm-level dataset can be limited. This is not surprising, as this is one of the motivations of this research. Table 1 summarizes the county-level data in the province of Manitoba, Canada. The statistical summary of the county-level data across the U.S. is collected in Section A of the online appendix.

3 Relational Model: Stage 1

The first stage of our proposed relational model requires us to determine an "optimal" county from the reference country that we will be borrowing information from. We adopt the Euclidean distance metric to determine the optimality. Suppose there are two points $\boldsymbol{p} = (p_1, p_2, \ldots, p_n)$ and $\boldsymbol{q} = (q_1, q_2, \ldots, q_n)$, each comprised of n variables. The Euclidean distance, $d_{\boldsymbol{p},\boldsymbol{q}}$, between these two points is given by[1]

$$d_{\boldsymbol{p},\boldsymbol{q}} = \sqrt{\sum_{i=1}^{n}(p_i - q_i)^2}. \tag{1}$$

[1]A detailed introduction of similarity measures is available in Goshtasby (2012).

Figure 1: Geographical Locations of Reference Counties in the U.S.

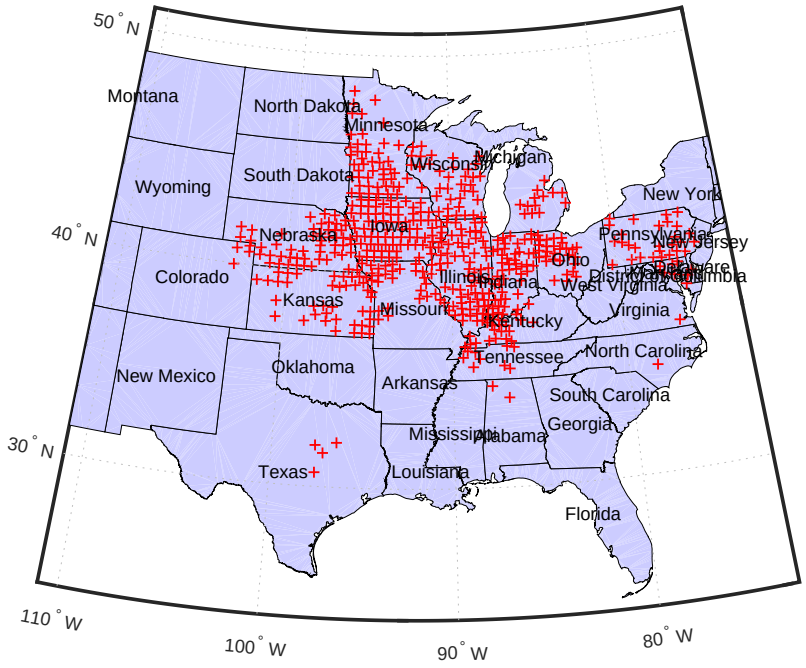


Table 1: Summary Statistics of the County-Level Corn Yield Data of Manitoba, Canada.

County Name	County Size	Avg Farm Size	Mean	S.D.	Skewness	Kurtosis
Brokenhead	1497.06	115.16	101.95	55.20	0.27	2.26
Desalaberry	7522.74	250.76	121.39	59.20	-0.12	1.96
Dufferin	40829.35	250.49	149.43	59.86	-0.61	3.18
Gery	7124.51	209.54	140.68	68.18	-0.08	2.36
Hanover	13823.28	189.36	106.18	50.93	-0.03	2.16
Labroquerie	2380.54	198.38	109.41	51.96	0.03	2.12
Montcalm	9641.50	260.58	121.65	44.04	-0.93	3.08
Northnorfolk	14414.71	205.92	168.00	66.73	-0.54	3.30
Pembina	2848.51	219.12	114.48	41.35	-0.79	3.15
Portagelaprairie	12387.70	247.75	131.06	47.72	-0.67	3.39
Rhineland	37780.85	188.90	131.40	53.06	-0.87	3.08
Roland	19717.77	226.64	143.91	51.96	-1.01	3.81
Steanne	5407.78	193.14	91.31	48.80	-0.01	1.75
Southnorfolk	4639.01	132.54	143.83	55.05	-0.65	3.43
Stanley	25851.60	211.90	134.73	42.06	-1.35	5.03
Tache	3427.60	163.22	79.75	37.16	-0.07	2.18
Thompson	10571.04	192.20	143.52	46.04	-1.23	4.82
Whitemouth	1603.17	145.74	120.87	68.48	0.18	2.37

Clearly, small $d_{\boldsymbol{p},\boldsymbol{q}}$ signifies that both points $\boldsymbol{p}$ and $\boldsymbol{q}$ are close to each other in term of the characteristics reflected by the given n variables. The above distance metric could be biased arising from the variations and scale differences among the variables. To account for these issues, it is desirable to standardize the original data before applying the standard Euclidean distance (1). The standardization of the data is often accomplished via the following transformation

$$\text{standardized value} = \frac{\text{original value} - \text{mean}}{\text{standard deviation}},$$

which ensures the standardized values have zero mean and unit variance. The Euclidean distance calculated using the above standardized values is known as the normalized Euclidean distance. The standardization is particularly important when we use the Euclidean distance to quantify the closeness/similiarity of two counties.

To measure the closeness/similarity of two counties, the selected variables should reflect the underlying characteristics of the county, which in our context relates to agricultural production. While there may be a number of factors that impact agricultural production in a county, for the purpose of this empirical exercise, the following variables are considered:

- Growing season weather information;
- Average farm size;
- County size;
- County's coefficient of variation, which is defined as the ratio of standard deviation over mean.

For the weather information, empirical evidence shows that temperature and precipitation during the growing season are important considerations for crop yields (Turvey, 2001; 2005; Woodard and Garcia, 2008a;b) and since the growing season in Canada and the U.S. are fairly similar, starting in April with harvest in approximately November, the monthly mean

temperature and cumulative precipitation from April to November are part of the variables for assessing the closeness of the counties. In order to consider the non-linearity effect of the temperature, the average growing degree days (GDD) variable over the growing season is also included. Together with the annual yield output, average farm size, county size, and the county coefficient of variation, a total 21 variables are used to determine the closeness of the counties via the normalized Euclidean distance. Therefore the optimal reference county corresponds to the county that gives the smallest normalized Euclidean distance relative to the base county. This, in turn, implies that the optimal reference county is most similar to the base county in terms of the 21 selected variables.

As an illustration, suppose our base county is "Dufferin" in the province of Manitoba, Canada and we are interested in finding the optimal reference county across the U.S. This entails calculating the normalized Euclidean distance over all the counties in the U.S. The Euclidean distance approach adopted in this study does not explicitly consider the geographical location of the counties to each other in order to allow for the possibility that the most similar information may belong to a county that is located a considerable geographical distance away from the farm being predicted, as illustrated in Figure 2. The x- and y-axis denote, respectively, the approximated geographical distance (in kilometres)[2] of the reference counties and the base county (i.e., "Dufferin") and the corresponding calculated Euclidean distance relative to the county (i.e., "Dufferin"). It is noteworthy that the optimal reference county selected by the Euclidean distance metric (which is identified by a point in the square in Figure 2, and labeled the "Min EucDistance") is different from the county with the smallest

[2]More specifically, the geographical distance is calculated as the great-circle distance between two locations on earth (i.e., the shortest distance over the earth's surface). Given the latitude and longitude pairs of two locations 1 and 2 denoted by (ϕ_1, γ_1) and (ϕ_2, γ_2), respectively in radians of locations, their geographical distance, D, can be calculated as

$$D = R \cdot 2 \cdot \arcsin\sqrt{\sin^2\left(\frac{\Delta\phi}{2}\right) + \cos\phi_1 \cdot \cos\phi_2 \cdot \sin^2\left(\frac{\Delta\gamma}{2}\right)},$$

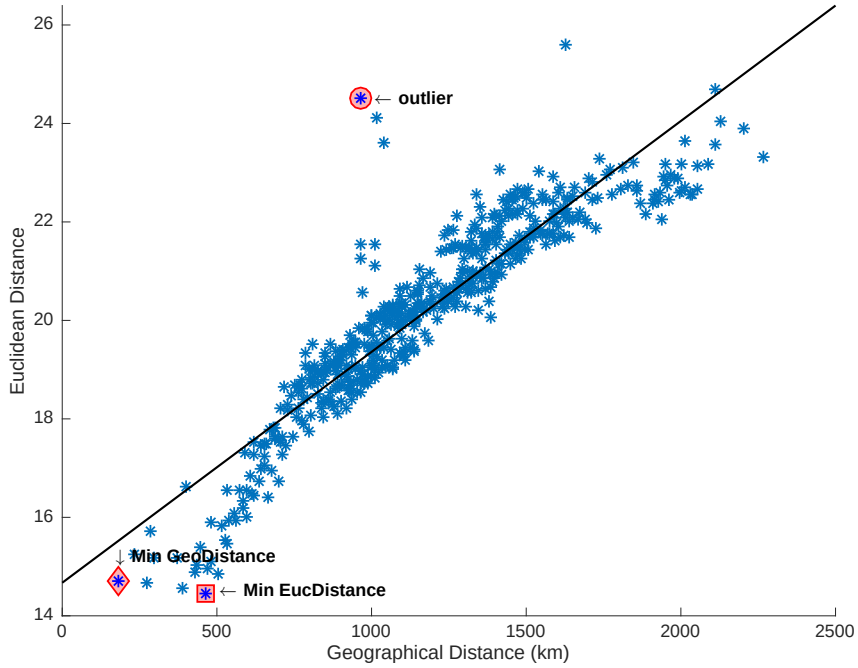
where $\Delta\phi = \phi_2 - \phi_1$ and $\Delta\gamma = \gamma_2 - \gamma_1$; and R is the earth's radium (approximated as 6371 km).

geographical distance (which is identified by a point in the diamond in Figure 2, and labeled the "Min GeoDistance"). However, when we examine the correlation between the geographical distance and Euclidean distance, the correlation between the two variables is 0.9 (p-value: 0.0000), indicating that the Euclidean distance calculated from selected variables between the base county and reference counties are significantly correlated with their geographical locations. In addition, there are also a few "outliers" (which is identified by a point in the circle in Figure 2, and labeled as "outlier"), which have relatively small GeoDistance, but considerable large EucDistance. The specific locations of these counties are displayed in the map in Figure 3. The discrepancy highlights the importance of using the Euclidean distance to select the optimal reference county, since the proposed Euclidean distance takes into consideration the selected characteristics of the counties as well as weather variables and other important information. Robustness analysis to be presented in Section 6.3 provides further support on the approach of selecting the optimal reference county in our proposed relational model.

4 Relational Model: Stage 2

We begin this section by introducing additional notation. Since we are dealing with information from Canada and the U.S., it is useful to use the superscript "CA" or "US" to differentiate the country of origin. Let $y_{i,t}^{\mathrm{X}}$ denote the year-t yield for county i in country X. Here $\mathrm{X} \in \{\mathrm{CA}, \mathrm{US}\}$ and $t \in \mathcal{T}^{\mathrm{X}}$, where $\mathcal{T}^{\mathrm{CA}} = \{1996, 1997, \ldots, 2011\}$ and $\mathcal{T}^{\mathrm{US}} = \{1975, 1976, \ldots, 2013\}$. The notation for the farm-level yield is defined similarly except that we need an additional subscript j to identify the j-th farm within the county i in country X; i.e. $y_{i,j,t}^{\mathrm{X}}$ where $j \in \mathcal{F}_i^{\mathrm{X}}$ and $t \in \mathcal{T}_i^{\mathrm{X}}$. Here $\mathcal{F}_i^{\mathrm{X}} \in \{1, 2, \ldots, n_i^{\mathrm{X}}\}$ where n_i^{X} represents the maximum number of farms within county i in country X, and $\mathcal{T}_i^{\mathrm{X}}$ captures the years for which farm-level yields exist in county i and country X. As pointed out earlier, farm-level yields are typically quite limited and sparse, hence $\mathcal{T}_i^{\mathrm{X}} \subseteq \mathcal{T}^{\mathrm{X}}$.

Figure 2: Scatter Plot of the locations of counties.



Using this notation, Subsection 4.1 specifies the yield models that we will use and Subsection 4.2 describes the assumptions underlying the relational model and the procedure for predicting the farm-level yields within the base county.

4.1 Yield Models Specification

For the county-level and farm-level corn yield data, we consider both log-linear and log-quadratic trend models to detrend the yield. More specifically, we have

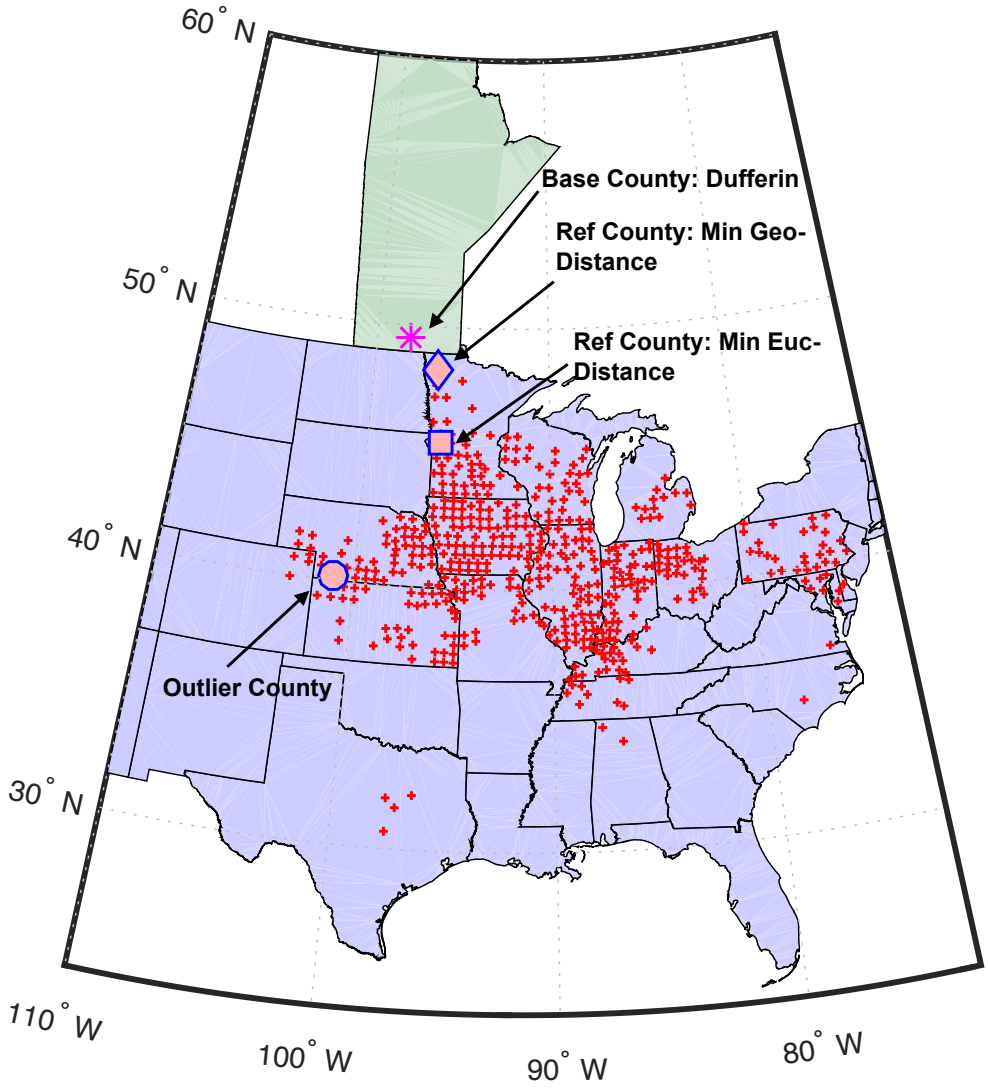
$$\text{Log-Linear:} \qquad \log\left(y_{i,t}^{\mathrm{X}}\right) = a_i^{\mathrm{X}} + b_i^{\mathrm{X}}(t - t_{i,0}^{\mathrm{X}}) + \varepsilon_{i,t}^{\mathrm{X}}, \tag{2}$$

$$\text{Log-Quadratic:} \quad \log\left(y_{i,t}^{\mathrm{X}}\right) = a_i^{\mathrm{X}} + b_i^{\mathrm{X}}(t - t_{i,0}^{\mathrm{X}}) + c_i^{\mathrm{X}}(t - t_{i,0}^{\mathrm{X}})^2 + \varepsilon_{i,t}^{\mathrm{X}}, \tag{3}$$

$$t \in \mathcal{T}^{\mathrm{X}}, \mathrm{X} \in \{\mathrm{CA}, \mathrm{US}\},$$

In this model, the yields in log scale are captured by two terms: the deterministic component $\mu_{i,t}^{\mathrm{X}}$ and the random component $\varepsilon_{i,t}^{\mathrm{X}}$. For the log-linear model (2), the deterministic

Figure 3: Geographical locations of the countries used for illustrative purposes. The asteroid shows the base county (i.e. "Dufferin"), which is in Canada, the diamond shows the reference county, which is selected based on the minimum geographical distance (Min GeoDistance), the square shows the reference county, which is selected based on the minimum Euclidean distance (Min EucDistance), and the circle shows an outlier from Figure 2.



component is comprised of a_i^{X}, which measures the central tendency of the log-yields, and $b_i^{\mathrm{X}}(t-t_{i,0}^{\mathrm{X}})$, which reflects the log-linear time trend attributed to factors such as technological changes, where $t_{i,0}^{\mathrm{X}}$ is the first year for which the data is available for country X; i.e. 1975 for the U.S. and 1996 for Canada. For the log-quadratic model (3), an additional quadratic term $c_i^{\mathrm{X}}(t-t_{i,0}^{\mathrm{X}})^2$ is added to reflect the effect not captured by the log-linear model (2). In the empirical study to be presented in Section 6, the optimal detrending model is determined based on Bayesian Information criterion (BIC).

The model for farm-level yields is defined similarily, except with the additional subscript j to identify the corresponding farm, and the parameters in the above model are interpreted as in the county-level model:

$$\text{Log-Linear:} \qquad \log\big(y_{i,j,t}^{\mathrm{X}}\big) = a_{i,j}^{\mathrm{X}} + b_{i,j}^{\mathrm{X}}(t-t_{i,j,0}^{\mathrm{X}}) + \varepsilon_{i,j,t}^{\mathrm{X}}, \tag{4}$$

$$\text{Log-Quadratic:} \quad \log\big(y_{i,j,t}^{\mathrm{X}}\big) = a_{i,j}^{\mathrm{X}} + b_{i,j}^{\mathrm{X}}(t-t_{i,j,0}^{\mathrm{X}}) + c_{i,j}^{\mathrm{X}}(t-t_{i,j,0}^{\mathrm{X}})^2 + \varepsilon_{i,t}^{\mathrm{X}}, \tag{5}$$

$$t \in \mathcal{T}_i^{\mathrm{X}}, j \in \mathcal{F}_i^{\mathrm{X}}, \mathrm{X} \in \{\mathrm{CA}, \mathrm{US}\}.$$

Using either of the above trend models, the predicted value of yield, predicted value of yield, $\hat{y}_{i,t}^{\mathrm{X}}$, $\hat{y}_{i,j,t}^{\mathrm{X}}$, can be obtained, and the corresponding detrended yields for prediction year $\tilde{t}$ can be calculated as:

$$y_{i,t}^{\text{Det-X}} = \frac{y_{i,t}^{\mathrm{X}}}{\hat{y}_{i,t}^{\mathrm{X}}} \cdot \hat{y}_{i,\tilde{t}}^{\mathrm{X}}, \quad t, \tilde{t} \in \mathcal{T}^{\mathrm{X}}, \mathrm{X} \in \{\mathrm{CA}, \mathrm{US}\}, \tag{6}$$

$$y_{i,j,t}^{\text{Det-X}} = \frac{y_{i,j,t}^{\mathrm{X}}}{\hat{y}_{i,j,t}^{\mathrm{X}}} \cdot \hat{y}_{i,j,\tilde{t}}^{\mathrm{X}}, \quad t, \tilde{t} \in \mathcal{T}_i^{\mathrm{X}}, j \in \mathcal{F}_i^{\mathrm{X}}, \mathrm{X} \in \{\mathrm{CA}, \mathrm{US}\}. \tag{7}$$

4.2 Relational Prediction Model

Let $i_*^{\mathrm{CA}} \equiv b^*$ represent the base county for which we are interested in estimating its corresponding farm-level yield distribution. Suppose the optimal county in the reference country has been identified in Stage 1 and is denoted by county $i_*^{\mathrm{US}} \equiv r^*$. Based on the model spec-

ification in Subsection 4.1, the detrended county- and farm-level yield for reference county r^* are obtained as $y_{r^*,t}^{\text{Det-US}}$ and $y_{r^*,j,t}^{\text{Det-US}}$. It should not be surprising that both $y_{r^*,t}^{\text{Det-US}}$ and $y_{r^*,j,t}^{\text{Det-US}}$ are dependent, and the key idea of the relational prediction model is to estimate the farm-level yield in the base county utilizing this dependence relationship. The challenge then is to accurately estimate the joint distribution of $y_{r^*,t}^{\text{Det-US}}$ and $y_{r^*,j,t}^{\text{Det-US}}$.

This paper utilizes a bivariate coupla approach to address the challenging issue of joint distribution estimation. To be more specific, according to Sklar Theorem (Sklar, 1959), the joint distribution of two random variables X and Y can be decomposed into two parts: the marginal distributions and the dependence structure, i.e., the copula function.[3] As a result, the joint distribution of $y_{r^*,t}^{\text{Det-US}}$ and $y_{r^*,j,t}^{\text{Det-US}}$, denoted as $F_{r^*,j}\left(x_{r^*}, x_{r^*,j}\right)$, is decomposed as

$$F_{r^*,j}\left(x_{r^*}, x_{r^*,j}\right) = C_{r^*,j}\big(F_{r^*}(x_{r^*}; \mu_{r^*}^{\text{US}}, \sigma_{r^*}^{\text{US}}), F_{r^*,j}(x_{r^*,j}; \mu_{r^*,j}^{\text{US}}, \sigma_{r^*,j}^{\text{US}}); \rho_{r^*,j}^{\text{US}}\big), \tag{8}$$

where $C_{r^*,j}\left(\cdot,\cdot;\rho_{r^*,j}^{\text{US}}\right)$ is the copula function with dependence parameter $\rho_{r^*,j}^{\text{US}}$; $F_{r^*}(x_{r^*}; \mu_{r^*}^{\text{US}}, \sigma_{r^*}^{\text{US}})$ is the marginal distribution of county-level yield $y_{r^*,t}^{\text{Det-US}}$, with mean $\mu_{r^*}^{\text{US}}$ and standard deviation $\sigma_{r^*}^{\text{US}}$; and $F_{r^*,j}(x_{r^*,j}; \mu_{r^*,j}^{\text{US}}, \sigma_{r^*,j}^{\text{US}})$ is the marginal distribution of the j-th farm-level yield $y_{r^*,j,t}^{\text{Det-US}}$, with mean $\mu_{r^*,j}^{\text{US}}$ and standard deviation $\sigma_{r^*,j}^{\text{US}}$.

To estimate the farm-level yield in the base county, the following additional assumptions are imposed:

$$\frac{\mu_{b^*,j}^{\text{CA}}}{\mu_{b^*}^{\text{CA}}} = \frac{\bar{\mu}_{r^*}^{\text{US}}}{\mu_{r^*}^{\text{US}}}, \qquad \forall j \in \mathcal{F}_{b^*}^{\text{CA}} \tag{9}$$

$$\frac{\sigma_{b^*,j}^{\text{CA}}}{\sigma_{b^*}^{\text{CA}}} = \frac{\bar{\sigma}_{r^*}^{\text{US}}}{\sigma_{r^*}^{\text{US}}}, \qquad \forall j \in \mathcal{F}_{b^*}^{\text{CA}} \tag{10}$$

$$\rho_{b^*,j}^{\text{CA}} = \bar{\rho}_{r^*}^{\text{US}}, \qquad \forall j \in \mathcal{F}_{b^*}^{\text{CA}} \tag{11}$$

[3]For a more detailed introduction of copulas, refer to Joe (1997) and McNeil et al. (2010). Copula models have become increasingly popular in the area of agricultural risk management and insurance (See, for example, Goodwin and Hungerford, 2014; Okhrin et al., 2013; Zhu et al., 2018).

where $\rho^{\mathrm{CA}}_{b^*,j}$ is the dependence parameter of the copula function between county- and farm-level yield in the base county b^*, namely,

$$F_{b^*,j}\left(x_{b^*}, x_{b^*,j}\right) = C_{b^*,j}\big(F_{b^*}(x_{b^*}; \mu^{\mathrm{CA}}_{b^*}, \sigma^{\mathrm{CA}}_{r^*}), F_{b^*,j}(x_{b^*,j}; \mu^{\mathrm{CA}}_{b^*,j}, \sigma^{\mathrm{CA}}_{b^*,j}); \rho^{\mathrm{CA}}_{b^*,j}\big), \tag{12}$$

and

$$\bar{\mu}^{\mathrm{US}}_{r^*} = \frac{1}{n^{\mathrm{US}}_{r^*}} \sum_{j=1}^{n^{\mathrm{US}}_{r^*}} \mu^{\mathrm{US}}_{r^*,j}, \tag{13}$$

$$(\bar{\sigma}^{\mathrm{US}}_{r^*})^2 = \frac{1}{n^{\mathrm{US}}_{r^*}} \sum_{j=1}^{n^{\mathrm{US}}_{r^*}} (\sigma^{\mathrm{US}}_{r^*,j})^2 \tag{14}$$

$$\bar{\rho}^{\mathrm{US}}_{r^*} = \frac{1}{n^{\mathrm{US}}_{r^*}} \sum_{j=1}^{n^{\mathrm{US}}_{r^*}} \rho^{\mathrm{US}}_{r^*,j}. \tag{15}$$

These assumptions imply that under the proposed relational model, the ratio of farm-level to county-level mean (standard deviation) in the base county is equal to the ratio of the average of the means (standard deviations) of all farms within the optimal reference county to the county-level mean (standard deviation) of the reference county (see (9), (10), (13) and (14)). In addition, the dependence between the base county-level log-yield and the individual farm-level log-yield, as measured by the copula parameter, is invariant across all the farms and is equal to the average of all corresponding dependence parameters in the optimal reference county (see (11) and (15)). Since the county-level mean for Canada, $\mu^{\mathrm{CA}}_{b^*}$, is known, we can calculate the farm-level mean as $\mu^{\mathrm{CA}}_{b^*,j} = \mu^{\mathrm{CA}}_{b^*} \cdot \frac{\bar{\mu}^{\mathrm{US}}_{r^*}}{\mu^{\mathrm{US}}_{r^*}}$. Similarly, the farm-level standard deviation can be calculated as $\sigma^{\mathrm{CA}}_{b^*,j} = \sigma^{\mathrm{CA}}_{b^*} \cdot \frac{\bar{\sigma}^{\mathrm{US}}_{r^*}}{\sigma^{\mathrm{US}}_{r^*}}$. Finally, the bivariate copula model (12) facilitates the determination of the individual farm-level yield distribution, hence the farm-level yield prediction. For example, suppose $(\hat{u}^k_{b^*}, \hat{u}^k_{b^*,j}), k = 1, \ldots, m$ are m random sample pairs drawn from the bivariate copula model with parameter $\rho^{\mathrm{CA}}_{b^*,j}$ as in (11), then it follows from (12), (13), and (14) that the farm-level yield in the prediction year $\tilde{t}$ can be

simulated according to

$$\hat{y}^{\text{Relational-CA}}_{b^*,j,k} = F^{-1}_{b^*,j}\left(\hat{u}^{k}_{b^*,j}; \mu^{\text{CA}}_{b^*,j}, \sigma^{\text{CA}}_{b^*,j}\right), \qquad k = 1, 2, \ldots, m, \ \forall j \in \mathcal{F}^{\text{CA}}_{b^*} \tag{16}$$

where $F^{-1}_{b^*,j}(\cdot)$ is the general inverse of the marginal distribution, $F_{b^*,j}(\cdot)$.

5 "Benchmark" Farm-Yield Prediction Model

Deng et al. (2007) attempt to improve the credibility of the farm-level crop yields by exploiting additional yield information from the county-level. The basic assumption of their approach is that the farm yield and the county yield are related multiplicatively. More specifically, suppose $\psi_{i,j,t}$ denotes the year-t multiplicative constant corresponding to farm j within county i; i.e.

$$\psi_{i,j,t} = \frac{y_{i,j,t}}{y_{i,t}} \qquad \forall j \in \mathcal{F}_i, t \in \mathcal{T}_i. \tag{17}$$

Note that we have suppressed the superscript country identifier since their model only deals with data within the same county (and hence the same country). The multiplicative constant $\psi_{i,j,t}$ can be interpreted as the idiosyncratic shock for farm j relative to the county as a whole. By calculating the idiosyncratic shocks (i.e. using (17)) from the available farm and county yields and assuming these shocks are equally applicable to all the observed county yields over the years, Deng et al. (2007) provide a simple resampling way of generating additional pseudo farm-level yields for the given farm within the county. Their method requires some empirical estimates of the idiosyncratic shocks, and hence still hinges on the availability of farm-level data, albeit limited. Our proposed relational model, in contrast, does not require such assumption and this is precisely the motivation for our current study.

Given the lack of farm-level data in the base country, the resampling method of Deng et al. (2007) can no longer be applied directly. Nevertheless we propose a generalization and

demonstrate that their resampling method can be adapted to exploit information from the reference country. As in our proposed relational method, the generalized method first requires the determination of an optimal reference county. Then we use (17) to estimate the idiosyncratic shocks, but, based on the data from the optimal reference county. While the estimated idiosyncratic shocks are of relevance to the reference county, which has the most similarity based on the Euclidean distance in the first stage, we assume that these shocks are equally applicable to the county-level yields in the base county to produce the pseudo sample farm-level yields. The algorithm below describes our generalized method of Deng et al. (2007), which we denote as the benchmark farm yield prediction model:

Algorithm 5.1 (**Benchmark farm yield prediction model**).

Given

- *Base county b^* and optimal reference county r^* obtained from Stage 1 of the relational method.*
- *Estimate $\hat{y}^{CA}_{b^*,t}$ and $\hat{y}^{US}_{r^*,t}$ based on (2) or (3) for $t \in \mathcal{T}^X$.*

Step 1: *Detrend the base and the reference county-level yields to the prediction year $\tilde{t}$ using (6):*

$$y^{Det\text{-}US}_{r^*,t} = \frac{y^{US}_{r^*,t}}{\hat{y}^{US}_{r^*,t}} \hat{y}^{US}_{r^*,\tilde{t}}, \qquad t, \tilde{t} \in \mathcal{T}^{US} \tag{18}$$

$$y^{Det\text{-}CA}_{b^*,t} = \frac{y^{CA}_{b^*,t}}{\hat{y}^{CA}_{b^*,t}} \hat{y}^{CA}_{b^*,\tilde{t}}, \qquad t, \tilde{t} \in \mathcal{T}^{CA}. \tag{19}$$

Step 2: *For each farm $j \in \mathcal{F}^{US}_{r^*}, t \in \mathcal{T}^{US}$, calculate its idiosyncratic shock relative to the reference county r^* via:*

$$\psi^{US}_{r^*,j,t} = \frac{y^{US}_{r^*,j,t}}{y^{US}_{r^*,t}}. \tag{20}$$

Step 3: *By assuming the idiosyncratic shocks obtained from the reference county are equally applicable to the base county, then its corresponding pseudo farm-level yields,* $\boldsymbol{y}^{Pseudo\text{-}CA}$*, can be simulated as follows:*

$$\hat{\boldsymbol{y}}^{Pseudo\text{-}CA} \;=\; \boldsymbol{y}_{b^*}^{Det\text{-}CA} \times (\boldsymbol{\psi}_{r^*}^{US})', \tag{21}$$

where $\boldsymbol{y}_{b^*}^{Det\text{-}CA}$ *and* $\boldsymbol{\psi}_{r^*}^{US}$ *are column vectors that collect all the detrended county yields calculated from* (19) *and all the idiosyncratic shocks estimated from* (20)*, respectively.*

By denoting n_b and n_r as the dimensions of $\boldsymbol{y}_{b^*}^{\text{Det-CA}}$ and $\boldsymbol{\psi}_{r^*}^{\text{US}}$, respectively, then the vector multiplication (21) produces $\hat{\boldsymbol{y}}^{\text{Pseudo-CA}}$, which is a matrix of dimension $n_b \times n_r$. All the $n_b \times n_r$ entries in $\hat{\boldsymbol{y}}^{\text{Pseudo-CA}}$ are considered as the pseudo farm yields of the base county. Parametric or non-parametric farm-level yield distributions can then be calibrated to these pseudo farm yields. Furthermore, the mean of these pseudo samples is our best estimate for the farm-level yield of the base county b^*. The above algorithm adapts the method of Deng et al. (2007) so that it can be used to predict farm-level yields even though the farm-level data within the base county does not exist. We will use this method as the benchmark against our proposed relational prediction model. For this reason, we refer this method as the benchmark farm yield prediction model.

6 Validation and Robustness Analysis

Numerous assumptions have been imposed in our proposed relational farm-level yield prediction model. It is therefore crucial to provide some validation to the proposed model, as well as to address issues associated with robustness and sensitivity. To accomplish these analysis, we proceed as follows. We note that our data set actually contains some farm-level yields within our base county even though such information has never been exploited for implementing the relational and the benchmark farm-level yield prediction models. The availability of these empirical data, even though limited and incomplete, can be used to

"backtest" the predicted farm-level yields from these models. To quantify the prediction errors, we adopt a metric known as the root-mean-square-error (RMSE). For each county i, the RMSE of the target statistic Q that we are predicting is calculated as follows:

$$RMSE_i(Q) = \sqrt{\frac{1}{n_i}\sum_{j=1}^{n_i}\left(\frac{\hat{Q}_j - \tilde{Q}_j}{\tilde{Q}_j}\right)^2}, \tag{22}$$

where $\tilde{Q}$ is the estimator of Q calculated from the actual farm-level data within the base county, and $\hat{Q}$ is an estimator of Q calculated from the farm-level yield prediction model of interest. Given the limited availability of the farm-level data and to ensure the resulting RMSE metric is reliable, the target statistic Q is taken to be either the mean or the standard deviation of farm-level yields, and also applies to those farms with at least 6 years of observations. Hence n_i in (22) represents the total number of farms in county i that satisfies this criterion.

With the above setup, Subsection 6.1 compares the performance of the relational model relative to the benchmark model. Both the "backtest" as well as the out-of-sample forecast results are provided. Subsections 6.2 focuses on the importance of the "ratio" assumptions we impose on the relational model, while Subsection 6.3 addresses the robustness of the optimal reference county selection via the Euclidean distance metric.

6.1 Relational model v.s. benchmark model

The objective of this subsection is to assess the performance of the farm-level yield prediction results from the relational and the benchmark models. Modeling the distribution of the underlying crop yields is fundamental to pricing and other risk management processes for agricultural insurance, given that selecting an incorrect distribution may lead to mispricing, as well as misleading conclusions (Sherrick et al., 2004). In literature researchers have attempted to model yields with parametric distributional attributes, including the normal distribution (Just and Weninger, 1999), log-normal distribution (Goodwin et al., 2000;

Jung and Ramezani, 1999; Stocks, 2000; Tirupattur et al., 1996), beta distribution (Nelson and Preckel, 1989; Sherrick et al., 2004; Tirupattur et al., 1996), gamma distribution (Gallagher, 1986), and Weibull distribution (Sherrick et al., 2004). Other research focuses on nonparametric estimation methods (Goodwin and Ker, 1998; Ker and Goodwin, 2000). In order to achieve a balance between overfitting and flexibility, parametric mixture models are also studied (Lanoue et al., 2010; Porth et al., 2014; Woodard and Sherrick, 2011; Yang, 2011).

Therefore, in this paper a comprehensive empirical study is conducted by first identifying the most appropriate joint distribution for the county- and farm-level yields. We accomplish this by calibrating different marginal distributions, as well as copula models, to the empirical data, and based on BIC the optimal combined model is determined. For the marginal distributions we consider distributions such as Normal, Gamma, Weibull, and Beta, while for the bivariate dependence modeling we use Archimedean copula models with varying tail dependence properties (i.e. Gumbel, Clayton, and Frank). The optimal calibrated models for the 18 counties of Manitoba are displayed in the final three columns of Table 2. We only report results for 18 counties as these are the only counties in Manitoba that cover data for the entire period of 1996-2011. The calibration results indicate that linear trending is sufficient for 12 of the 18 counties, and gamma distribution is optimal for modelling most of the marginal distributions. For the bivariate copula model, Clayton is optimal for half of the counties, Gumbel is optimal for six of the counties, and Frank is optimal for the remaining three counties. These results suggest that half of the counties exhibit lower tail dependence, six counties exhibit upper tail dependence, and the remaining three counties have no tail dependence.

Based on the optimally determined combined model, the ability of each model to predict farm-level yields is assessed. By taking the test statistics to be the mean and the standard deviation of the farm-level yields, columns 2-5 of Table 2 display the RMSEs for the

Table 2: The performance of the relational models relative to the banchmark model. RMSEs for mean and standard deviation estimations across 18 counties are reported in columns 2-5. The corresponding optimal trend models, marginal models, and copula models for each county are reported in the last three columns.

County Name	Relational		Benchmark		Optimal Models		
	Mean	S.D.	Mean	S.D.	Trend	Margin	Copula
Brokenhead	0.0061	0.4826	0.5258	0.7037	Linear	Gamma	Clayton
Desalaberry	0.3647	0.5215	0.4137	0.3955	Linear	Weibul	Clayton
Dufferin	0.0001	0.5120	0.0963	0.5838	Linear	Weibul	Gumbel
Gery	0.1880	0.4201	0.3301	0.1080	Quadratic	Normal	Gumbel
Hanover	0.0335	0.2044	0.5796	0.6327	Linear	Gamma	Gumbel
Labroquerie	0.1623	0.1760	0.3001	0.4115	Linear	Gamma	Gumbel
Montcalm	0.6910	0.6396	0.8413	0.8010	Quadratic	Gamma	Clayton
Northnorfolk	0.5006	0.8257	0.4867	0.3090	Linear	Normal	Clayton
Pembina	0.0850	0.5236	0.6137	0.7529	Quadratic	Gamma	Clayton
Portagelaprairie	0.0238	0.0067	0.3354	0.3052	Linear	Gamma	Clayton
Rhineland	0.2549	0.2420	0.8456	0.9119	Linear	Gamma	Gumbel
Roland	0.1829	0.4091	0.8159	0.8462	Quadratic	Gamma	Frank
Steanne	0.0360	0.3027	0.4545	0.5206	Linear	Gamma	Frank
Southnorfolk	0.1786	0.4840	0.8730	0.9193	Linear	Gamma	Clayton
Stanley	0.2010	2.1290	0.2593	1.0106	Linear	Normal	Clayton
Tache	0.2201	0.1544	0.3377	0.4361	Linear	Gamma	Gumbel
Thompson	0.0188	0.1452	0.4220	0.4900	Quadratic	Gamma	Frank
Whitemouth	0.1185	0.1805	0.8108	0.7550	Quadratic	Gamma	Clayton
Average	0.1814	0.4644	0.5190	0.6052	-	-	-
Minimum	0.0001	0.0067	0.0963	0.1080	-	-	-
Maximum	0.6910	2.1290	0.8730	1.0106	-	-	-

farm-level backtesting results from the respective models. The RMSEs of the relational model over the 18 counties range from 0.0001 to 0.6910 for mean estimation, and 0.0067 to 2.1290 for standard deviation estimation. The corresponding RMSEs (lowest, highest) for the benchmark model are (0.0963, 0.8730) and (0.1080, 1.0106). In terms of the effectiveness of the relational model and the benchmark farm yield prediction model for estimating mean and standard deviation, both methods have difficulty in accurately estimating the standard deviation. The performance of the two models for mean estimation are more encouraging. In particular, the relational model is capable of achieving RMSE of less than 0.1 for 7 out of 18 counties, which is much more accurate compared to the benchmark model where only 1 county has RMSE of less than 0.1. Furthermore, the relational model produces better prediction results in all counties, with the exception of Northnorfolk county. Regardless of whether the models are evaluated based on mean or standard deviation, on average the relational model outperforms the benchmark model. For the relational model, the average RMSEs across all 18 counties are 0.1814 and 0.4644 for mean and standard deviation estimations, respectively. In contrast, the respective averages for the benchmark model are 186% and 30% larger.

We now evaluate their out-of-sample prediction performance by using 1996-2010 as the in-sample training data and the year 2011 as the out-of-sample forecasting data. The one year out-of-sample forecasting results are displayed in Table 3. The effectiveness of the out-of-sample prediction from the relational model is clearly demonstrated. Among the 18 counties, the out-of-sample RMSEs of the relational model's mean estimation range from 0.0163 to 0.6935, with an overall average of 0.2525. Each of these results are superior compared to the benchmark model. For example, the lowest out-of-sample RMSEs of the mean estimation of the benchmark model is 0.3458. Only 5 of the 18 counties from the relational model produce RMSEs that are worse than 0.3458 of the benchmark model.

To conclude this subsection, it should be reminded that the original model of Deng et al.

Table 3: One year out-of-sample forecasting results of the relational model relative to the benchmark model. 1996-2010 are used as in-sample data to predict the year 2011. RMSEs for mean and standard deviation estimations across 18 counties are reported.

County Name	Relational		Benchmark	
	Mean	S.D.	Mean	S.D.
Brokenhead	0.1056	3.7726	0.5437	2.0005
Desalaberry	0.0466	2.7045	0.4364	1.3672
Dufferin	0.4119	1.8372	0.7317	0.9827
Gery	0.3381	2.3072	0.4470	0.5240
Hanover	0.0225	2.1348	0.4581	1.0341
Labroquerie	0.1130	0.1661	0.4602	0.1419
Montcalm	0.2308	3.2089	0.4587	1.1768
Northnorfolk	0.6032	0.0958	0.7072	0.7656
Pembina	0.0934	0.2015	0.5041	0.2215
Portagelaprairie	0.2799	1.3501	0.5619	0.0715
Rhineland	0.3821	0.9967	0.7280	1.2638
Roland	0.2456	1.7465	0.4740	1.0841
Steanne	0.1262	1.9408	0.4373	1.3606
Southnorfolk	0.6935	0.1082	0.7704	0.9166
Stanley	0.4837	1.5067	0.7223	0.9444
Tache	0.0163	0.9723	0.3458	0.5689
Thompson	0.1887	4.0280	0.5041	1.3639
Whitemouth	0.1640	1.9766	0.5536	0.4094
Average	0.2525	1.7252	0.5469	0.8999
Minimum	0.0163	0.0958	0.3458	0.0715
Maximum	0.6935	4.0280	0.7704	2.0005

(2007) generates additional pseudo farm yields by exploiting county-level yields together with the limited farm-level yields that are within the same county. The generalized model (i.e. the benchmark model) implemented in this paper extends this method by Deng et al. (2007), and instead develops a model that can borrow information from another county in a different country. As such, stringent assumptions, such as both counties (in different countries) having the same idiosyncratic shocks, are imposed. This, in part, may explain why the benchmark model is not as effective at farm-level yield prediction.

6.2 Robustness analysis on ratio assumptions

Recall that to estimate the farm-level yield in the base county, we have imposed the ratio equality assumption in the sense that the ratio of farm-level to county-level mean (standard deviation) of the log-yield in the base county is equal to the ratio of the average of the means (standard deviations) of all farms' log-yield within the optimal reference county to its county-level mean (standard deviation). For example, see (9) and (10), which defines the ratio equality assumption. The objective of this subsection is to investigate the sensitivity of the ratio equality assumption. To do this we consider the generalized ratios of (9) and (10) as follows:

$$\frac{\mu^{\mathrm{CA}}_{b^*,j}}{\mu^{\mathrm{CA}}_{b^*}} = \alpha \cdot \frac{\bar{\mu}^{\mathrm{US}}_{r^*}}{\mu^{\mathrm{US}}_{r^*}}, \qquad \forall j \in \mathcal{F}^{\mathrm{CA}}_{b^*} \tag{23}$$

$$\frac{\sigma^{\mathrm{CA}}_{b^*,j}}{\sigma^{\mathrm{CA}}_{b^*}} = \beta \cdot \frac{\bar{\sigma}^{\mathrm{US}}_{r^*}}{\sigma^{\mathrm{US}}_{r^*}}, \qquad \forall j \in \mathcal{F}^{\mathrm{CA}}_{b^*}, \tag{24}$$

where α and β are the proportional constants. In the special case with $\alpha = \beta = 1$, the above ratios recover the ratio equality assumption (9) and (10), respectively. Since our data set contains the actual farm-level yields of our base county, this in turn facilitates the estimation of these proportional constants for model validation and sensitivity analysis. For each individual farm, we can obtain the optimal α and β by numerically minimizing the

following RMSE:

$$\mathcal{L}(\alpha,\beta) \;=\; \mathsf{E}\left(\frac{y^{\text{CA}}_{b^*,j,t}-\hat{y}^{\text{Relational-CA}}_{b^*,j,t}}{y^{\text{CA}}_{b^*,j,t}}\right)^2. \tag{25}$$

Therefore if both coefficients α and β are significantly different from one, then the ratio equality assumption becomes questionable.

By repeating the above RMSE minimization procedure for all individual farms within the base county, we obtain optimal pairs of α and β. These optimal coefficients are averaged to produce the county-level optimal α^* and β^*, as reported in columns 5-6 of Table 4. From the optimal pairs of α^* and β^* we then use the generalized relational model given by (23) and (24) to estimate means and standard deviations, as we did in Table 2. The resulting RMSEs are reported in columns 6-7 of Table 4. We refer to this model as the *optimal relational model* to differentiate it from the standard relational method, which we have implemented in the preceding subsection with the implicit assumptions of $\alpha = \beta = 1$. For ease of comparison, the results from the standard relational model (i.e. imposing $\alpha = \beta = 1$) are reproduced in Table 4.

From the results depicted in Table 4, we make the following remarks. First, optimal α^* and β^* exhibit quite a bit of variability across the counties. It should be emphasized these optimal values rely on the existence of the farm-level yields of the base county, which in practice may not be available, and hence the optimal values in this case would not be able to be recovered. However, it is reassuring that on average, the optimal α^* and β^* across all 18 counties are 0.9801 and 1.0892, respectively. These values are reasonably close to one and hence justify the ratio equality assumption. The deterioration from using the standard relational model to estimate means and standard estimations are also within the tolerance. For example, while the average RMSE for the mean estimation is 3.6 times larger from the standard relational model (relative to the optimal relational model), its average RMSE is

18.1%, which is still reasonable in view of the lack of data.

Table 4: The impact of α and β on the relational models. RMSEs for mean and standard deviation (S.D.) estimations from standard and optimal implementations.

County Name	Standard Relational Model		Optimal Relational Model			
	Mean	S.D.	α^*	β^*	Mean	S.D.
Brokenhead	0.0061	0.4826	0.1560	1.3793	0.0145	0.0882
Desalaberry	0.3647	0.5215	0.2593	1.3512	0.0233	0.2433
Dufferin	0.0001	0.5120	1.6047	1.4195	0.0002	0.2398
Gery	0.1880	0.4201	0.5553	0.5482	0.0107	0.0085
Hanover	0.0335	0.2044	0.8830	1.1248	0.0257	0.0114
Labroquerie	0.1623	0.1760	1.2918	1.2650	0.0380	0.0247
Montcalm	0.6910	0.6396	1.6098	1.8988	0.1563	0.0999
Northnorfolk	0.5006	0.8257	1.7862	1.1567	0.0123	0.4655
Pembina	0.0850	0.5236	1.0002	0.9998	0.0730	0.2748
Portagelaprairie	0.0238	0.0067	1.0143	0.9857	0.0116	0.0044
Rhineland	0.2549	0.2420	1.0464	0.9933	0.1628	0.0580
Roland	0.1829	0.4091	0.6105	1.3924	0.0214	0.0328
Steanne	0.0360	0.3027	1.0015	0.9982	0.0149	0.2092
Southnorfolk	0.1786	0.4840	1.4196	0.9685	0.0494	0.0087
Stanley	0.2010	2.1290	0.4282	0.2718	0.1632	1.0700
Tache	0.2201	0.1544	0.8318	0.9707	0.1010	0.1444
Thompson	0.0188	0.1452	0.7536	1.0957	0.0049	0.0487
Whitemouth	0.1185	0.1805	1.3896	0.7864	0.0143	0.0444
Average	0.1814	0.4644	0.9801	1.0892	0.0499	0.1709
Minimum	0.0001	0.0067	0.1560	0.2718	0.0002	0.0044
Maximum	0.6910	2.1290	1.7862	1.8988	0.1632	1.0700

6.3 Sensitivity analysis in term of county selection

This subsection examines the importance of the optimal selection of the reference county on the proposed relational model. Recall that Stage 1 defines the optimal reference county as one that gives the lowest Euclidean distance. In this subsection, the importance of the choice of the reference county on the overall effectiveness of the proposed relational model is examined. In particular, rather than using the optimal reference county, here we examine the impact of randomly selecting the reference county. Additionally, forecasting results based on the reference county with the minimum geographical distance (i.e., county closest to the

base county) are also compared.

Table 5 provides some insights to this issue. In this table we produce three sets of results that are based on the standard implementation of the relational model. The first set of results, as shown in columns 2-3, predicts the mean and standard deviation based on the reference county optimally selected via the Euclidean distance minimization, which are the results previously discussed. The second set of results is obtained by randomly selecting a county from the reference country (i.e. the U.S.) and this in turn is used as the reference county in the relational model for predicting the farm-level yields in the base county. For each base county, this procedure is repeated 100 times, and the average RMSEs are reported in columns 4-5 of Table 5. The third set of results (columns 6-7 of Table 5) shows the prediction results with the optimal reference county selected based on the geographically closest distance. The results in Table 5 clearly highlight the sensitivity of the prediction results based on the reference county selected. The results also provide strong evidence in support of implementing the relational model based on the Euclidean distance optimality. Recall that the average RMSE for the mean estimation across 18 counties is 18.1%. In contrast, the comparable figures based on the reference county selected randomly, as well as selected based on the closest geographical distance are 241.9% and 238.7%, respectively, which is 13 times larger.

7 Application to Agricultural Insurance Pricing

The ultimate objective of imputing the farm-level crop yield distribution is to facilitate pricing agricultural insurance contracts, including the individual farm-level yield product (referred to as the MPCI contract). The premium rate of an agricultural insurance contract is often calculated based on the loss-cost ratio (LCR) random variable, which is defined as the ratio of indemnity (I) over liability (L). The liability L, which is known at the inception of the contract, measures the maximum exposure or the maximum payout of the

Table 5: The impact of county selection on the relational models. Results from counties selected based on Euclidean distance minimization, randomly, or geographical distance minimization are displayed. RMSEs for mean and standard deviation estimations across 18 counties are reported. For the random approach, the reported RMSE for each county is the average of 100 randomly selected reference counties.

County Name	**Relational Method with County Selected Based on**					
	Euclidean Distance		Random Approach		Geographic Distance	
	Mean	S.D.	Mean	S.D.	Mean	S.D.
Brokenhead	0.0061	0.4826	4.1580	2.2629	6.4129	2.2321
Desalaberry	0.3647	0.5215	0.5934	0.7836	1.1627	1.6656
Dufferin	0.0001	0.5120	0.3752	0.5892	0.1086	0.6371
Gery	0.1880	0.4201	0.8574	2.2086	1.1012	1.4766
Hanover	0.0335	0.2044	5.3030	4.5518	2.7001	1.8640
Labroquerie	0.1623	0.1760	5.9665	4.6305	7.2617	3.9963
Montcalm	0.6910	0.6396	0.4070	0.5614	0.1972	0.8953
Northnorfolk	0.5006	0.8257	1.1470	1.9430	0.5013	0.8249
Pembina	0.0850	0.5236	1.6318	0.9119	1.8033	0.2138
Portagelaprairie	0.0238	0.0067	4.7992	5.2649	2.5327	2.4709
Rhineland	0.2549	0.2420	0.9150	0.5437	0.8973	0.1971
Roland	0.1829	0.4091	1.7748	1.4445	1.5926	0.5517
Steanne	0.0360	0.3027	4.5505	3.6590	6.0700	3.3392
Southnorfolk	0.1786	0.4840	0.5640	0.6116	0.1791	0.4829
Stanley	0.2010	2.1290	0.5004	1.8994	0.9562	4.0763
Tache	0.2201	0.1544	5.4311	3.9886	3.2971	1.9726
Thompson	0.0188	0.1452	1.9592	1.9411	2.0459	1.2112
Whitemouth	0.1185	0.1805	2.6140	3.1488	4.1498	6.6064
Average	0.1814	0.4644	2.4193	2.2747	2.3872	1.9285
Minimum	0.0001	0.0067	0.3752	0.5437	0.1086	0.1971
Maximum	0.6910	2.1290	5.9665	5.2649	7.2617	6.6064

agricultural insurance contract and is typically defined as some percentage of the crop yield the farmer is expected to produce. If the actual farm yield for the year falls short of L, then the agricultural insurance contract will be triggered and the farmer will be indemnified for the magnitude of shortfall in yield. Formally, for a given farm j in county i we have

$$LCR_{ij} = \frac{I}{L} = \frac{\max(c \cdot \mathsf{E}(y_{ijt}) - y_{ijt}, 0)}{c \cdot \mathsf{E}(y_{ijt})}, \tag{26}$$

where $c \in (0, 1)$ is known as the coverage ratio.

Given a premium principle, the premium rate for the above agricultural contract can be calculated accordingly. For example, the premium rate is calculated as

$$\text{Premium Rate} = \mathsf{E}(LCR)(1 + \theta) \tag{27}$$

for the expectation premium principle, and

$$\text{Premium Rate} = \mathsf{E}(LCR) + \theta\sqrt{\mathsf{Var}(LCR)} \tag{28}$$

for the standard deviation premium principle. In both cases, θ is the risk loading parameter and it depends on the adopted premium principle, even though we have used the same notation for both premium principles.

In our numerical examples, we calculate premium based on the two aforementioned premium principles for the first county in Table 2, which has the lowest RMSE for mean prediction. In addition, the following parameter values are considered: the loading parameter $\theta \in \{0.1, 0.2, 0.3, 0.4\}$ and the coverage ratio $c \in \{0.5, 0.65, 0.75.0.85, 0.95\}$. These values are consistent with market practices (see Porth et al. (2013; 2014)). For each premium principle, we compare premium rates calculated based on the following three methods: the historical county-level yields, the predicted farm-level yields (using the relational model), and the his-

torical farm-level yields. We refer to the first method as the "county model", and collectively refer the second and third methods as the "farm model".

The computed premium rates for the assumed θ and c are reported in Table 6.[4] For both premium principles, the premium rates increase with the coverage level c and the risk loading parameter θ, which is to be expected. It is reassuring that the premium rates estimated from the yield prediction relational method (labelled as "Farm-Predict") are close to that based on the direct historical farm yield data (labelled as "Farm-Actual"). It should be emphasized that the yield prediction relational method does not rely on any of the available farm-level data in the base county (though it uses farm-level data from the reference county). In most cases the premium rates from the relational model are higher compared to those from the actual farm-level yields, indicating that pricing based on the relational method can be slightly more conservative. This can be interpreted as a positive attribute in view of the parameter and model uncertainties. Finally, and most importantly, the results caution using only the county-level yield data to approximate the farm-level premium rates, even though this is a common market practice (Gerlt et al., 2014). The discrepancy in the premium rates obtained from the "county model" and the "farm model" not only can be substantial, in many cases the county-level model can severely underestimate the actual results at the farm-level. To see this, the relative difference (in percentage) in premium rates between the "county model" and the "farm model" are computed, and these values are shown in parenthesis in Table 6. Most of these values are positive, indicating that the premium rates calculated from the "county model" are smaller than that from the "farm model". The degree of underestimation can be as high as over 60% in a few cases.

The large discrepancy between the models based on county and farm is a form of basis risk and is a consequence of the so-called aggregation bias. Figure 4 provides a graphical

[4]Due to the length consideration, we only produce the pricing results for the illustrative county Portage-laprairie, which has the best overall relational prediction results considering of both mean and standard deviation prediction as shown in Table 2. Pricing results for the other 17 counties are available in Section B of the online appendix.

Table 6: Agricultural insurance pricing results for county "Portagelaprairie". Premium rates are calculated using historical county yields, farm-level prediction yields (i.e. relational model), and historical farm-level yields, under expectation and standard deviation premium principles. The value in parenthesis denotes the relative difference (in percentage) in premium rate between the "county model" and the "farm model".

θ	Method	$c = 0.5$	$c = 0.65$	$c = 0.75$	$c = 0.85$	$c = 0.95$
Expectation Premium Principle						
	County	0.0349	0.0742	0.1075	0.1448	0.1849
0.1	Farm-Predict	0.1154 (69.78%)	0.1711 (56.61%)	0.2033 (47.09%)	0.2279 (36.44%)	0.2611 (29.18%)
	Farm-Actual	0.0956 (63.50%)	0.1488 (50.09%)	0.1850 (41.88%)	0.2239 (35.32%)	0.2606 (29.04%)
	County	0.0381	0.0810	0.1173	0.1580	0.2017
0.2	Farm-Predict	0.1259 (69.78%)	0.1866 (56.61%)	0.2218 (47.09%)	0.2486 (36.44%)	0.2848 (29.18%)
	Farm-Actual	0.1045 (63.57%)	0.1617 (49.90%)	0.2008 (41.58%)	0.2429 (34.95%)	0.2845 (29.11%)
	County	0.0412	0.0877	0.1271	0.1712	0.2185
0.3	Farm-Predict	0.1364 (69.78%)	0.2022 (56.61%)	0.2402 (47.09%)	0.2693 (36.44%)	0.3086 (29.18%)
	Farm-Actual	0.1124 (63.33%)	0.1752 (49.93%)	0.2188 (41.90%)	0.2639 (35.14%)	0.3091 (29.29%)
	County	0.0444	0.0945	0.1369	0.1843	0.2353
0.4	Farm-Predict	0.1469 (69.78%)	0.2177 (56.61%)	0.2587 (47.09%)	0.2900 (36.44%)	0.3323 (29.18%)
	Farm-Actual	0.1214 (63.44%)	0.1889 (49.98%)	0.2358 (41.95%)	0.2839 (35.07%)	0.3327 (29.26%)
Standard Deviation Premium Principle						
	County	0.0422	0.0826	0.1155	0.1517	0.1899
0.1	Farm-Predict	0.1275 (66.87%)	0.1815 (54.50%)	0.2129 (45.74%)	0.2373 (36.07%)	0.2684 (29.23%)
	Farm-Actual	0.1063 (60.25%)	0.1585 (47.91%)	0.1934 (40.27%)	0.2304 (34.15%)	0.2649 (28.30%)
	County	0.0528	0.0976	0.1333	0.1717	0.2117
0.2	Farm-Predict	0.1501 (64.84%)	0.2074 (52.93%)	0.2410 (44.70%)	0.2674 (35.78%)	0.2994 (29.27%)
	Farm-Actual	0.1258 (58.06%)	0.1812 (46.10%)	0.2177 (38.78%)	0.2558 (32.88%)	0.2931 (27.76%)
	County	0.0633	0.1127	0.1510	0.1917	0.2336
0.3	Farm-Predict	0.1727 (63.34%)	0.2334 (51.70%)	0.2691 (43.88%)	0.2974 (35.54%)	0.3304 (29.30%)
	Farm-Actual	0.1444 (56.16%)	0.2047 (44.95%)	0.2439 (38.07%)	0.2836 (32.41%)	0.3218 (27.42%)
	County	0.0738	0.1278	0.1688	0.2117	0.2554
0.4	Farm-Predict	0.1952 (62.18%)	0.2593 (50.72%)	0.2972 (43.22%)	0.3275 (35.36%)	0.3613 (29.32%)
	Farm-Actual	0.1643 (55.07%)	0.2280 (43.95%)	0.2695 (37.37%)	0.3099 (31.67%)	0.3497 (26.97%)

visualization of the aggregation by comparing the histograms of the county- and farm-level yields. The aggregation bias can also be quantified by comparing the difference between the variance of the county-level yields and the variance of the predicted farm-level yields. The difference can be expressed absolutely or relatively, as shown below:

$$\text{Absolute Aggregation Bias}_i = \sigma_{i,j}^2 - \sigma_i^2, \tag{29}$$

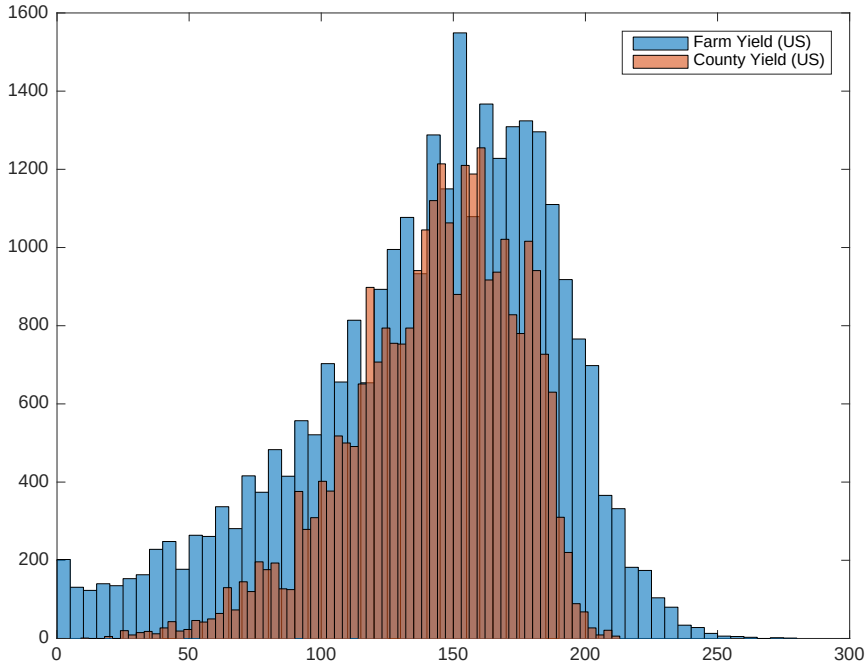
$$\text{Relative Aggregation Bias}_i = \frac{\sigma_{i,j}^2 - \sigma_i^2}{\sigma_{i,j}^2}. \tag{30}$$

In the above formulas, σ_i^2 denotes the sample variance estimated from the county i's county-level yield data and $\sigma_{i,j}^2$ is the variance of the farm-level yields of the relational yield prediction model. Table 7 reports both the absolute and relative aggregation biases for each county. Intuitively, aggregating data leads to possible cancellation of idiosyncratic risks, and hence results in smaller total risk. This implies that higher yield variance at the farm level relative to the county level should be expected. This is confirmed by the results in Table 7 which show that most counties have positive aggregation bias.

8 Conclusion

This paper develops a new data-matching relational model to contribute to the area of predictive analytics, and address the challenges regarding data scarcity and credibility, which are major difficulties facing the Property & Casualty insurance sector. More limited loss experience data for those infrequent, but, extreme events makes it difficult for the industry to build credible models and calculate the real risk and confidently provide coverage. This can lead to a gap between risk appetites and exposures being faced. Using crop insurance as an example, a new relational data-matching model developed in this paper is used to predict individual farm yields in the absence of farm-level data. The relational model defines a similarity measure based on a Euclidean distance metric that considers weather information, farm size, county size and the coefficient of variation of yield, to search for the most "similar"

Figure 4: Illustration of aggregation biases. Histograms of county- level and farm-level crop yields across all available counties and farms are displayed in the top figure for the U.S. and bottom figure for Canada (in 2011 level).



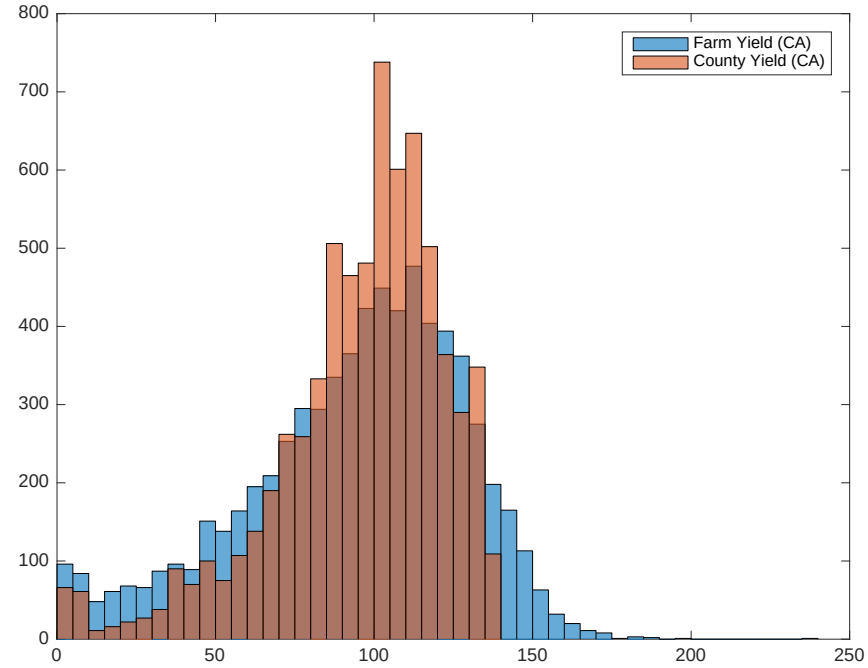


Table 7: Absolute and relative aggregation biases across the counties.

County	Relative Bias	Absolute Bias
Brokenhead	13.52%	7.03
Desalaberry	−118.08%	−18.66
Dufferin	40.09%	19.50
Gery	−30.99%	−15.61
Hanover	7.06%	3.67
Labroquerie	1.45%	0.63
Montcalm	38.83%	36.02
Northnorfolk	30.61%	13.78
Pembina	35.95%	28.20
Portagelaprairie	−10.01%	−4.01
Rhineland	80.82%	221.22
Roland	44.36%	61.10
Steanne	12.07%	5.95
Southnorfolk	71.82%	120.97
Stanley	51.65%	28.66
Tache	3.21%	1.36
Thompson	36.40%	37.26
Whitemouth	58.79%	130.22

region in a different country in order to borrow individual loss experience data that is otherwise not available. A unique data set from the U.S. and Canada are utilized to empirically validate the model. By optimally selecting marginal distributions and the bivariate copula function, the empirical results show that the proposed relational model is able to predict the farm-level yield with a reasonable degree of accuracy. The relational model achieves lower mean (standard deviation) prediction error of 18.14% (46.44%) compared to the benchmark model of 51.90% (60.52%), and the premium rates computed based on the relational model are found to recover the actual premium rates based on the historical loss experience data more closely compared to using only county-level data. The results also highlight the importance of the first step of the relational model algorithm, which selects an optimal reference county for borrowing the information based on a Euclidean distance metric, showing that the yield prediction results are significantly improved (more than 13 times better) compared to that based on random selection or closest geographical distance. Lastly, the results also

confirm the existence of aggregation bias and caution using only county-level yield data to approximate farm-level premium rates, as the degree of underestimation can be higher than 60% in some cases.

This research provides a new approach for insurers, reinsurers and governments to enhance individual loss experience, helping to overcome data scarcity and credibility issues that present substantial challenges in risk modelling, pricing and developing new insurance programs, particularly in developing countries. In order to implement this approach in practice, government or private insurance companies must have access to detailed farm and county-level data from other countries. In reality, however, most countries do not currently make agriculture related data public, particularly at the farm-level where confidentiality issues are a concern. Therefore, countries with crop insurance programs should consider their policies on making some data public in order to make advancements in predictive analytics regarding yield forecasting and pricing approaches that required shared data. More recently, several initiatives have been underway to address this issue, and platforms such as Ag-Analytics (Woodard, 2016a;b;c) have begun to standardize crop insurance data and make them available in the domain. Future research may consider examining additional selection algorithms beyond the Euclidean distance metric used in this research. Further, additional variables may be considered in the future in order to generalize the relational model and possibly improve the prediction ability.

References

Brockett, P. L., Cooper, W. W., Golden, L. L., Rousseau, J. J., Wang, Y., 2005. Financial intermediary versus production approach to efficiency of marketing distribution systems and organizational structure of insurance companies. Journal of Risk and Insurance 72 (3), 393–412.

Claassen, R., Just, R. E., 2009. Heterogeneity and distributional forms of farm-level yields.

American Journal of Agricultural Economics 93 (1), 144–160.

Coble, K., Barnett, B. J., 2008. Implications of integrated commodity programs and crop insurance. Journal of Agricultural and Resource Economics 40 (2), 431–442.

Cooper, J., Zulauf, C., Langemeier, M., Schnitkey, G., 2012. Implications of within county yield heterogeneity for modeling crop insurance premiums. Agricultural Finance Review 72 (1), 134–155.

Deng, X., Barnett, B. J., Vedenov, D. V., 2007. Is there a viable market for area-based crop insurance? American Journal of Agricultural Economics 89 (2), 508–519.

Gallagher, P., 1986. U. S. corn yield capacity and probability: Estimation and forecasting with none-symmetric distributions. North Central Journal of Agriculure Economics 8, 109–122.

Gerlt, S., Thompson, W., Miller, D., 2014. Exploiting the relationship between farm-level yields and county-level yields for applied analysis. Journal of Agricultural and Resource Economics 39 (2), 253–270.

Goodwin, B. K., Hungerford, A., 2014. Copula-based models of systemic risk in u.s. agriculture: Implications for crop insurance and reinsurance contracts. American Journal of Agriculture Economics.

Goodwin, B. K., Ker, A. P., 1998. Nonparametric estimation of crop yield distributions: Implications for rating croup-risk (GRP) crop insurance contracts. American Journal of Agricultural Economics 80, 139–153.

Goodwin, B. K., Roberts, M. C., Coble, K. H., 2000. Measurement of price risk in revenue insurance: implications of distributional assumptions. Journal of Agricultural and Resource Economics 25 (1), 195–214.

Goshtasby, A. A., 2012. Image Registration. Springer, Ch. 2: Similarity and Dissimilarity Measures.

Joe, H., 1997. Multivariate Models and Dependence Concepts. Taylor & Francis.

Jung, A. R., Ramezani, C. A., 1999. Valuing risk management tools as complex derivatives: An application to revenue insurance. Journal of Financial Engineering 8, 99–120.

Just, R., Weninger, Q., 1999. Are crop yields normally distributed? American Journal of Agricultural Economics 81 (2), 287–304.

Ker, A., Coble, K., 2003. Modeling conditional yield densities. American Journal of Agriculture Economics 85 (2), 291–304.

Ker, A., Goodwin, B. K., May 2000. Nonparametric estimation of crop insurance rates revisited. American Journal of Agricultural Economics 83 (2), 463–478.

Lanoue, C., Sherrick, B. J., Woodard, J. D., Paulson, N. D., July 25-27 2010. Evaluating yield models for crop insurance rating. In: Selected Paper prepared for presentation at the Agricultural & Applied Economics Associations 2010 AAEA, CAES & WAEA Joint Annual Meeting.

Mahul, O., 2000. Optimum crop insurance under joint yield and price risk. Journal of Risk and Insurance 67 (1), 109–122.

Marsh & McClennan, September 2014. Ahead of the curve: Understanding emerging risks. emerging risks report. Available at https://www.mmc.com/content/dam/mmc-web/Files/AheadoftheCurve-UnderstandingEmergingRisks.pdf, accessed: 05-14-2017.

McNeil, A. J., Frey, R., Embrechts, P., 2010. Quantitative Risk Management: Concepts, Techniques, and Tools. Princeton university press.

Miranda, M. J., 1991. Area yield crop insurance reconsidered. American Journal of Agricultural Economics 73, 233–242.

Nelson, C. H., Preckel, P. V., 1989. The conditional distribution as a stochastic production function. American Journal of Agricultural Economics 71, 370–378.

Okhrin, O., Odening, M., Xu, W., 2013. Systemic weather risk and crop insurance: The case of china. Journal of Risk and Insurance 80 (2), 351–372.

Pai, J., Boyd, M., Porth, L., 2014. Insurance premium calculation using credibility analysis: An example from livestock mortality insurance. Journal of Risk and Insurance 82 (2), 341–357.

Porth, L., Tan, K. S., Weng, C., 2013. Optimal reinsurance analysis from a crop insurer's perspective. Agriculture Finance Review 73 (2), 310 – 328.

Porth, L., Zhu, W., Tan, K. S., 2014. A credibility-based erlang mixture model for pricing crop reinsurance. Agricultural Finance Review 74 (2), 162–187.

Ramaswami, B., Roe, T., 2004. Aggregation in area-yield crop insurance: The linear additive model. American Journal of Agriculture Economics 86 (2), 420–431.

Schneider, K., Roth, M., 2013. Insider quarterly, Q-Re. Growing Premium. Accessed January, 2015, available at: www.insiderquarterly.com/growing-premium.

Sherrick, B., Zanini, F., Schnitkey, G., Irwin, S., 2004. Crop insurance valuation under alternative yield distributions. American Journal of Agricultural Economics 86, 406–419.

Sklar, A., 1959. Fonctions dé repartition án dimension et leurs marges. Publ. Inst. Statist. Univ. Paris 8, 229–231.

Stelmakowich, A., 2015. Lack of cyber data a current barrier to more comprehensive offerings: Kovacs. Available at http://www.canadianunderwriter.ca/insurance/

lack-of-cyber-data-a-current-barrier-to-more-comprehensive-offerings-kovacs-1003648329/, accessed: 05-14-2017.

Stocks, J. R., 2000. A derivative security approach to setting crop revenue coverage insurance premiums. Journal of Agricultural and Rsource Economics 25, 159–176.

Swiss Re, 2013. New swiss re sigma study puts the spotlight on the role that insurance can play in improving food security for over 850 million people globally. Accessed January, 2015, available at: www.media.swissre.com/documents/sigma1_2013_en.pdf.

Tirupattur, V., Hauser, R. J., Chaherli, N. M., December 1996. Crop yield and price distributional effects on revenue hedging. In: OFOR Paper Number 96-05.

Turvey, C. G., 2001. Weather derivatives for specific event risks in agriculture. Review of Agricultural Economics 23 (2), 333–351.

Turvey, C. G., 2005. The pricing of degree-day weather options. Agricultural Finance Review 65 (1), 59–85.

Turvey, C. G., 2012. Whole farm income insurance. Journal of Risk and Insurance 79 (2), 515–540.

Wang, H. H., Zhang, H., 2003. On the possibility of a private crop insurance market: a spatial statistics approach. Journal of Risk and Insurance 70 (1), 111–124.

Woodard, J. D., 2016a. Ag-analytics.org data warehouse. Available at ag-analytics.org, accessed: 09-30-2015.

Woodard, J. D., 2016b. Big data and ag-analytics: an open source, open data platform for agricultural & environmental finance, insurance, and risk. Agricultural Finance Review 76 (1), 15–26.

Woodard, J. D., 2016c. Data science and management for large scale empirical applications in agricultural and applied economics research. Applied Economic Perspectives and Policy.

Woodard, J. D., Garcia, P., 2008a. Basis risk and weather hedging effectiveness. Agricultural Finance Review 68 (1), 99–117.

Woodard, J. D., Garcia, P., 2008b. Weather derivatives, spatial aggregation, and systemic risk: Implications for reinsurance hedging. Journal of Agricultural and Rsource Economics 33 (1), 34–51.

Woodard, J. D., Schnitkey, G. D., Sherrick, B. J., Lozano-Gracia, N., Anselin, L., 2012. A spatial econometric analysis of loss experience in the U.S. crop insurance program. Journal of Risk and Insurance 79 (1), 261–286.

Woodard, J. D., Sherrick, B. J., 2011. Estimation of Mixture Models using Cross-Validation Optimization: Implications for Crop Yield Distribution Modeling. American Journal of Agricultural Economics 93 (4), 968–982.

Yang, T., July 24-26 2011. Measurement of yield distributions: time-varying mixture distribution models. In: Prepared for Agricultural & Applied Economics Association's 2011 Annual Meeting, Pittsburgh, Pennsylvania.

Zhu, W., Tan, K. S., Porth, L., Wang, C.-W., 2018. Spatial dependence and aggregation in weather risk hedging: A lévy subordinated hierarchical archimedean copulas (lshac) approach. ASTIN Bulletin: The Journal of the IAA 48 (2), 779–815.