6-2024

Utilizing Extreme Value Theory to Uncover Yield Distributions from Farm and County Level Historical Corn Yields

Gerald H. Van Tassell
University of Nebraska-Lincoln, vantas10@msu.edu

UTILIZING EXTREME VALUE THEORY TO UNCOVER YIELD DISTRIBUTIONS FROM FARM AND COUNTY LEVEL HISTORICAL CORN YIELDS

by

Gerald H. Van Tassell

A THESIS

Presented to the Faculty of

The Graduate College at the University of Nebraska

In Partial Fulfillment of Requirements

For the Degree of Master of Science

Major: Agricultural Economics

Under the Supervision of Professor Cory G. Walters

Lincoln, Nebraska

June 2024

UTILIZING EXTREME VALUE THEORY TO UNCOVER YIELD DISTRIBUTIONS FROM FARM AND COUNTY LEVEL HISTORICAL CORN YIELDS

Gerald H. Van Tassell M.S.

University of Nebraska, 2024

Advisor: Cory G. Walters

Yield risk represents a major portion of the financial risk facing corn producers and is found in the left tail of the yield distribution. Traditional methods for generating yield distributions fall into two categories: parametric and non-parametric. The shape and behavior of the tail of parametric yield distributions are determined by distributional assumptions. Non-parametric distributions fail to account for the possibility of as yet unseen extreme events, often referred to as “Black Swans”. Extreme Value Theory (EVT) rectifies these issues by providing an empirical, parametric estimate of the risk of extreme events, regardless of the underlying distribution of corn yields.

A new method for generating complete yield distributions using EVT and Kernel Density Estimation (KDE) is proposed. EVT is used to estimate the tails of the yield distribution and KDE is used to estimate the body of the yield distribution. The new method combines the EVT estimate of the tails of the yield distribution with the KDE of the body of the yield distribution into a complete yield distribution.

County-level yield data is often used instead of farm-level yield data due to the paucity of farm-level yield data. The aggregation from farm to county-level data changes the shape of the yield distribution and reduces its variance. The new method for generating complete crop yield distributions is implemented on four datasets from three

aggregation levels. Data from Preston Farms in Hardin County, Kentucky is compared to county-level yield data from Hardin County. Data from the Knorr-Holden Plot, a research plot in Scotts Bluff County, Nebraska is compared to county-level yield data from Scotts Bluff County. This paper showcases best practices for the application of EVT to small sample crop yield data.

The relationship between farm, field, and county-level yield distributions is heterogenous. While the Hardin County yield distribution well approximates the Preston Farms yield distribution, the Scotts Bluff County yield distribution is a poor approximation of the Knorr-Holden Plot yield distribution.

In the future, the improved method for estimating yield distributions will be applied to a net income model to improve producer decision making under uncertainty.

TABLE OF CONTENTS

LIST OF TABLES AND FIGURES

CHAPTER 1: INTRODUCTION

An agricultural producer's risk management strategy is a major factor in the long-term survival of the operation (Kim et al., 2019). Price and yield risk are two of the most important risks facing the producer. Price risk can be managed through futures markets, forward contracts, and crop insurance. Yield risk is primarily managed via crop insurance.

While the costs of hedging and crop insurance may be clear, the risk manager must be able to make an accurate assessment of the benefits to create an optimized risk management strategy. The benefits of crop insurance and hedging are in the left tail of the price and yield distributions.

1.1 Motivations

The role of minimum yields represents the foundation to risk management design as consequential events exist in the extreme. Little is known about the behavior of minimum yields as they are, by construction, rarely observed. Extreme Value Theory (EVT) provides a framework for understanding the behavior of minimum yields, leading to improved risk management design, an important step as minimum yields and risk management design directly affects the probability of farm survival.

The fundamental idea behind EVT is that the minimum (or maximum) has a behavior that can be modeled (Coles et al., 2001). The main challenge inherent in estimating the tails of a distribution is a paucity of data. Yields are only observed once per year, at harvest, which is compounded by the rarity of extreme events in the tail.

There are also events, often referred to as “Black Swans”, that we have yet to experience. Extreme Value Theory (EVT) enables empirical estimation of the probability of events that have yet to occur but are nonetheless possible. For the risk manager, improving the understanding of rare, financially devastating events impacts risk management decision making and investment.

Our work builds upon several papers that have used EVT to better underwrite crop insurance (Park et al., 2019) or to create price distributions (Morgan et al., 2012), but EVT has not previously been used in a model to improve producers' decision making process. EVT will be combined with Kernel Density Estimation (KDE) to create a complete yield distribution. One of the goals of this project is to apply the yield distributions to improving the net income model developed by Walters and Preston (2018). An improved estimate of the probability of low yield events will improve modeling of the probability of extreme low-income events.

An additional application of the method for generating yield distributions with EVT is to investigate the relationship between yield data aggregated at different levels. County-level yield data is often used instead of farm-level yield data for crop insurance and farm policy research due to the lack of availability of farm-level data (Claasen & Just, 2011). County-level yield can significantly underestimate the risk of extreme low yield events. A common solution to this is to multiply the county-level yield distribution by a coefficient of variation (Cooper et al., 2009). This assumes that the county and farm-level distributions have the same tail structure. Applying EVT to farm and county-level yield data uncovers the tail structure of both the farm and county-level yield distributions.

1.2 Outline

Our work is presented in two major sections. The first section contains the method for generating crop yield distributions. The second section applies that method to four datasets. There are three levels of aggregation represented in the four datasets. Field (plot) level data comes from the Knorr-Holden Plot in Scotts Bluff County, Nebraska. Farm-level data is available from Preston Farms in Hardin County, Kentucky. We compare these two levels of aggregation to their county-level yield distributions, Scotts Bluff and Hardin Counties respectively.

CHAPTER 2: METHODS

Understanding producers' underlying yield distribution helps reveal the true amount of risk producers face. Over the past century, agricultural producers have experienced improvement in yields, obscuring identification of the underlying yield distribution. Yields have been increasing over time due to technological advancement, improvements in knowledge, and input intensification. To identify the underlying yield distribution, the effect of increasing yields must be removed. Increasing yields are removed by a process called detrending. The goal of detrending is to isolate deviations from potential yield caused by random events such as weather and disease. While the idea of detrending is straightforward, many methods are available to detrend data. When choosing from the suite of detrending methods, the researcher must balance simplicity with overfitting, where overfitting results in random events being captured in the trend. Errors made in detrending can compound in later steps of the analysis.

2.1 Residual Ratios

There are two components to our method for uncovering the yield distribution. In the first component, we detrend the data by fitting the data using one of the detrending methods with regression. The second component addresses heteroskedasticity using residual ratios. Residual ratios are defined as:

$$RR_t = \frac{y_t - \hat{y}_t}{\hat{y}_t}$$

where y_t represents the observed yield in year t and $\hat{y}_t$ represents the trend yield in year t. Residual ratios represent a percent deviation from the trend yield. As yields increase over time, so does the size of the residuals. Normalizing the residuals obtained from the first stage of detrending removes the trend of increasing variance. The use of residual ratios requires assuming a multiplicative error structure. An alternative approach would be to use an additive error structure. In our case, we find that an additive error structure results in heteroskedasticity (Figure 2.1). Our data exhibits a trend in yields which causes the variance of the non-standardized residuals to increase with time and thus requires the use of residual ratios.

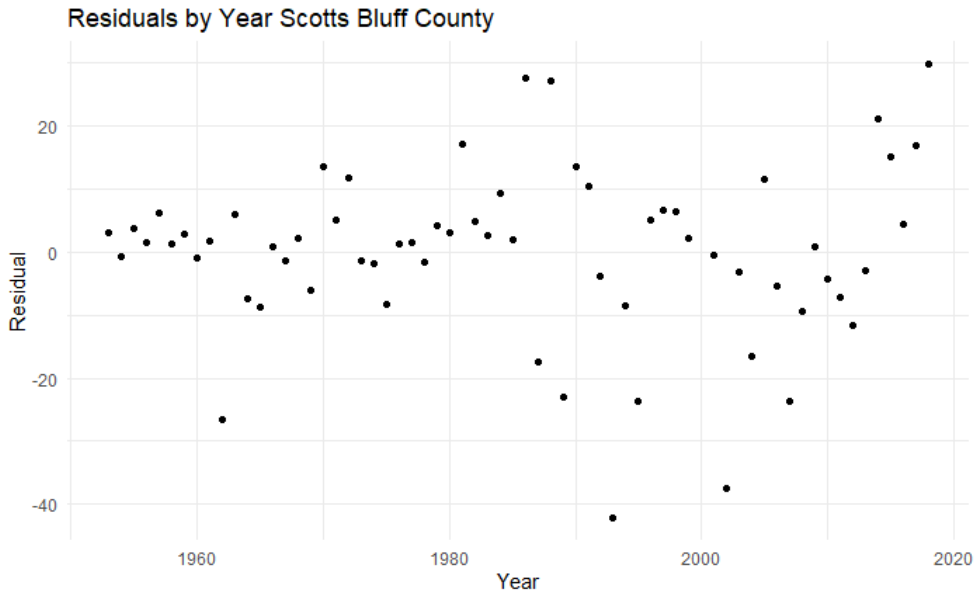

Figure 2.1 Scotts Bluff County Residuals by Year

After detrending (and constructing residual ratios), the distribution of percent deviations from trend yield should be weakly stationary (Ye et al., 2015). A weakly stationary time series has constant mean, variance, and autocorrelation structure. If our

estimation of trend is correct, then any deviations from that trend should be due to random events like weather and disease. We contend that it is reasonable to assume that the effect of these events on yield is constant throughout time (Deng, Barnett, and Vedenov 2007, Ker and Coble 2003). This is certainly not true in practice. For example, climate change is causing shifts in the expected effects of weather on crops. There are methods for estimating trends in moments above the mean (Tolhurst and Ker 2015), but they are beyond the objectives of this paper. See Harri et al. (2011) for a discussion on the impacts of heteroskedasticity assumptions on estimated yield distributions.

2.2 Linear Detrending Methods

Two general approaches have been chosen for the detrending process. The first approach is straightforward and involves a choice between ordinary least squares (OLS), one-knot piecewise linear spline, and two-knot piecewise linear spline. The Risk Management Agency (RMA) relies upon OLS, one-knot splines, and 2-knot linear splines for detrending historical yield data.[1] OLS and one-knot piecewise linear splines have been used by Tolhurst and Ker (2015) and Ker and Coble (2003). We include the two-knot piecewise spline as it is also straightforward to use while also allowing for additional flexibility to control for additional technological and knowledge advancements. From here forward we refer to the straightforward approach as linear methods. The simple linear regression is:

[1] RMA uses these detrending methods in area-based policies such as Area Risk Protection Insurance (ARPI), Supplemental Coverage Option (SCO), and Enhanced Coverage Option (ECO).

$$\hat{y}_t = \beta_0 + \beta_1 t + \varepsilon_t.$$

Where β_0 represents the intercept; β_1 represents the trend and ε_t represents the error term. Equation 1 is estimated using Ordinary Least Squares (OLS). Piecewise linear splines divide the data range into distinct intervals and fit separate linear regressions to each interval. The one-knot linear spline is:

$$\hat{y}_t = \begin{cases} \gamma_0 + \gamma_1 t + e_t & for\ t \le k_1 \\ \gamma_0 + \gamma_1 t + \gamma_2 (t - k_1) + e_t & for\ t > k_1 \end{cases}$$

where k_1 represents a predetermined year that separates the two different linear trends, commonly referred to as the "knot". k_1 is selected by minimizing the average of the residual ratios, defined in Eq. 1, squared. The equation is formulated such that it is continuous, but not differentiable at the knot location. The two-knot piecewise linear spline is:

$$\hat{y}_t = \begin{cases} \alpha_0 + \alpha_1 t + \epsilon_t & for\ t \le h_1 \\ \alpha_0 + \alpha_1 t + \alpha_2 (t - h_1) + \epsilon_t & for\ h_1 > t \le h_2, where\ h_1 < h_2 \\ \alpha_0 + \alpha_1 t + \alpha_2 (t - h_1) + \alpha_3 (t - h_2) + \epsilon_t & for\ t > h_2 \end{cases}$$

where h_1 and h_2 represent pre-determined knot locations using the same procedure defined in the one-knot linear spline.

An Augmented Dickey-Fuller (ADF) test is performed on the residual ratios resulting from the three different detrending methods. The ADF test tests the null hypothesis that a time series sample is non-stationary against the alternative hypothesis that the sample is a stationary series. To avoid overfitting, the model with the least complexity (number of knots) that produces a stationary series is chosen.

2.3 Locally Estimated Scatterplot Smoothing (LOESS)

The second detrending method is locally estimated scatterplot smoothing (LOESS). With LOESS, for each observation, y_t, the nearest neighboring points are used in a weighted quadratic least squares regression to estimate $\hat{y}_t$, trend yield. The number of neighboring points used to estimate the trend yield for each year is predetermined and is based on the bandwidth parameter, *b*, the percent of total observations to be used in each weighted regression. The tri-cube weight function is used to weight the observations within each weighted regression. The tri-cube weight function is the standard for use in the weighted regressions that make up the LOESS method (Cleveland, 1979). The tri-cube weight function assigns higher weights to observations closer to the year for which the yield is being estimated, reducing the weight for observations further away in time. The bandwidth can range from zero to one. A larger (smaller) bandwidth parameter results in a smoother (more jagged) final LOESS curve. The bandwidth parameter is chosen through 10-fold cross-validation to avoid overfitting as in (Lu, Carbone, and Gao 2017). The final LOESS curve is created by combining the trend yield estimated by each weighted quadratic least squares regression. The ADF test is performed on the residual ratios derived from the LOESS to ensure stationarity.

The two detrending methods, linear methods (three types) and LOESS, were used to assess the effects of different methods on the shape of the yield distribution. A practitioner will be required to make a subjective choice between the two methods to create a final yield distribution as there is no clear test to identify which method to use. As a result, the decision on which detrending method to select must be based primarily on

a visual comparison of the models. However, the linear methods approach to detrending has several distinct advantages to the practitioner. The first advantage is simplicity. The linear methods result in a simple linear trend that is easy to interpret and communicate. The LOESS method outputs a series of fitted values that vary non-linearly. This non-linearity provides LOESS with an advantage in the case of a high curvature as seen in some corn yield trends. Flexibility to non-linear yield trends is a key advantage of the LOESS method.

Another key advantage of the linear methods is that they can be used to extrapolate trend yield into the future. LOESS cannot be used to extrapolate into the future. This does not eliminate the usefulness of the LOESS method. If LOESS is subjectively determined to be the best fit for the data, it should still be used to generate the final yield distribution. The distribution of residual ratios is centered at 0 and is generated based on data up to year t. To estimate the yield distribution for year t+1, we need an estimate for the trend yield in year t+1. The support for the distribution of residual ratios multiplied by the estimate of the trend yield in year t+1 plus the estimated trend yield in year t+1 results in the "true" yield distribution for year t+1. It would be reasonable if the best method for detrending the historical yield data allows for an estimate of trend yield in year t+1, but if this is not the case, the best model of historical yields should be combined with the best estimate of future yields to generate the final distribution. If the LOESS method best represents the trend in historical yields, it should be used to create the residual ratio distribution. The linear methods could then be used to predict the future trend yield (year t+1) and create the final yield distribution. Both

methods require pre-determining parameters whether it be knot location or bandwidth parameter.

2.4 Extreme Value Theory

Extreme Value Theory (EVT) provides an empirically based estimate of the area below the lowest value observation in a yield distribution. The heart of EVT lies in the Fisher-Tippet-Gnedko (FTG) theorem. The FTG theorem states that the distribution of the maxima (or minima) of a sequence of i.i.d. random variables will come from one of three types of extreme value distributions, regardless of the distribution of the underlying random variable (Coles et al., 2001). The three distributions are Fréchet, Gumbel, and Weibull.

To identify maxima (or minima), the FTG theorem, when first introduced, relied upon the Block Maxima (BM) approach, which fits a sequence of maxima to the Generalized Extreme Value (GEV) distribution. By construction, the block maxima identify the largest (smallest) value within a predetermined range, thereby ignoring information from observations with similar, but, in the case of maxima (minima), smaller (larger) values. To account for the ignored information, the Peaks over Threshold (PoT) model has since been developed. The PoT model is based on an extension to the FTG theorem known as the Pickands–Balkema–De Haan (PBD) theorem. The PDM theorem states that the distribution of the exceedances of a random variable above some threshold, *u,* is the Generalized Pareto Distribution (GPD). While in theory both the PoT and BM approach, in the limit, provide the same result, in practice it has been shown that the GPD provides a better fit due to its more efficient use of data (Madsen et al., 1997).

The GPD is defined over three parameters: shape (ξ), scale (σ), and location (u). The shape parameter, ξ, controls the rate of decay of the tail of the distribution. The three types of extreme value distributions also correspond to the rate of decay in the tail. The GPD combines all three distributions into one, based on ξ. The Fréchet distribution ($\xi > 0$) has an unbounded, heavy tail (i.e., student-t). The Weibull distribution ($\xi < 0$) has a thin, bounded tail (i.e., beta distribution). The Gumbel distribution ($\xi = 0$) has a light tail and is equivalent to the exponential distribution when $u=0$. The CDF of the GPD is given by

$$\hat{F}(rr) = \begin{cases} 1 - \left(1 + \dfrac{\xi(rr - \mu)}{\sigma}\right)^{-\frac{1}{\xi}} & for\ \xi \neq 0 \\ 1 - e^{-\frac{rr-\mu}{\sigma}} & for\ \xi = 0 \end{cases}$$

where rr is residual ratios. Once the threshold has been selected the shape and scale parameters are estimated with maximum likelihood estimation. This is done using the package "evir" in R. The above formula is for the GPD in the right tail of the distribution. For use in the left tail, the negative of the residual ratios is used to estimate ξ and σ.

When using parametric distributions, observations in the middle of the distribution have a large effect on the shape of the tail. The choice of parametric distribution also affects the resulting distribution of the tail. EVT uses only the observations in the tail to estimate the tail of the distribution. Using EVT means that the data guides the shape of the tail, rather than the assumptions of the modeler. Non-parametric distributions, like empirical distributions or Kernel Density Estimates, do not suffer the same problems as parametric distributions. For both empirical distributions and KDEs, only observations in the tail are used to estimate the tail of the distribution. These

non-parametric measures, however, have their own faulty assumptions as well. In the case of an empirical distribution, there is no probability mass above (below) the maximum (minimum) observation. It is always possible for future events to be more extreme than past events, but empirical distributions disregard such possibilities. KDEs have a similar issue, placing a normal distribution around observations is an improvement to the empirical distribution, but it is still imposing an assumption about the structure of the tail. EVT removes the need for making assumptions about the shape or structure of the tail of the distribution by estimating the structure of the tail based on the observations in the tail.

EVT is applied to both the upper and lower tails of the yield distribution. More attention is paid to the fitting of the GPD to the lower tail of the yield distribution as the risk of extreme low yield events is contained in the lower (left) tail of the yield distribution. While we focus on the application of EVT to the risk of extreme low yield events, EVT also improves the estimate of the upper tail of the yield distribution. The KDE often over-smooths the upper tail of the yield distribution, assigning probabilities of unreasonably high yield events. EVT corrects this over-smoothing by providing a bounded upper tail.

2.5 Threshold Selection

Threshold selection is a classic bias-variance tradeoff (Benito et al 2023). The PBD theorem applies in the limit, as u approaches infinity, so selecting a threshold, u, too close to the mean will lead to an inadequate representation of the tail by the GPD. If the threshold is chosen to be far in the extremes of the distribution, there will not be enough

data to reliably parameterize the GPD. We consider thresholds in the first quartile of the data for left tail threshold selection. Within the possible threshold range, the threshold choice has a significant effect on the shape parameter, ξ. Due to the small sample sizes from historical crop yield data, the bias-variance tradeoff is magnified. The additional variance associated with using a traditional threshold for extreme events, like the fifth percentile results in inconsistent parameter estimates. A threshold at the fifth percentile would contain only three observations from the Scotts Bluff County datasets and only two observations from the Hardin County datasets. As a result, we consider thresholds up to the 25^{th} percentile, the upper limit of what could be considered the tail of the yield distribution.

Theoretically, the shape parameter should vary linearly with u. However, when working with small samples, linearity may be hard to obtain, resulting in a shape parameter varying wildly with the threshold. This makes threshold selection a critical step in fitting the GPD to historical crop yields. Tools such as the Hill estimator, mean excess plots, parameter stability plots, and qq plots have been developed to assist in the process of threshold selection. For small sample sizes, the mean excess and parameter stability plots were found to be the most useful.

A mean excess plot graphs the relationship between the threshold level (u), and its associated average exceedance. The usefulness of the mean excess plot is based in part on the fact that beyond a threshold u_0, for which the GPD is a good approximation, the mean

excess of a given threshold should vary linearly in u[2]. Thus, a threshold, u_0, above which the mean excess plot is linear in u can be accurately approximated by the GPD (Coles et al., 2001). For very high threshold levels, there are not enough exceedances to get a reliable estimate of the mean exceedance. For this reason, the mean excess of the most extreme thresholds should be ignored when analyzing the mean excess plot. It is tempting to use footnote 2 as evidence that the slope of the mean excess curve can be used as an estimate of ξ. For reasons laid out in Ghosh & Resnick (2010), there are reasons to doubt the consistency of such an estimator. It is still possible to use the sign of the estimated slope as an indication of whether ξ will be positive (Fréchet), negative (Weibull), or close to 0 (Gumbel) (Das & Ghosh, 2016). The effect of threshold choice on the slope of the mean excess plot is the best way to understand the effect of threshold selection on the shape parameter, ξ.

A parameter stability graph plots the estimated shape parameter, ξ, against the threshold from which the shape parameter was estimated. If the exceedances of a threshold, u_0, can be modeled by the GPD, then exceedances of a threshold greater than u_0 should also be well approximated by the GPD and should have the same shape parameter as the exceedances of u_0 (Coles et al., 2001). For this reason, if beyond a threshold u_0 the estimate of ξ is relatively stable, u_0 should be used as the threshold for the GPD. It is important to keep in mind that when fitting the GPD to the tails of a crop yield distribution there is an extremely limited number of observations. Thus, the

[2] This result comes from the fact that $E[X - u|X > u] = \frac{\sigma_{u_0} + \xi u}{1-\xi}$. See Coles pg. 79.

practitioner must treat these two tools as mere heuristics and use them to make the best choice possible even when the results are unclear.

2.6 Kernel Density Estimation

While EVT is used to model the tails of the yield distribution, a Kernel Density Estimate (KDE) is used to model the body of the yield distribution. The KDE is a non-parametric density estimator that allows the authors to refrain from choosing a parametric distribution to represent crop yields. A KDE can be thought of as a smoothed empirical distribution. A Gaussian KDE is constructed by creating n normal distributions with mean RR_i and variance h^2, summing all n normal distributions, and dividing by n to ensure that the resulting PDF integrates to one. Each residual ratio (RR_i) is used to predict the PDF in its neighborhood. The influence of a residual ratio on the PDF decreases with distance from the observation, according to the kernel being used. We use the Gaussian kernel, the kernel typically used when estimating crop yield distributions using KDEs (Goodwin and Ker 1998). The choice of kernel has a limited effect on the final KDE (Chen 2017). The KDE estimates the PDF of the residual ratios as:

$$\hat{f}(rr) = \frac{1}{n}\sum_{i=1}^{n}\frac{1}{h}K\left(\frac{rr - RR_i}{h}\right)$$

where $\hat{f}(rr)$ is the estimated density at point rr, n is the number of observations, h is the smoothing parameter, RR_i are the observed residual ratios, and K is the Gaussian kernel. h, the smoothing parameter governs the width of the kernel and thus the degree to which the distribution is smoothed. A larger (smaller) h leads to a smoother (more jagged)

density estimate. The choice of smoothing parameter is a bias-variance tradeoff. A small h can cause oversensitivity to individual data points creating jagged peaks, capturing more noise than signal. A large h can over-smooth the distribution, leaving out detail and resulting in a flatter distribution. There are three types of methods for solving the bias-variance tradeoff and selecting the smoothing parameter: cross-validation, rules of thumb, and plug-in approaches. The rules of thumb rely on an expectation that the data comes from a close to a normal distribution (Scott 2010). To avoid making such assumptions, the Sheather Jones plug-in is used for the smoothing parameter as it is shown to be an improvement over cross-validation methods (Sheather 2004).

2.7 Combining KDE and EVT

Once the KDE and the pareto tails have been estimated it is necessary to combine them into a single distribution. This can be done in two ways, depending on the preferred way of describing the final yield distribution. The first method is to sample from the KDE, then remove samples from below the lower threshold and above the upper threshold and replace those samples with samples from the estimated GPDs. The resulting sample can be used in further net income analysis. This method is the simplest and best way of producing a sample from the combined yield distribution.

We create a CDF of the complete yield distribution to provide an easy visual comparison of yield distributions from different data sources. The second method for combining the KDE and Pareto tails into a complete yield distribution is to construct a CDF of the complete distribution which can be used to generate a sample of the complete

yield distribution. This requires the construction of a CDF of the GPDs as well as the CDF of the KDE. The CDF of the GPD is given as

$$\hat{F}(rr) = \begin{cases} 1 - \left(1 + \frac{\xi(rr-\mu)}{\sigma}\right)^{-\frac{1}{\xi}} & for\ \xi \neq 0 \\ 1 - e^{-\frac{rr-\mu}{\sigma}} & for\ \xi = 0 \end{cases}$$

This is the formula for the CDF of the GPD for the upper tail. To account for this in the lower tail, the GPD is fit to the negative of the residual ratios. The true CDF is given by $\hat{G}(rr)$, where

$$\hat{G}(rr) = 1 - \hat{F}(-rr)$$

Because yields cannot be lower than 0, which is equivalent to a residual ratio of -1, $\hat{G}(rr)$ is defined for $rr \epsilon (-1, \mu)$. A yield of 0 is assumed to occur with probability $\hat{G}(-1)$. $\hat{G}(\mu) = 1$.

The CDF of a KDE is equal to the integral of the PDF of the KDE

$$\hat{F}(rr) = \int_{-\infty}^{rr} \hat{f}(x)\, drr$$

where $\hat{f}(x)$ is the PDF of the Gaussian KDE

$$\hat{F}(rr) = \int_{-\infty}^{rr} \frac{1}{n} \sum_{i=1}^{n} \frac{1}{h} K\left(\frac{rr - RR_i}{h}\right) drr$$

Because both integration and summation are linear operators, they can be interchanged.

$$\hat{F}(rr) = \frac{1}{n}\sum_{i=1}^{n}\int_{-\infty}^{rr}\frac{1}{h}K\left(\frac{rr - RR_i}{h}\right)drr$$

The integral of the Gaussian kernel is equal to the CDF of the Gaussian distribution

$$\hat{F}(rr) = \frac{1}{n}\sum_{i=1}^{n}\Phi\left(\frac{rr - RR_i}{h}\right)$$

where Φ is the CDF of the Gaussian distribution.

The PDF of a Gaussian KDE is constructed by placing a Gaussian PDF on each observation. At each point *rr* in the support, the density estimate is the average of the values of the Gaussian PDFs at *rr.* Similarly, the CDF of a Gaussian KDE is constructed by placing a Gaussian CDF on each observation. At each point *rr* in the support, the cumulative density estimate is the average value of the Gaussian CDFs at *rr.*

To create a complete yield distribution, the CDF of the GPDs and the CDF of the Gaussian KDE must be combined. The combined CDF is given as

$$\hat{F}(rr) = \begin{cases} H(\mu_L)G_L(rr), & rr \leq \mu_L \\ H(rr), & \mu_U > rr > \mu_L \\ H(\mu_U)G_U(rr), & rr \geq \mu_U \end{cases}$$

where H is the CDF of the KDE, G_U is the CDF of the GPD for the upper tail, G_L is the CDF of the GPD for the upper tail, μ_L is the threshold of the lower tail GPD, and μ_U is the threshold of the GPD for the upper tail. The combined CDF normalizes the GPDs by the KDE's estimate for the probability of exceeding the threshold. This results in a continuous combined CDF. The normalized lower tail GPD is used below the lower

threshold, the CDF of the KDE is used above the lower threshold and below the upper threshold, and the normalized upper tail GPD is used above the upper threshold. This CDF is equivalent to the sampling method described at the beginning of this section.

CHAPTER 3: APPLICATION

3.1 Knorr-Holden Plot Yield Data

The Knorr-Holden Plot is located in Scottsbluff, Nebraska, and has been continuously planted in corn since 1912. It is administered by the University of Nebraska-Lincoln (UNL) Panhandle Research and Extension Center, Scottsbluff, NE. The Knorr-Holden Plot is the oldest irrigated corn research plot in the U.S. Since 1912, the UNL Panhandle Research and Extension Center has been running a randomized control trial to determine the sustainability of continuous corn in the Sandhills region of Western Nebraska. To facilitate this, starting in 1953, the plot was divided into subplots with distinct nutrient management practices. We chose to use the data from the subplot with the nutrient management practice that most closely resembles that used by producers in the Sandhills region: no manure, 120 lbs. per acre of nitrogen, and 40 lbs. per acre of phosphate. Yield data is available from 1953-2018. Data from the year 2000 is missing due to the plots having been harvested by a combine that did not work properly.

3.2 Scotts Bluff County Yield Data

County-level irrigated corn yield data from Scotts Bluff County is used to compare to the results from the irrigated Knorr-Holden Plot. Yield data for Scotts Bluff County comes from a mixture of NASS and RMA data. NASS data is collected by surveying a representative sample of producers within the county. Participation in these surveys has declined over time and there were only two years after 2012 when there were enough responses to publish a county-level yield estimate. The RMA data is instead

collected from all producers that purchase federally subsidized crop insurance. Accurate reporting to the RMA is mandated by law. The RMA data is only available starting in 1990. To account for the issues with both datasets, it was decided to use NASS data until the year in which the percentage of acres insured in the county passed 60% for the last time. This occurs in 2001 for Scotts Bluff County, so 2001 is the first year in which RMA data is used instead of NASS. From 1990-2001 the difference between RMA and NASS yields is minimal and fluctuates around 0. See Figure 3.1 for a graph of the difference between RMA and NASS yields from 1990-2001. Yield data for the Knorr-Holden Plot is available from 1953-2018, excluding the year 2000. To facilitate a consistent comparison between the data from the two levels of aggregation, we use data from 1953-2018 and exclude the year 2000 from our analysis of Scotts Bluff County.

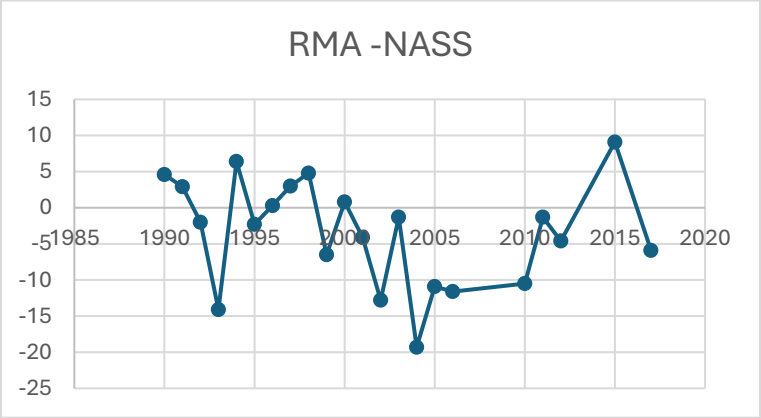

Figure 3.1 Difference Between RMA and NASS Average Yields for Scotts Bluff County

3.3 Preston Farms Yield Data

Preston Farms is in Hardin County, Kentucky and yield data is available from 1980-2023. The county level yield data for Hardin County is available only up to 2022. For comparability, the year 2023 will be excluded from the Preston Farms data.

3.4 Hardin County Yield Data

County-level corn yield data from Hardin County will be used to compare to the results from the Preston Farms data. Yield data for Hardin County comes from a mixture of NASS and RMA data. The NASS data is used until the year in which the percentage of acres insured in the county passed 60% for the last time. This occurs in 2002 for Hardin County, so 2002 is the first year in which RMA data is used instead of NASS data. See Figure 3.2 for a graph of the difference between RMA and NASS yields from 1993 to 2021. From 1998-2002 the difference between RMA and NASS yields is minimal and fluctuates around 0. The difference between RMA and NASS data is minimized in the time around the switch point, 2002. This provides support for our choice to switch to using RMA data in 2002. Yield data for Preston Farms is available from 1980 to 2023, excluding the year 2000. RMA data is available only up to 2022 for Hardin County. To facilitate a consistent comparison between the data from the two levels of aggregation, Hardin County and Preston Farms, we use data from 1980 to 2022 for our analysis of Hardin County.

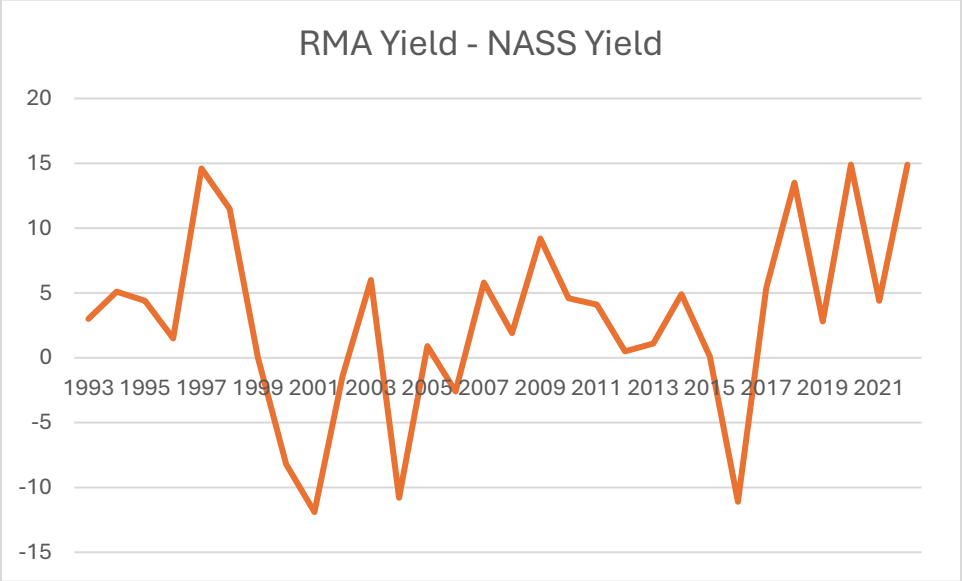

Figure 3.2 Difference Between RMA and NASS Average Yields for Scotts Bluff County

3.5 Knorr-Holden Plot Detrending

The linear detrending methods include OLS, one-knot linear spline, and two-knot linear spline. The first step in detrending the data using linear methods is to fit the one- and two-knot splines to the data. This is done by finding the knot location(s) that minimize the average of the residual ratios squared. The optimal knot location for the one-knot linear spline is 1989. The optimal locations for the two-knot linear spline are 1969 and 1995. OLS is the third of the linear methods but does not require choosing any pre-specified knot locations.

Next, to choose between the three linear methods, the Augmented Dickey-Fuller (ADF) Test is used to test for a stationary distribution of the residual ratios derived from the three different linear methods. The null hypothesis of the ADF test is that the sequence is non-stationary. OLS resulted in an ADF test with a p-value of 0.6044857,

meaning that the test failed to reject the null hypothesis of non-stationarity at the 5% level. The one-knot linear spline resulted in an ADF test with a p-value of 0.08574216, meaning that the test failed to reject the null hypothesis of non-stationarity at the 5% level. The two-knot linear spline was the only one of the linear methods to result in a stationary sequence of residual ratios. The two-knot spline resulted in an ADF test with a p-value of 0.01, rejecting the null hypothesis of non-stationarity of the residual ratios. Thus, the two-knot linear spline is the best detrending model of the linear methods and will be compared to the LOESS method moving forward.

The span parameter used to fit the LOESS curve is chosen through 10-fold cross-validation. A span parameter of 1.0 was found to minimize the average squared error. The ADF test was performed to ensure that the LOESS detrending method resulted in stationary residual ratios. The ADF test resulted in a p-value of 0.03118368, meaning that the null hypothesis of non-stationarity was rejected at the 5% level.

A comparison of the three linear methods and the LOESS method can be viewed below in Figure 3.3. The resulting KDEs from the four detrending methods can be seen in Figure 3.4

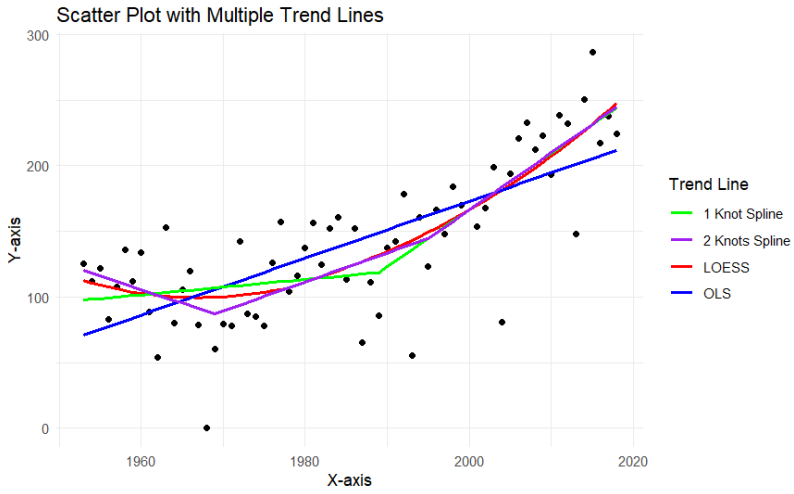

Figure 3.3 Knorr-Holden Plot Trend Yields

The trend yields estimated by the two-knot spline and LOESS models are similar. The KDEs from the two-knot linear spline and LOESS detrending methods are also similar.

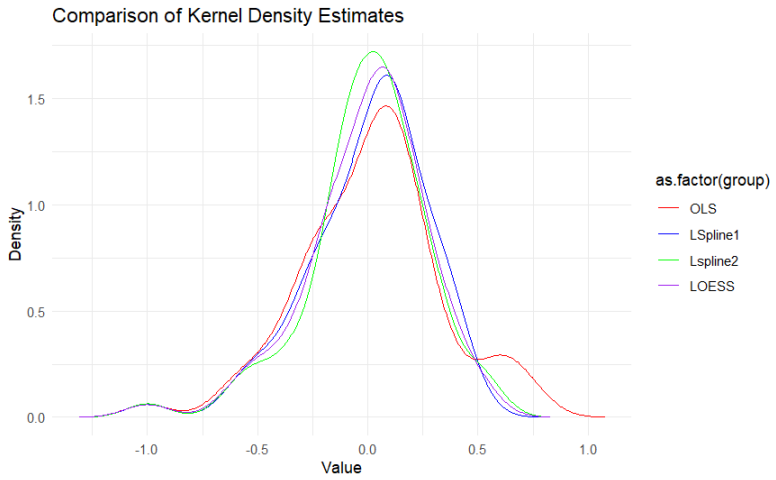

Figure 3.4 Knorr-Holden Plot Kernel Density Estimates

For the Knorr-Holden Plot yield data, the choice of detrending method has little effect on the XI parameters estimated by the GPD. While the parameter estimates are not the same between the two detrending methods, they follow the same pattern and are

generally close to each other. A threshold of 5 does not result in parameter estimates for the GPD because these data are not Pareto distributed. This is due to the small sample size (only five observations). Because the data is not Pareto distributed, the numerical optimization used to fit the GPD fails to converge. See Table 3.1 for the estimated parameters depending on detrending method and threshold choice.

Table 3.1 Knorr-Holden Plot GPD Parameters by Detrending Method

	Lspline2		LOESS	
Threshold	XI	Beta	XI	Beta
4	0.08571	0.177146	0.034683	0.189817
5	NA	NA	-0.29691	0.299013
6	-0.29128	0.316677	-0.31299	0.319107
7	-0.18719	0.278189	-0.14942	0.254594
8	-0.22557	0.301838	-0.41153	0.400775
9	-0.29996	0.350502	-0.31493	0.354632
10	-0.32496	0.376487	-0.18202	0.288721
11	-0.31004	0.376618	-0.12088	0.263047
12	-0.20652	0.319366	-0.02764	0.226352
13	-0.10660	0.271583	0.05718	0.197796
14	-0.03307	0.241680	-0.02289	0.223621
15	0.10596	0.196004	0.05721	0.195916
16	0.19397	0.172159	-0.06019	0.238069

Similarly, for the Knorr-Holden Plot yield data, the choice of detrending method has little effect on the expected shortfall and minimum yields of the final yield distribution. When we reference expected shortfall, we refer to the 5% expected shortfall, or the mean yield conditional on being in the 5% tail of the yield distribution. The 5% expected shortfall provides a simple summary of the risk of extreme, low yield events. See Table 3.2 for the effects of the detrending method on ES and minimum yield. Because the final yield distributions are robust to the choice of detrending method, only

the two-knot linear spline detrending method will be used moving forward to compare with the Scottsbluff County yield distributions.

Table 3.2 Knorr-Holden Plot Minimum Yield and Expected Shortfall by Detrending Method

	Lspline2		LOESS	
Threshold	ES	Min Yield	ES	Min Yield
4	-0.69066	-1	-0.70959	-1
5	NA	NA	-0.74823	-1
6	-0.73679	-1	-0.75904	-1
7	-0.71615	-1	-0.73468	-1
8	-0.72152	-1	-0.79959	-1
9	-0.75807	-1	-0.77594	-1
10	-0.77694	-1	-0.74820	-1
11	-0.78325	-1	-0.73212	-1
12	-0.75672	-1	-0.70930	-1
13	-0.72121	-1	-0.70462	-1
14	-0.70571	-1	-0.70392	-1
15	-0.70713	-1	-0.68905	-1
16	-0.70292	-1	-0.71414	-1

3.6 Scotts Bluff County Detrending

For Scotts Bluff County the optimal knot location for the one-knot spline is the year 1986. For the two-knot spline, the optimal locations are 1975 and 1981. The least complex linear method that results in a stationary sequence of residual ratios is chosen for further analysis. In this case, the two-knot linear spline was the only linear method to produce a stationary sequence of residual ratios.

The LOESS detrending method also has a predetermined parameter, the span parameter, that needs to be chosen through 10-fold cross-validation. For Scottsbluff County, a span parameter of 0.7 was found to be optimal.

A comparison of the three linear methods and the LOESS method can be viewed below in Figure 3.5. The resulting KDEs from the four detrending methods can be seen in Figure 3.6.

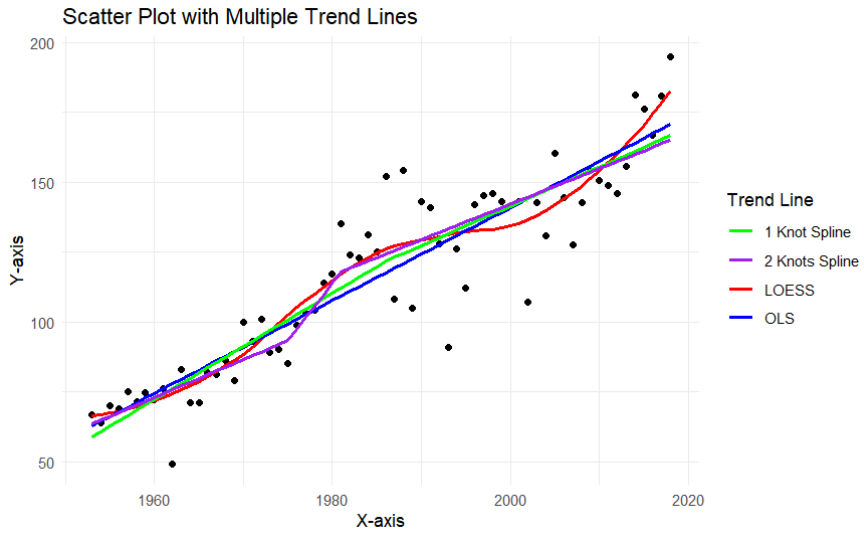

Figure 3.5 Scotts Bluff County Trend Yields

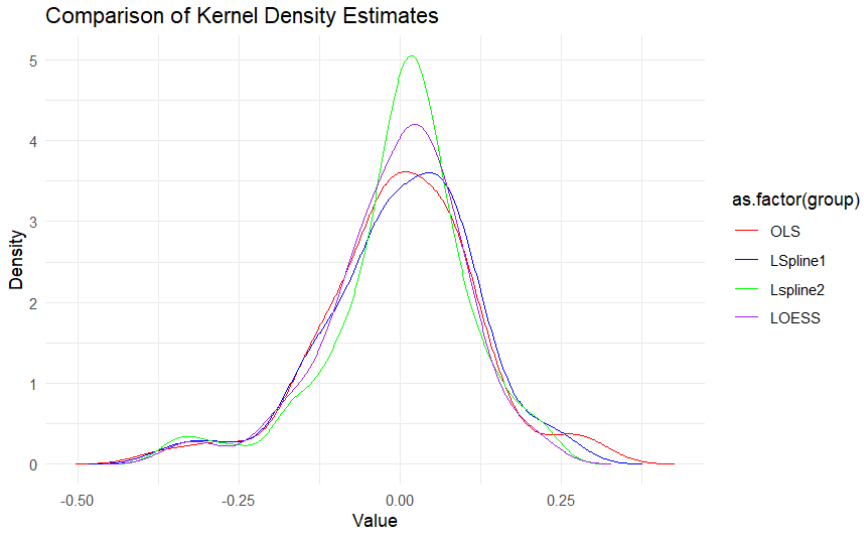

Figure 3.6 Scotts Bluff County Kernel Density Estimates

While the parameter estimates generated from the two detrending methods are not the same, they are similar and follow the same pattern as the threshold changes. The

expected shortfalls and minimum yields are also very similar between the two detrending methods. For this reason, the two-knot linear spline detrending method will be used when comparing the final yield distributions generated from the Knorr-Holden Plots and Scotts Bluff County yield data.

Table 3.3 Scotts Bluff County GPD Parameters by Detrending Method

	Lspline2		LOESS	
Threshold	Xl	Beta	Xl	Beta
10	-0.56101	0.169383	-0.50745	0.151155
11	-0.55507	0.176586	-0.36835	0.130284
12	-0.32666	0.135035	-0.23307	0.110263
13	-0.30776	0.134587	-0.12686	0.095878
14	-0.14223	0.108311	-0.19380	0.106767
15	-0.29635	0.137129	-0.11996	0.096161
16	-0.32198	0.145179	-0.16377	0.103882

Table 3.4 Scotts Bluff County Expected Shortfall and Minimum Yield by Detrending Method

	Lspline2		LOESS	
Threshold	ES	Min Yield	ES	Min Yield
10	-0.2899547	-0.3886244	-0.28637	-0.37891
11	-0.2929531	-0.3908529	-0.27963	-0.41841
12	-0.2856194	-0.4731377	-0.28050	-0.49791
13	-0.2848165	-0.4757767	-0.27596	-0.67538
14	-0.2836967	-0.6741074	-0.27802	-0.54807
15	-0.2871539	-0.4849473	-0.27249	-0.57740
16	-0.2905964	-0.4756361	-0.27658	-0.56101

3.7 Preston Farms Detrending

For Preston Farms, the optimal knot location for the one-knot spline is 1997. The optimal knot locations for the two-knot linear spline are 1997 and 2003. All the linear methods fail to produce a stationary sequence of residual ratios. The one-knot spline is the closest to producing a stationary series of residuals and will be used for further

analysis. The ADF test of the series of residual ratios from the one-knot spline produces a p-value of 0.0502, only just failing to reject the null hypothesis of non-stationarity at the 5% level. For the LOESS method, a span parameter of 1.0 is chosen. The LOESS method produces a stationary series of residual ratios.

A comparison of the detrending methods can be viewed in Figure 3.7. The KDEs resulting from the detrending methods can be viewed in Figure 3.8. The one-knot linear spline and LOESS methods produce very similar KDEs.

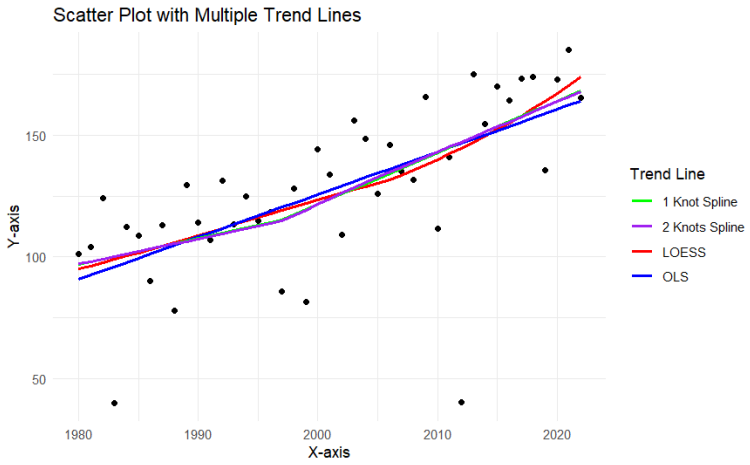

Figure 3.7 Preston Farms Trend Yields

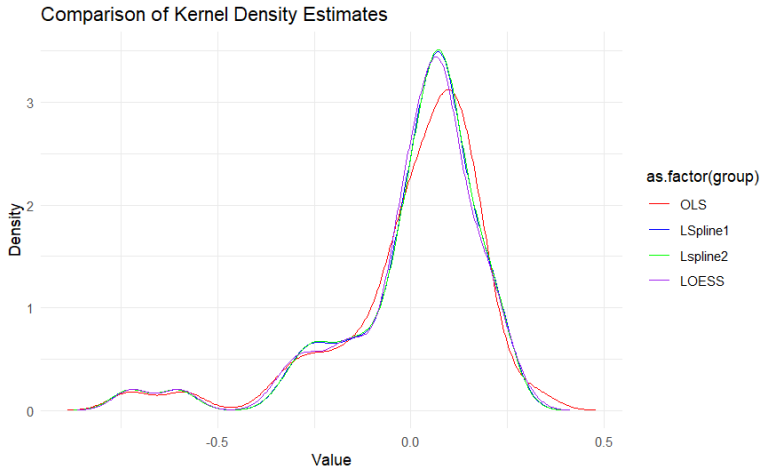

Figure 3.8 Hardin Farms Kernel Density Estimates

The shape parameters estimated from the two detrending methods are similar. The expected shortfalls and minimum yields are also similar. One-knot linear spline will be used for further analysis.

Table 3.5 Preston Farms GPD Parameters by Detrending Method

	Lspline1		LOESS	
Threshold	Xl	Beta	Xl	Beta
7	-0.62798	0.418198	NA	NA
8	-0.08637	0.224881	-0.22346	0.263731
9	-0.53672	0.423015	-0.59818	0.449755
10	-0.29334	0.313432	-0.43535	0.378284
11	-0.24453	0.298161	-0.25931	0.301752

Table 3.6 Preston Farms Expected Shortfall and Minimum Yields by Detrending Method

	Lspline1		LOESS	
Threshold	ES	Min Yield	ES	Min Yield
7	-0.61508	-0.796	NA	NA
8	-0.58993	-1	-0.59454	-1
9	-0.62637	-0.82901	-0.6284	-0.7978
10	-0.61277	-1	-0.62088	-0.88299
11	-0.61028	-1	-0.61462	-1

3.8 Hardin County Detrending

For Hardin County, the optimal knot location for the one-knot linear spline is 1988. For the two-knot linear spline, the optimal knot locations are 1986 and 2012. The two-knot linear spline is the only linear detrending method to result in a stationary sequence of residual ratios and will be used in further analysis. For the LOESS detrending method the optimal span parameter is 1.0.

A comparison of the detrending methods can be viewed in Figure 3.9. A comparison of the KDEs resulting from the different detrending methods can be viewed in Figure 3.10. The LOESS and two-knot linear spline methods produce similar KDEs.

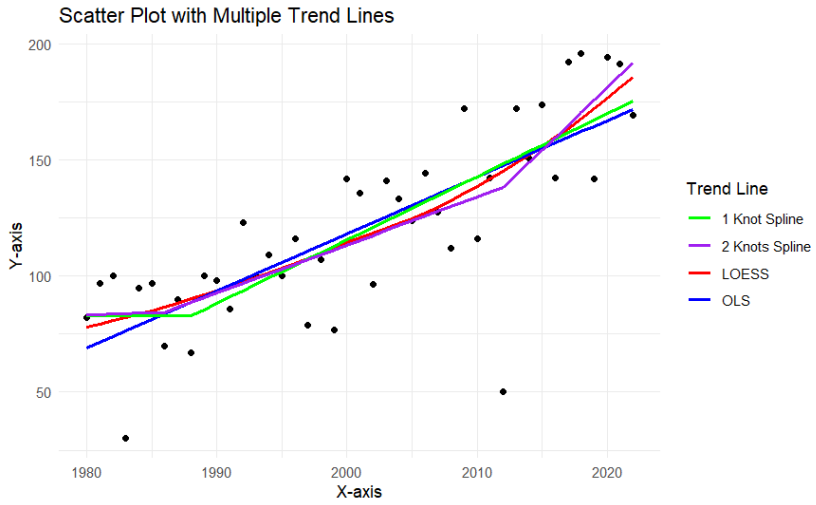

Figure 3.9 Hardin County Trend Yields

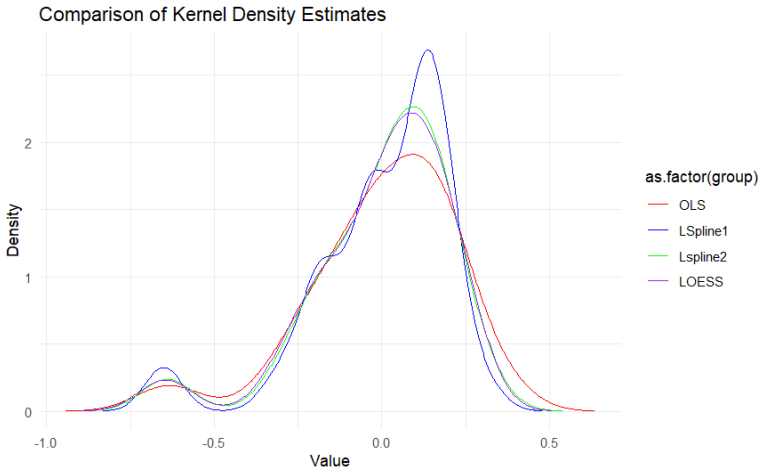

Figure 3.10 Hardin County Kernel Density Estimates

Both detrending methods produced a similar range of parameter estimates. The two methods resulted in shape parameters around 0.1. The expected shortfall and minimum yield estimates from the two detrending methods are similar. The two-knot linear spline method will be used for further analysis.

Table 3.7 Hardin County GPD Parameters by Detrending Method

	Lspline2		LOESS	
Threshold	Xl	Beta	Xl	Beta
7	0.11971	0.161019	0.12028	0.160801
8	-0.10267	0.210597	0.19325	0.142371
9	0.12223	0.154747	0.37575	0.107032
10	0.10741	0.155358	-0.12873	0.216979
11	0.15400	0.142610	-0.03557	0.189656

Table 3.8 Hardin County Expected Shortfall and Minimum Yields by Detrending Method

	Lspline2		LOESS	
Threshold	ES	Min Yield	ES	Min Yield
7	-0.58142	-1	-0.59287	-1
8	-0.59134	-1	-0.58687	-1
9	-0.57359	-1	-0.57330	-1
10	-0.58880	-1	-0.59094	-1
11	-0.56513	-1	-0.58141	-1

3.9 Knorr-Holden Plot Threshold Selection

Figure 3.11 contains the mean excess plot, the parameter stability graph, and a plot of the data by inverse rank for the Knorr-Holden Plot. In the parameter stability graph, the shape parameter has a negative linear trend in threshold until a threshold of 12, after which the estimate of ξ stabilizes around -0.25. This stability beyond the threshold of 12 provides evidence for the use of a threshold of 12. The red vertical line in the mean excess plot represents the 25^{th} percentile of residual ratios. We do not consider thresholds beyond the 25^{th} percentile. The mean excess plot has two clear trends. From a threshold of 16 (the first observation to the right of the red vertical line) until observation 12 there is a clear linear increase in mean excess. From the threshold of 12 on there is a clear linear decrease with threshold. The evidence for choosing a threshold of 12 is strengthened by the linearity of the mean excess above this value, which is not maintained beyond the threshold of 16. A threshold of 12 will be used for further analysis.

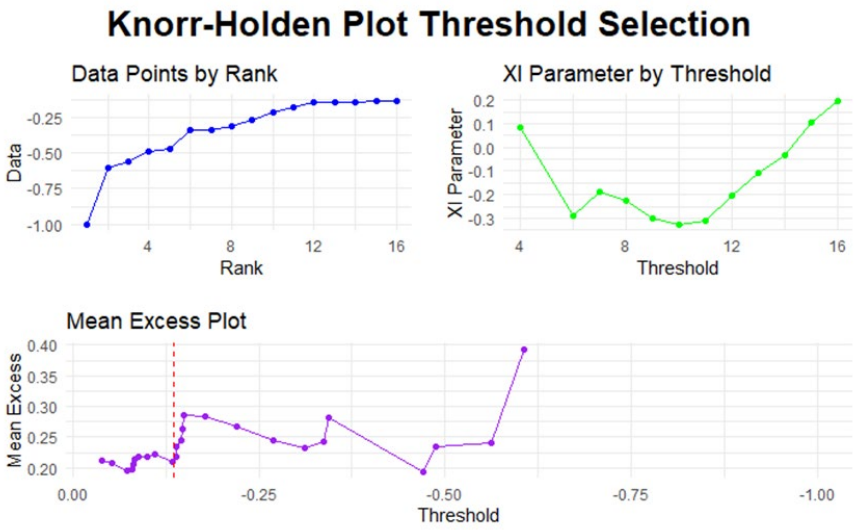

Figure 3.11 Knorr-Holden Plot Threshold Selection

3.10 Scotts Bluff County Threshold Selection

Figure 3.12 displays the mean excess plot, parameter stability graph, and a graph of the residual ratios by inverse rank for Scotts Bluff County. There is stability in the estimated shape parameter beyond the highest possible threshold, 16. There is also a clear linear decrease in mean excess beyond the threshold of 16. Both these factors provide corroborating evidence for the use of threshold 16 for further analysis.

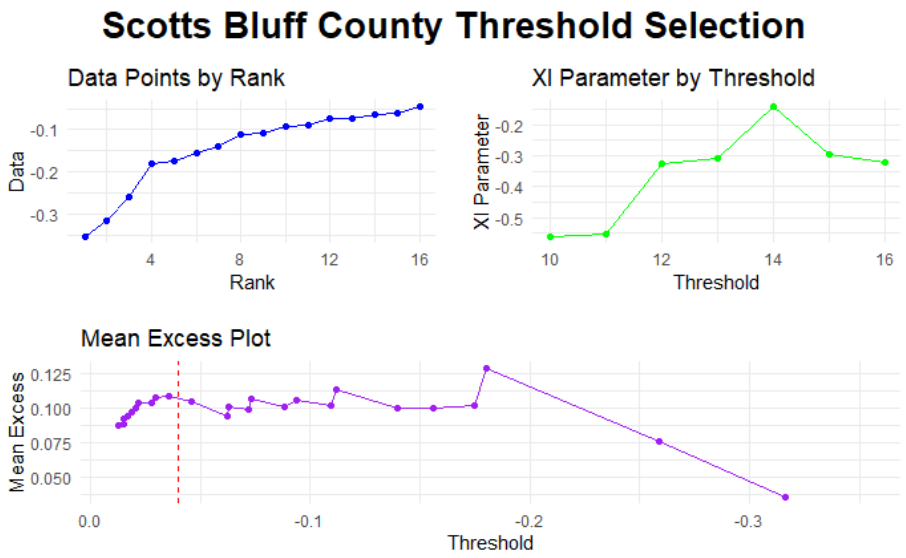

Figure 3.12 Scotts Bluff County Threshold Selection

3.11 Preston Farms Threshold Selection

Figure 3.13 contains the mean excess plot, parameter stability plot, and a plot of residual ratios by inverse rank for Preston Farms. The parameter stability plot shows stability past a threshold of 11 around a ξ value of -0.3. The mean excess plot also shows a stable linear decline in mean excess beyond a threshold of 11. In the mean excess plot, the vertical red line characterizes the threshold beyond which the data is no longer Pareto-distributed. Beyond the threshold of 11, the mean excess is no longer linear in threshold. There is clear evidence to select a threshold of 11 for further analysis.

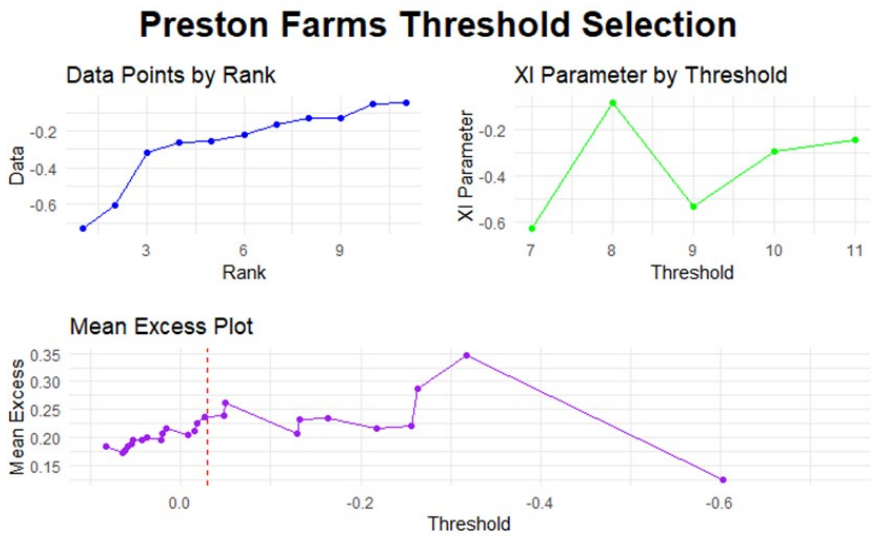

Figure 3.13 Preston Farms Threshold Selection

3.12 Hardin County Threshold Selection

Figure 3.14 contains the mean excess plot, parameter stability plot, and a plot of residual ratios by inverse rank for Hardin County. There is stability in the parameter estimates beyond threshold 11 around a ξ value of 0.1. There is also a clear, positive linear trend in mean excess beyond the threshold of 11. The red vertical line in the mean excess plot represents the threshold beyond which the residual ratios are no longer Pareto distributed. Beyond the threshold represented by the red vertical line, there is no longer a clear linear relationship between threshold and mean excess. The parameter stability graph and mean excess plot both provide evidence for using a threshold of 11 for further analysis.

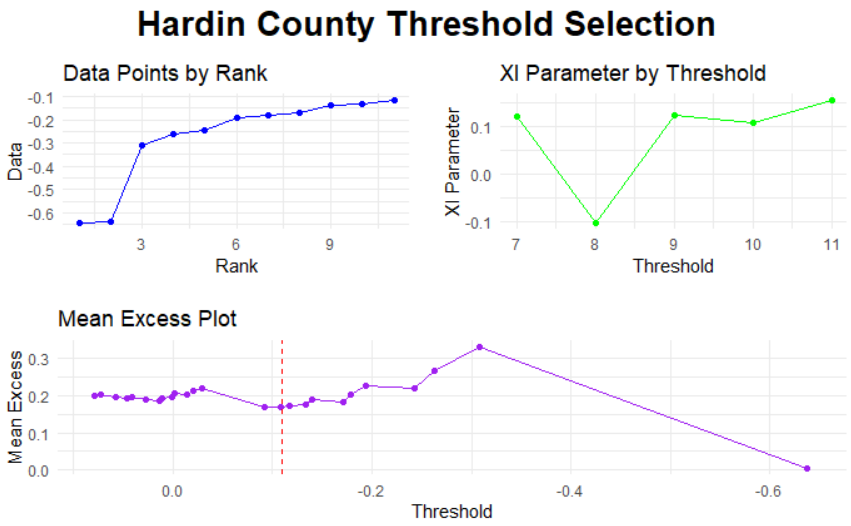

Figure 3.14 Hardin County Threshold Selection

CHAPTER 4: RESULTS AND CONCLUSION

In this study, the yield distributions are defined over residual ratios, not bushels per acre. This is to remove the effect of different trend yields on our comparison of the distributions. The shape of the distribution is what is important in our analysis. To convert the distribution defined over residual ratios to one defined over yield (bu/a), simply multiply the residual ratios by the trend yield.

4.1 Hardin County and Preston Farms

The Hardin County and Preston Farms yield distributions are strikingly similar. One measure of this similarity is the average (integrated) squared difference (ASD) between the two CDFs. The ASD for Hardin County and Preston Farms is 0.0007 residual ratios squared. The average distance is then .026 residual ratios or only about 2% of trend yield. The two distributions have almost the same variance. The Preston Farms yield distribution is skewed more to the left and thus has a lower expected shortfall by about 5% of trend yield. The Preston Farms distribution has a large mass just above trend yield, is skewed to the left, and has a long left tail. This corresponds to the historical yields from Preston Farms. The threshold selected is at the 25th percentile of the residual ratios but is just less than trend yield. In contrast, Scotts Bluff County has a more symmetric distribution relative to Preston Farms. The Scotts Bluff County yield distribution is still skewed to the left but has the potential to exceed trend yield by more than Preston Farms. Overall, Preston Farms' yield distribution would be well approximated by the yield distribution of Hardin County.

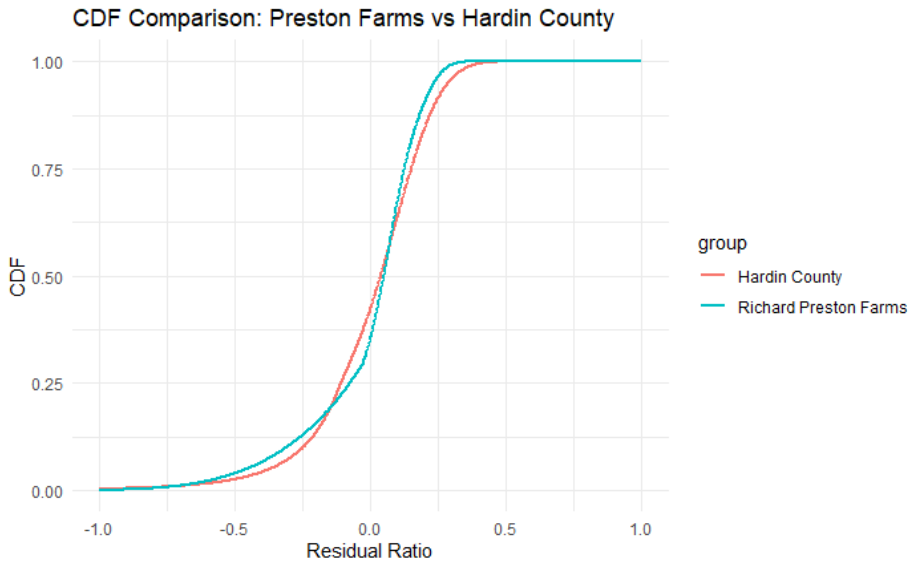

Figure 4.1 Preston Farms and Hardin County CDFs

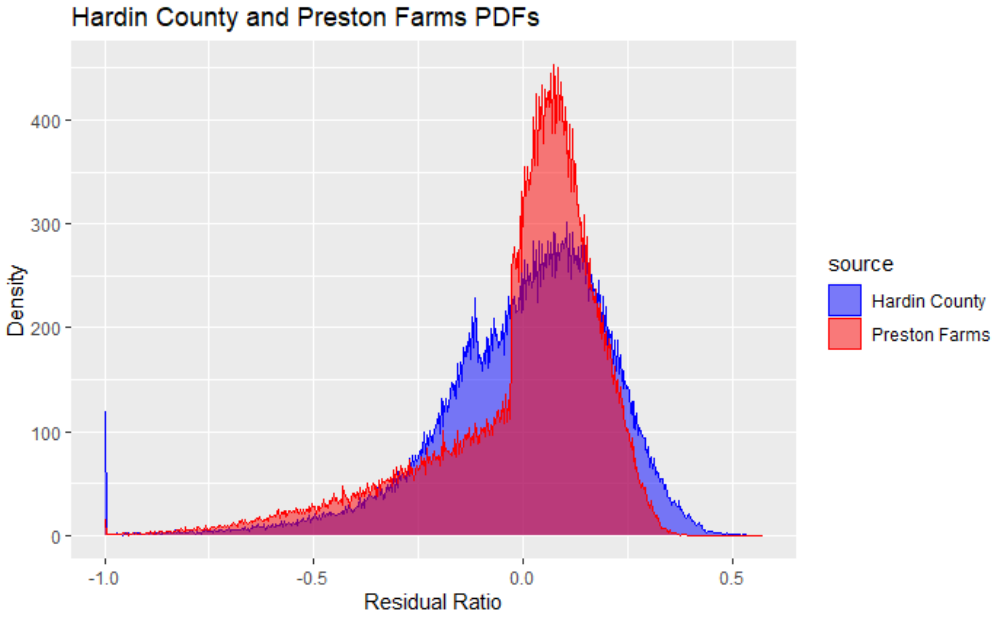

Figure 4.2 Preston Farms and Hardin County PDFs

Table 4.1 Summary Statistics for Preston Farms Sampled Values

Statistic	Value
Minimum	-1.000000000
1st Quartile (Q1)	-0.078000000
Median	0.046000000
3rd Quartile (Q3)	0.122000000
Maximum	0.418000000
Mean	-0.009110333
Standard Deviation	0.209525714
Skewness	-1.477927766
Kurtosis	5.365097337
Expected Shortfall	-0.610442634

Table 4.2 Summary Statistics for Hardin County Sampled Values

Statistic	Value
Minimum	-1.0000000
1st Quartile (Q1)	-0.1120000
Median	0.0340000
3rd Quartile (Q3)	0.1480000
Maximum	0.5740000
Mean	0.0032921
Standard Deviation	0.2104398
Skewness	-1.1450479
Kurtosis	5.5924728
Expected Shortfall	-0.5638167

4.2 Scotts Bluff County and Knorr-Holden Plot

The Scotts Bluff County and Knorr-Holden Plot yield distributions are substantially different from each other. The ASD between the Scotts Bluff County and the Knorr Holden Plot yield distributions is 0.0069 residual ratios squared. This amounts to an average difference between the two CDFs of 8.3% of trend yield. The standard deviation of the Knorr-Holden Plot yield distribution is over twice as large as that for the Scotts Bluff County yield distribution. The Scotts Bluff County distribution is symmetrical with a skew of close to 0. The Knorr-Holden Plot is skewed to the left and has a long tail. The Scotts Bluff County yield distribution would be an extremely poor approximation of the Knorr-Holden Plot yield distribution. Adjusting the Scotts Bluff

County distribution using the coefficient of variation method would be a better approximation, but still fails to well approximate the Knorr-Holden Plot data. Scotts Bluff County has a symmetric yield distribution while the Knorr-Holden plot has a left skewed distribution.

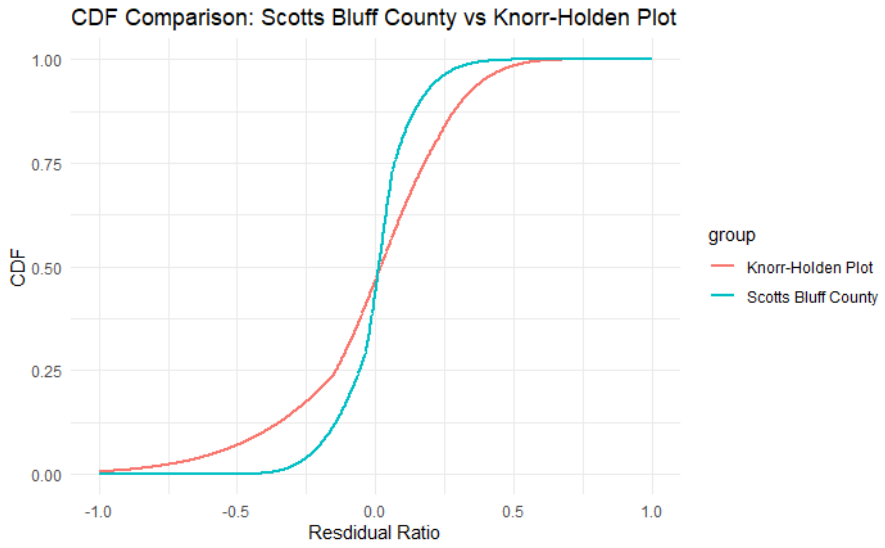

Figure 4.3 Scotts Bluff County and Knorr-Holden Plot CDFs

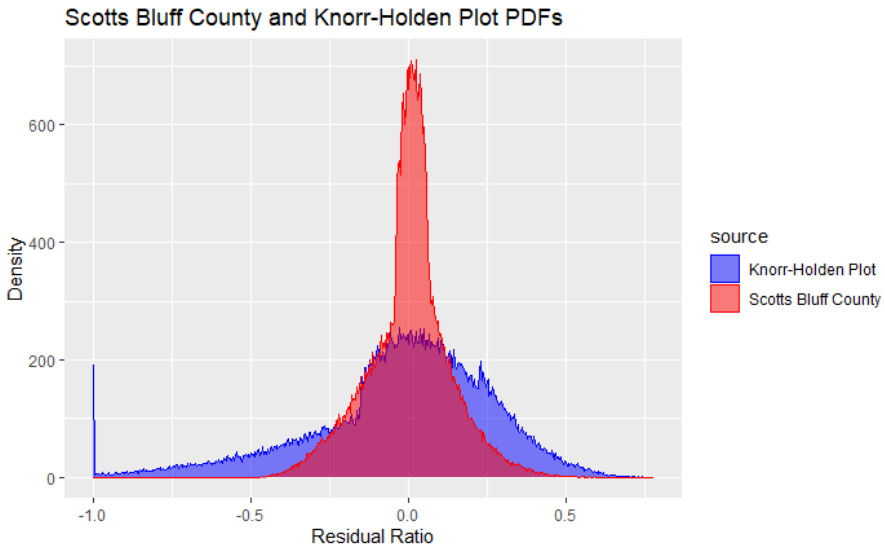

Figure 4.4 Scotts Bluff County and Knorr-Holden Plot PDFs

Table 4.3 Summary Statistics for Knorr-Holden Plot Sampled Values

Statistic	Value
Minimum	-1.0000000
1st Quartile (Q1)	-0.1440000
Median	0.0160000
3rd Quartile (Q3)	0.1760000
Maximum	0.7800000
Mean	-0.0161746
Standard Deviation	0.2872016
Skewness	-0.7874022
Kurtosis	3.8954562
Expected Shortfall	-0.7552289

Table 4.4 Summary Statistics for Scotts Bluff County Sampled Values

Statistic	Value
Minimum	-0.468000000
1st Quartile (Q1)	-0.060000000
Median	0.010000000
3rd Quartile (Q3)	0.068000000
Maximum	0.650000000
Mean	0.004124567
Standard Deviation	0.130842383
Skewness	-0.051420392
Kurtosis	4.004751113
Expected Shortfall	-0.292868230

4.3 Discussion

An important question that arises from these results is why Preston Farms and Hardin County have similar yield distributions while the Knorr-Holden Plot and Scotts Bluff County distributions are substantially different. Farm size represents the primary reason. Preston Farms accounts for about 1% of the corn acres in Hardin County and the acreage is relatively spread out within the county. The Knorr-Holden Plot, being just a field, makes up a negligible portion of the irrigated corn acres in Scotts Bluff County.

Another reason for the similarity between Hardin County and Preston Farms is that corn is grown without irrigation throughout Hardin County and on Preston Farms. Thus, drought is the biggest risk to yields for both Hardin County and Preston Farms. Drought is relatively non-local; if Hardin County experiences drought, so does Preston Farms, and vice versa. In contrast, the Knorr-Holden plot is irrigated and is being compared to Scotts Bluff County's irrigated corn yield. Irrigation limits the effects of drought on yield. While drought is generally a non-local phenomenon, hail is a localized risk. Hail is the primary source of yield risk for the Knorr-Holden Plot. In any year only a small portion of the area of Scotts Bluff County will be hailed on, so any effects on yield will be averaged out of the county-level yield data. The risk of hail only shows up in more disaggregated levels of yield data.

Additional determinants of the relationship between farm and county level yield distributions include the farming practices implemented by the producer and the soil characteristics of the land being farmed. Preston Farms has a diversity of land quality and planting times. These practices were implemented prior to the implementation of revenue

protection crop insurance and were designed to be anti-fragile. It is important for the risk manager to consider non-financial methods of risk mitigation.

There are drawbacks associated with using the field-level data from the Knorr-Holden research plot. First, it may be argued that, as a research plot, it is not operated with the goal of maximizing profits, in contrast to the county-level data to which it is being compared. While this is true, the plot is being operated in a way that responds to changing conditions to at least somewhat approximate the behavior of the producers in the surrounding Scotts Bluff County. Also, we chose to use data from the trial that best approximates the fertilizer application practices of the surrounding Scotts Bluff County. Additionally, while the Knorr-Holden plot is very small, we believe that it reasonably well approximates a section or quarter section that is often insured as a unit. Spatial heterogeneity is an important consideration in both the producers understanding of yield risk as well as the rating of crop insurance.

4.4 Conclusion

In this paper, a new method for generating yield distributions from historical yield was proposed and applied to four series of historical yields. The method employs EVT to accurately estimate the probability of extreme, low yield events. EVT is combined with a KDE to form a complete yield distribution.

Before applying EVT and KDEs, the historical yield data was detrended using LOESS and linear detrending methods. The two detrending methods performed similarly. Linear methods were used for later analysis because of the similar performance and their

ability to predict future yields. Linear methods are also simpler to implement and are less susceptible to overfitting. Once the model of trend yields was established, a series of residual ratios was created to overcome the issue of heteroskedasticity. Yield distributions were then fit to the series of residual ratios generated from each historical dataset.

The next step in the process was to fit EVT to the tails of the residual ratios. A key contribution of this paper is an in-depth explanation of how to apply EVT to small sample historical yield data. Threshold selection is crucial to accurately estimating the probabilities of extreme events. For small samples, mean excess and parameter stability plots were the best methods for threshold selection. KDEs were then fit to the residual ratios to model the main body of the yield distribution. The KDEs were combined with the GPDs to form a complete yield distribution

When applying our method to four different datasets, we found a heterogenous relationship between county and farm/field level distributions. Several explanations were hypothesized. The primary factor determining the relationship between the farm and county-level yield distribution is the size and geographic diversity of the farm within the county. Secondary influences include the primary climactic risks to yield and farmers' management practices.

For future research, our method could be applied to more farm/county combinations to improve the understanding of the relationship between tail risk in the farm and county level yield distributions. Further, including weather events from surrounding farms that the evaluated farms did not experience would improve EVT estimation, leading to improvements in farm yield estimation. Additionally, the improved

yield distributions will be applied to the risk management model of Walters and Preston to improve their estimate of the risk of extreme income events and improve producer decision making under uncertainty.

REFERENCES

Benito, S., López-Martín, C., & Navarro, M. Á. (2023). Assessing the importance of the choice threshold in quantifying market risk under the POT approach (EVT). *Risk Management*, *25*(1), 6.

Claassen, R., & Just, R. E. (2011). Heterogeneity and distributional form of farm-level yields. *American journal of agricultural economics*, *93*(1), 144-160.

Cleveland, W. S. (1979). Robust locally weighted regression and smoothing scatterplots. *Journal of the American Statistical Association*, *74*(368), 829-836.

Coles, S., Bawa, J., Trenner, L., & Dorazio, P. (2001). An introduction to statistical modeling of extreme values (Vol. 208, p. 208). London: Springer.

Cooper, J. C., Langemeier, M. R., Schnitkey, G. D., & Zulauf, C. R. (2009). Constructing farm level yield densities from aggregated data: analysis and comparison of approaches.

Das, B., & Ghosh, S. (2016). Detecting tail behavior: mean excess plots with confidence bounds. *Extremes*, *19*, 325-349.

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.

Ghosh, S., & Resnick, S. (2010). A discussion on mean excess plots. *Stochastic Processes and their Applications*, *120*(8), 1492-1517.

Harri, A., Coble, K. H., Ker, A. P., & Goodwin, B. J. (2011). Relaxing heteroscedasticity assumptions in area-yield crop insurance rating. *American Journal of Agricultural Economics*, *93*(3), 707-717.

Morgan, W., Cotter, J., & Dowd, K. (2012). Extreme measures of agricultural financial risk. *Journal of Agricultural Economics*, 63(1), 65-82.

Ker, A. P., & Coble, K. (2003). Modeling conditional yield densities. *American Journal of Agricultural Economics*, *85*(2), 291-304.

Kim, Y., Yu, J., & Pendell, D. L. (2020). Effects of crop insurance on farm disinvestment and exit decisions. *European Review of Agricultural Economics*, *47*(1), 324-347.

Lu, J., Carbone, G. J., & Gao, P. (2017). Detrending crop yield data for spatial visualization of drought impacts in the United States, 1895–2014. *Agricultural and forest meteorology*, *237*, 196-208.

Madsen, H., Rasmussen, P. F., & Rosbjerg, D. (1997). Comparison of annual maximum series and partial duration series methods for modeling extreme hydrologic events: 1. At-site modeling. *Water resources research*, *33*(4), 747-757.

Park, E., Brorsen, B. W., & Harri, A. (2019). Using Bayesian Kriging for spatial smoothing in crop insurance rating. *American Journal of Agricultural Economics*, 101(1), 330-351.

Pfaff B, McNeil A (2018). _evir: Extreme Values in R_. R package version 1.7-4,

<https://CRAN.R-project.org/package=evir>.

Tolhurst, T. N., & Ker, A. P. (2015). On technological change in crop yields. *American Journal of Agricultural Economics*, *97*(1), 137-158.

Walters, C., & Preston, R. (2018). Net income risk, crop insurance and hedging. *Agricultural Finance Review*, 78(1), 135-151.

Ye, T., Nie, J., Wang, J., Shi, P., & Wang, Z. (2015). Performance of detrending models of crop yield risk assessment: evaluation on real and hypothetical yield data. *Stochastic environmental research and risk assessment*, *29*, 109-117.