NBER WORKING PAPER SERIES

BREAKING BOUNDARIES: LOWER TAIL DEPENDENCE CAN TRIPLE THE ECONOMIC VALUE OF INDEX INSURANCE FOR RURAL HOUSEHOLDS

Enrique Estefania-Salazar
Michael Carter
Eva Iglesias
Álvaro Escribano

Working Paper 32618
http://www.nber.org/papers/w32618

NATIONAL BUREAU OF ECONOMIC RESEARCH
1050 Massachusetts Avenue
Cambridge, MA 02138
June 2024

Enrique Estefania-Salazar was supported by the Programa Propio de I+D+i 2021 of the Universidad Politécnica de Madrid (Grant # PREDOC-21-KB36HA-52-2L7YEQ). The views expressed herein are those of the authors and do not necessarily reflect the views of the National Bureau of Economic Research.

Breaking Boundaries: Lower Tail Dependence Can Triple the Economic Value of Index Insurance for Rural Households

Enrique Estefania-Salazar, Michael Carter, Eva Iglesias, and Álvaro Escribano

NBER Working Paper No. 32618
June 2024
JEL No. G22,O13,O16

ABSTRACT

Despite its promise to help low-wealth households manage climate risk, index insurance remains hampered by downside basis risk, meaning that an insured party suffers a loss, but receives no payment because the insurance index fails to register a loss. While efforts to reduce basis risk focus on the creation of indices that better predict losses, this paper focuses on the creation of statistically rigorous insurance zones that minimize downside basis risk. In contrast to our approach, most existing index insurance contracts use statistically ad hoc administrative boundaries to demarcate insurance zones. To improve on this practice, we develop a machine learning algorithm that forms zones to maximizes lower tail dependence within the zone. After exploring the logic for using lower tail dependence, we apply our algorithm to the long-running index-based livestock insurance contract in Northern Kenya. We show that compared to the currently employed administrative insurance areas, our algorithm creates an insurance contract that increases the expected utility value of the insurance by 60-200% even when keeping the number of zones the same. Optimizing the number of zones using our method can further increase the economic value of index insurance to its beneficiaries.

Enrique Estefania-Salazar
Research Centre for the Management of Agricultural and Environmental Risks (CEIGRAM)
Universidad Politécnica de Madrid
Madrid, 28040
Spain
enrique.esalazar@upm.es

Eva Iglesias
Research Centre for the Management of Agricultural and Environmental Risks (CEIGRAM)
Universidad Politécnica de Madrid
Madrid, 28040
Spain
eva.iglesias@upm.es

Michael Carter
Department of Agricultural and Resource Economics
University of California, Davis
One Shields Avenue
Davis, CA 95616
and NBER
mrcarter@ucdavis.edu

Álvaro Escribano
Department of Economics
Universidad Carlos III de Madrid
Getafe, 28093
Spain
alvaro.escribano@uc3m.es

Breaking Boundaries: Lower Tail Dependence Can Triple the Economic Value of Index Insurance for Rural Households

Enrique Estefania-Salazar, Michael R. Carter, Eva Iglesias and Álvaro Escribano*

Abstract

Despite its promise to help low-wealth households manage climate risk, index insurance remains hampered by problems of downside basis risk, meaning that an insured party suffers a loss, but receives no payment because the underlying insurance index fails to register a loss. While most efforts to reduce basis risk have focussed on the creation of indices that better predict losses, this paper focusses on the creation of statistically rigorous insurance zones that minimize downside basis risk for a given loss predictor. In contrast to our approach, most existing index insurance contracts use convenient, but statistically *ad hoc* administrative boundaries (*e.g.*, counties or districts) to demarcate insurance zones. To improve on this practice, we develop a machine learning algorithm that maximizes lower tail dependence within the insurance zones. We show that maximizing lower tail dependence is statistically equivalent to minimizing downside basis risk. After exploring the logic for using the maximization of lower tail dependence to form zones, we apply our algorithm to a long-running index-based livestock insurance contract in Northern Kenya. We show that compared to the currently employed administrative insurance areas, our algorithm creates an insurance contract that increases the expected utility value of the insurance by 60-200% even when keeping the number of zones the same. Optimizing the number of zones using our method can further increase the economic value of index insurance to its beneficiaries.

Keywords: Index insurance, Basis risk, Machine Learning, Drought, Lower tail dependence

*Estefania-Salazar and Iglesias, Research Center for the Management of Agricultural and Environmental Risks (CEIGRAM), Madrid, 28040, Spain andDepartment of Agricultural Economics, Statistics and Business, Universidad Politécnica de Madrid, Madrid, 28040, Spain; Carter NBER, University of Cape Town and Department of Agricultural and Resource Economics, University of California Davis, Davis, CA, 95616 USA; Escribano, Department of Economics, Universidad Carlos III de Madrid, Getafe, 28093, Spain

1 Introduction

While agricultural index insurance was first introduced over 100 years ago, it has remerged over the last two decades as a potential solution to the age-old problem that uninsured risk dampens technological change and productivity growth, especially amongst small-scale agriculturalists (see Benami et al., 2021). Index insurance operates by grouping farmers into insurance zones. Payments are then made based on a single verifiable index that proxies for farmer losses in the zone (*e.g.*, an index of weather conditions, plant growth or crop yields). The index averages or otherwise aggregates multiple pieces of information from within the zone. Area yield indexes, for example, average yields across all farm units within the insurance zone, while indices based on high resolution earth observation data aggregate signals derived from all earth observation pixels located within the zone.[1] Aggregation is critical to controlling moral hazard by assuring that no single farmer can significantly influence the zone index by his or her actions or inactions. The fundamental strength of index insurance is that it does not require costly individual loss verification, making it possible to offer insurance to small-scale agriculturalists whose small coverage levels make loss verification uneconomical.

Unfortunately, this fundamental advantage of index insurance is also its Achilles heel as the index may fail to accurately capture losses experienced by the insured farmers. This inaccuracy in insurance index is typically labelled basis risk (Clarke, 2016), and it comes in two flavors. Downside basis risk (or a "false negative" signal) occurs when the index does not register an insurable loss and issues no payment when in fact the insured has actually suffered a loss. Upside basis risk (or a "false positive" signal) occurs when the index signals a loss and issues payment when in fact the individual has not suffered a loss. Both false negatives and false positives diminish the economic value of insurance when measured using an expected utility welfare metric (Clarke, 2016, Kenduiywo et al., 2021 and (Lichtenberg and Iglesias, 2022)).[2]

[1] Benami et al., 2021 discuss the history of satellite-based earth observation and increasingly high resolution of measures of plant growth, soil moisture and weather. For example the bio-mass growth measure (NDVI) used in Section 5's empirical analysis is currently available in resolutions as high as 3 meters by 3 meters, with many index contracts relying on the 10 meters by 10 meters pixels available from the European Sentinel satellites.

[2] Although it may seem that upside basis risk is a good thing for the insured, Kenduiywo et al. (2021) analyze why it is not. Intuitively, for realistic, actuarially unfavorable insurance, the insured pays more than a dollar to get 1 dollar

Following Elabed et al. (2013), basis risk of both flavors can be traced to one of two components, design-based basis risk and idiosyncratic or zone-based basis risk. Design-based basis risk occurs when the index fails to accurately capture *average* losses in the insurance zone. Idiosyncratic or zone-based basis risk occurs when the losses of an individual within the insurance zone deviates from the average losses in the zone. Because the extent of idiosyncratic risk depends on the size and construction of the insurance zone, we will also refer to this type of basis risk as zonal risk. Individual studies have shown both types of basis risk to be important in particular contexts. Clarke et al. (2012) show that a weather index used in India poorly predicts average losses, whereas Jensen et al. (2019) study of livestock index insurance in Kenya comment that idiosyncratic risk is the largest component of basis risk.

The rapid advance of remote sensing technologies and data analytics has underwritten efforts to create new indices that better predict losses within a zone and thereby reduce design-based basis risk (see Benami et al., 2021). Much less attention has been devoted to improving insurance zones, despite its potential importance, as discussed by Vrieling et al. (2014) and Benami and Carter (2021). Stigler and Lobell (2023) gauge the degree to which standard insurance zones, based on administrative boundaries, are responsible for basis risk. In their analysis, they calculate the correlation between units contained within these zones. They conclude that administrative zones are highly heterogeneous (exhibiting modest correlation amongst the individual units contained in the zone) and are as much responsible for basis risk as are bad indexes that poorly predict average outcomes in a zone. While this observation is valuable, this paper argues linear correlation is unlikely to be the most appropriate metric to either evaluate the quality of zone structures (as Kenduiywo et al. (2021) stress) or to optimally construct them.

Responding to this evidence, this paper's primary contribution is a method to break administrative boundaries and optimally design contiguous index insurance zones in order to minimize downside basis risk. We rely on non-linear statistics (specifically the coefficient of lower tail dependence, or LTD) to define these optimal insurance zones.[3] We show that using LTD to design insurance zones

in payout. It a wealth or income reducing shock made the shadow value of money, then paying more than dollar to get a dollar is utility. Paying more than a dollar to get a dollar in a good state of the world (when the shadow price of money is low and a "dollar is worth a dollar) is utility reducing.

[3]Conradt et al. (2015) and Bokusheva (2018) use lower tail-sensitive non-linear methods (quantile regression- and

is almost always preferable to using linear correlation to form zones. Intuitively, a relatively high correlation can occur even when correlation in the lower tail of the distribution is low. It is in fact the behavior of the lower tail that is most important for insurance value (see Kenduiywo et al. (2021)), and we show that designing zones in order to maximize the LTD amongst farm or pixel units in the zone is equivalent to minimizing the probability of "false negative" basis risk events. We also show that the methods developed here can be used to create correlation maximizing zones, but empirically we find that these zones result in lower quality insurance than LTD-based zones.

A second contribution of this paper is its use of an economic metric to value the impact of optimized insurance zone design on the quality of the index insurance. This approach allows us to show that optimal designs are not simply a statistical nicety, but in the case of our case study of livestock insurance in Kenya can lead to a tripling of the impact on insurance on the certainty equivalent well-being of insured families. Third and finally, the approach we develop has the advantage that it can consider a large number of variables/periods and be applied to very dense grids due to the use of powerful machine learning techniques. This is of particular relevance given the ever-increasing frequency and resolution of remote sensing data that be used to create insurance indexes.

The remainder of this paper is organized as follows. Section 2 covers the basic economics of index insurance, including an approach to measure the economic value of an index insurance contract. Section 3 introduces the key concept of lower tail dependence. Using simulated data that vary the correlation and the LTD between an individual and an index, the section also shows why in most instances LTD is more important than simple correlation for the quality of index insurance. It also shows that if the coefficient of LTD is below a threshold value, then there is no index insurance contract that is welfare-improving for moderately risk averse individuals.

Section 4 then shows how to aggregate multiple individuals or pixel units of observation into economically optimal insurance zones. The section begins with simulated data to explore the relatively simple case in which contiguity of the insurance zones does not matter. Building on this analysis, the second half of the section moves to the complexity of real world data in which zones must be

copula-based analysis) to better design or evaluate index insurance contracts. Nevertheless, non linear measures have not been used for improving the spatial-design quality of index insurance programs.

contiguous and the number of pixel or individual units may be very large, as they are in the case of indexes based on high resolution remote-sensing data. Using these building blocks, Section 5 then tests this methodology in the pastoral areas of Marsabit district in Northern Kenya, an area where a remote sensing-based index insurance is currently offered using 14 local administrative units as the basis for insurance zones. Using the LTD maximization criterion to group remote sensing pixels into relatively homogenous groups, we show that the resulting optimized insurance zones double the expected utility value of the livestock insurance for communities in Marsabit compared to the currently existing structure. To make this welfare valuation, we draw on 7 years livestock mortality data from 903 households in the region. We also show that redesigning the insurance zones to maximize the correlation amongst units (in the spirit of Stigler and Lobell (2023)) results in an improvement over the existing administrative zones, but that improvement is less than what is possible when zones are designed to maximize LTD. Finally, Section 6 concludes.

2 The Economics of Index Insurance

As a prelude to improving index insurance, this section summarizes the literature on index insurance and basis risk. It also summarizes an expected utility-based approach to measuring index insurance quality suggested by Carter and Chiu (2018) and Kenduiywo et al. (2021). This approach allows evaluation of whether any index insurance contract is worth buying (better than nothing) as well as determining which of multiple contract zone design schemes would be expected to most increase the economic well-being of the insured.

2.1 Index insurance and basis risk

Traditional agricultural insurance confronts many problems related to transaction costs, information asymmetries, adverse selection and moral hazard. Index insurance can in principal resolve these issues (Cole and Xiong, 2017). Moreover, index insurance is an especially suitable alternative to insure populations living in remote rural areas where the accessibility might be restricted or there are no qualified experts for the appraisals, which may be the case of many developing countries. As

a result, international organizations such as the World Bank are making strong efforts to promote the development of index insurance in developing countries (See Global Index Insurance Facility[4]). Index insurance is becoming increasingly important in Africa with the hope that it will increase increase the potential and adoption of new technologies (Farrin and Miranda, 2015; Carter et al., 2016; Tang et al., 2019). Likewise, the advance of digital technologies represents an opportunity for the implementation of this type of insurance in the rural world of sub-Saharan Africa (Benami and Carter, 2021) .

Despite the putative benefits of index insurance, its adoption in the developing world remains low (Sarris, 2013). Existing literature gives many reasons to explain the scarce take up. While there are multiple barriers to the adoption of any insurance contract (Cole and Xiong, 2017), basis risk stands out as a unique and important barrier to the take up and scale out of index insurance (Miranda and Farrin, 2012; Jensen et al., 2018). Moreover, although there is evidence that index insurance increases agricultural investment, basis risk is responsible for not harnessing the full potential (Karlan et al., 2014).

While basis risk is consequential from an expected utility perspective, it is even more consequential and demand-depressing for individuals who are both ambiguity and risk averse. Elabed and Carter (2015) note that index insurance is essentially a compound lottery: first the individual sees if they have a loss and then they see if the index properly captures that loss. There is ample behavioral evidence that many individuals cannot reduce compound lotteries to the corresponding simple lotteries and unsurprisingly many people are in fact compound lottery or ambiguity averse. Using elicited preference data from cotton farmers in Mali, Elabed and Carter (2015) show that basis risk has a magnified negative effect on the demand for index insurance for individuals who are compound lottery averse compared to individuals who are merely risk averse.

As discussed in the introduction, both components of basis risk (zonal and design) can play a significant role in creating a noisy relationship between the index and individual losses. Despite this observation, the literature has largely focussed on reducing design risk (by creating indexes that better predict losses given a zone structure), rather than studying whether the insurance area

[4] `https://www.indexinsuranceforum.org/`

can be improved by grouping the insured into more homogenous areas (Stigler and Lobell, 2023). There exists a vast literature dealing with which indices correlate best with losses. Bucheli et al. (2021) compare several variables (evapotranspiration, soil moisture and precipitation) and indices that can be formed from them in a winter wheat index insurance program in Eastern Germany. Jensen et al. (2019) provide a comprehensive analysis of different NDVI-based[5] indices for insuring pastoralist agains drought in Kenya. They find out that even if the indexes are highly correlated among them (Correlation > 0.95), there are substantial improvements when using those that become available earlier in the season. However, they confirm that there exists a huge heterogeneity among households and that relationships between the indices and livestock losses vary substantially across index units. Moreover, the index design risk can be also addressed by changing the strike level when indemnities are triggered or by providing flexible products (Jensen et al., 2016; Ceballos and Robles, 2020; Lichtenberg and Iglesias, 2022).

An exception to this focus on reducing design risk is the approach developed by Elabed et al. (2013). They propose a multi-scale index insurance contract, where the first level insurance areas are made small enough to reduce heterogeneity between insureds and make the index closer to farmers. However, if regions are too small, insureds could know each other, exposing insurance providers to moral hazard. In order to circumvent this problem, they introduce a second, larger area with a more forgiving trigger level to essentially audit the lower level for morally hazardous behavior. Under this scheme, indemnities will only be paid if both the indexes for both areas fall below their respective trigger levels.[6]

Other efforts to design index insurance to reduce zonal basis risk includes De Oto et al. (2019) proposal to make payments based on the number of pixels that fall below a certain level, allowing this contract to better account for spatial heterogeneity. Another effort to reduce zonal basis risk is illustrated by Rigo et al. (2022) who study an index insurance for rice in Laos and propose to cluster regions according to elevation and slope. In a similar spirit, Arias et al. (2018) group zones in Ecuador according to soil properties. It is true that soil properties, elevation or slope might be

[5]NDVI stands for Normalized Difference Vegetation Index. This index measures how green a given surface is and it is a great predictor for vegetation health

[6]This approach has the benefit of completely eliminating false positives.

related to the index used for the insurance contract. However, we propose to more straightforwardly address this problem by considering the statistical properties of the index itself rather than searching for proxies that my be related to those properties.

2.2 An Economic Measure of Index Insurance Quality

At the heart of our later analysis will be a time series of length T of observations on a measure of of realized economic well-being or capacity of an individual unit (household or perhaps community). This measure could be income, wealth or productive capital. Each unit is geo-spatially located at coordinates ℓ. The task before us is to define an optimal zone assignment scheme, $z_s(\ell)$ that assigns or maps each location to a particular index insurance zone z_s. The subscript s indexes the particular zone assignment scheme. We denote this measure at time t as as $y_{\ell z_s t}$, where the index z_s represents the insurance zone to which unit ℓ is assigned under scheme s.

Economic well-being or capacity is subject to random shocks and we write the relevant probability density and cumulative distributions functions for these shocks as f_y and F_y, respectively. We further assume that individual expected well-being can be represented by a standard, concave utility function, $v(\bullet)$, and we calculate uninsured expected economic well-being according to the standard postulates of expected utility theory:

$$V^n = \int v(y) f_y dy \approx \frac{1}{T} \sum_{t=1}^{T} \nu(y_{\ell z_s t})$$

where the superscript n indicates that the unit is not insured. To make this object more interpretable, we convert V^n to its certainty equivalent, which we denote as C^n.[7]

We next consider an insurance index, $x_{z_s t}$ that is signals the average economic capacity for individual units assigned to zone z_s. Lower values of $x_{z_s t}$ signal a shock that lowered average capacity in zone z_s.[8] As discussed in the introduction, researchers have devoted substantial efforts to identifying

[7]For example, if the utility function is of the constant relative risk aversion form $v(y) = \frac{y^{(1-\rho)}}{1-\rho}$,where ρ is the coefficient that measures the risk sensitivity

[8]While the analysis in this section studies a series of discrete realization, it is useful for later analysis to denote the probability density and cumulative distributions functions for x as f_x and F_x, respectively. We denote the joint probability distributions between the individual outcome and the index as f_{xy} and F_{xy}.

indexes based on remote sensing or meteorological data that adequately reflect economic outcome for individual units. In our analysis, we take such research for granted and instead focus on the other critical element of index insurance, namely zone formation.

We can define an index insurance contract by its premium π_{x_s} and its indemnity function:

$$I(x_{z_s t}) = S \times g\left(max[0, x_{z_s t} - d]\right) \tag{1}$$

where d is the trigger value or deductible of the index contract defined as $d = F_x^{-1}(u)$, where u is the distributional quantile at which insurance payoffs begin (*e.g.*, payouts are provided in the u% worst states of the world). The indemnity function $g(\bullet)$ maps shocks into financial payments per-unit insured. Finally S is the sum insured under the contract (*e.g.,* 3 cattle or 2 hectares of cotton).

Given a time series of data on $x_{z_s t}$, we can calculate the counterfactual level of expected economic well-being that an individual unit ℓ would have had if the unit had been insured under zone scheme s:

$$V^s = \frac{1}{T} \sum_{t=1}^{T} \nu\left(y_{\ell z_s t} - \pi_{x_s} + I(x_{z_s t})\right)$$

We denote the certainty equivalent of ℓ's expected well-being under zone assignment scheme s as C^s.

Given these pieces, we can define the relative gain in expected utility gain or loss under scheme s as:

$$\Delta^s = C^s - C^n \tag{2}$$

Kenduiywo et al. (2021) decompose Δ^s and discuss how both false positives and negatives affect its value.[9] Note that when $\Delta^s < 0$, the insurance contract lessens economic wellbeing and would never be purchased by an informed, expected utility maximizing individual.

Finally, we follow Kenduiywo et al. (2021) and normalize expression (2) by the certainty equivalent

[9]Kenduiywo et al. (2021) show that Δ^s can be decomposed into the difference in income between the insured and uninsured states marked up by the shadow price of money (or marginal utility of income) when the difference occurs. False negative are particularly damaging to the value of insurance because the insured individual has less money than she would have without insurance and the shadow value of money is high because a loss has just occurred. As mentioned, false positive do not offset this less because the shadow value of money is relatively low when false positive payments occur in good states of the world.

gain, Δ^p, that an individual could obtain from a hypothetical perfect insurance index that exhibits zero basis risk:

$$RIB^s = \frac{\Delta^s}{\Delta^p},$$

where $\Delta^p = V^p - V^n$. This RIB (Relative Insurance Benefit) measure has a convenient interpretation as it captures the fraction of the potential gain from insurance that can be achieved by index insurance contract with zone selection scheme s. In our subsequent analysis, we will use the RIB to gauge the relevant gain of different zone assignment schemes.

3 Lower Tail Dependence and the Design of Index Insurance Zones

As mentioned in the introduction, Stigler and Lobell (2023) use spatial correlation analysis to show that a poorly performing index insurance can result when the *ad hoc* definition of insurance zones based on administrative boundaries results in zones that are highly heterogeneous in the sense that outcomes for one individual within the zone may have a zero or even negative correlation with outcomes for other units in the same insurance zone. In the extreme, such zone heterogeneity could create a world in which the zonal index is quite stable over time even as some units in the zone experience extreme losses at different points of time. Our goal here is to replace the convenient or *ad hoc* definition of zones with a statistical approach that generates zones that will create the highest quality index insurance possible given the ability of the underlying index to accurately predict outcomes within a zone.[10]

As a first step towards this goal, this section proposes a statistical objective that can be used to optimally design insurance zones, specifically by maximizing the lower tail dependence (LTD) of the joint distribution of the zone index and the individual pixel or farm unit assigned to the zone. After defining LTD and establishing its relationship to downside basis risk, this section graphically explores the impact of LTD on basis risk. It then more rigorously shows how zone design based

[10]Zones based on administrative boundaries are convenient in the sense that it is relatively easy to assign the insured to their correct zone. When optimized zones cross administrative boundaries, greater care may be needed to assign individuals to their zone. As index insurance shifts to digital enrollment platforms, this assignment can be done by capturing the geo-spatial location of the client using the sales app.

on LTD impacts the quality of index insurance using the index insurance quality metric defined in Section 2.2.

3.1 Lower Tail Dependence and Basis Risk

The LTD coefficient between two random variables, x and y, is the probability that x is below a threshold value given that y is also below a threshold value. Assuming that these two threshold values are the same, we can formally define the lower tail dependence coefficient between random variables x and y as:

$$\lambda_{x,y}(u) = \lim_{u \to 0^+} P[X < F_x^{-1}(u) | Y < F_y^{-1}(u)] \tag{3}$$

where F_x and F_y are the univariate cumulative distribution functions for the two random variables and u is a threshold quantile value of the cumulative distribution used to define what is a "tail" event.

To put things in an index insurance context, let y measure the random yield of an individual farmer and x the area yield index for the insurance zone to which the farmer is assigned.[11] Note that lower values of both variables signal worse outcomes and larger losses. In insurance terms, u is the risk layer that the insurance contract is intended to cover. If $u = 10\%$, then the insurance is intended to cover the losses that occur in 10% worst states of the world. Lower tail dependence $\lambda_{x,y}(10\%)$ would thus measure the probability that the individuals receives an insurance payout when the individual is in the 10% lower tail of her or his yield distribution. When λ is close to zero, it indicates that the variables exhibit a very weak association between individual and group losses. Conversely, when the LTD coefficient approaches one, it indicates a strong association in the lower tail, suggesting that extreme values of one variable tend to occur simultaneously with extreme values of the other.

Lower tail dependence is commonly used in the analysis of financial returns (Hartmann et al., 2004; Zimmer, 2012; Engle, 2023). For example, imagine that x and y are the random prices of two assets

[11] Alternatively, x could be vegetative growth in a pixel while y is average vegetative growth across all pixels in an insurance zone.

that might be combined into a portfolio. To minimize the chance of bankruptcy, an investor might want to select assets for her portfolio that *minimize* the LTD coefficient so that there is minimal probability that the price of one asset hits its floor value when the other asset is at its floor value (De Luca and Zuccolotto, 2017).

The index insurance zone design problem bears important similarity to this portfolio design problem. An insurance zone can be seen as a portfolio of individual farm or pixel units and the design problem is deciding which units are best included in the "portfolio." But unlike the example of the financial portfolio, a well-constructed insurance zone will *maximize*, not minimize, LTD. To see the logic behind this approach, we follow Clarke (2016) and define r_{YX} as downside basis risk for a pixel or farm unit Y that is grouped into a zone with index X. That is, r_{XY} is the probability that the insured unit experiences losses and the index does not trigger a payout (false negatives). It is straightforward to show that: $r_{XY} = (1 - \lambda_{xy})F_y(\mu)$, where $F_y(\mu)$ is the probability that the individual loss measure falls below the insurance threshold value μ. We turn now to more deeply explore the relationship between LTD and basis risk.

3.2 Empirical Measurement of Lower Tail Dependence

Maximizing LTD to design low basis risk insurance zones of course requires a way to estimate the LTD coefficient, λ, given time series data on the candidate farms or pixel units to be grouped into a zone. To approach this problem, we build on the literature on bivariate copulas. Copulas are especially useful in finance and insurance applications as they present a flexible way to model the dependence between two random variables in the extremes of their distributions. Continuing with the notation from Section 3.1, the theorem by Sklar (1959) shows that the joint distribution between random variables y (*e.g.*, individual farmer yields) and x (*e.g.*, the zonal yield index), $F_{X,Y}(x, y)$, can be written as:

$$F_{XY}(x, y) = C(F_x(x), F_y(y)) \rightarrow C(u_x, u_y) = F_{XY}(F_x^{-1}(x), F_y^{-1}(y)) \tag{4}$$

where C is the copula function that links two random variables that are uniformly distributed over

the $0, 1$ interval. Note that the transformation of the original random variables by their cumulative distribution functions creates new random variables that by construction are uniformly distributed over the $0, 1$ interval. Intuitively, the copula allows flexibility in determining how the tails of the distribution covary. Finally, the right hand side of expression (4) shows that the copula captures precisely the information needed to measure LTD.

While data exhibiting a given LTD can be generated from a known, parametric copula (see Section 3.3 below), we will in general need to estimate the empirical copula given time series outcomes on individual farm outcomes or remote sensing pixels. Following Caillault and Guegan (2005), we can non-parametrically estimate the copula and the coefficient of LTD as:

$$C_T(\frac{t}{T}, \frac{t}{T}) = \frac{\#\{(x, y) \,| x \leq x_{(t)} \, and \, y \leq y_{(t)}\}}{T} \tag{5}$$

$$\hat{\lambda}(\frac{t}{T}) = \frac{C_T(\frac{t}{T}, \frac{t}{T})}{\frac{t}{T}} \tag{6}$$

where *x(t)* and *y(t)* represent the order statistics of observation t in the sample, $\#$ denotes the cardinality of the set and T the total number of observations of the time series. This estimation requires that we determine the quantile values that signal severe losses and trigger insurance payment. In equations (5) and (6), we assume that these quantiles are the same, but the analysis can of course be carried out with different threshold values for the two random variables.

A point to note is that, although the lower tail dependence appears to offer a more reasonable approach than correlation for capturing the dependence between the index and the individual, the LTD coefficient and correlation are not independent. For instance, a correlation of 1 will invariably result in a LTD coefficient of 1 as well. Conversely, if the LTD coefficient is 1, it is not possible to have a correlation equal to 0, since the presence of co-movement in the lower tail would imply a non-zero correlation. That said, aside from extreme values, for a given level of correlation between x and y, it is possible, within limits, to have a higher or lower LTD. In the next subsections, we explore these intermediate cases to help establish the relationship between LTD and basis risk in

index insurance contracts.

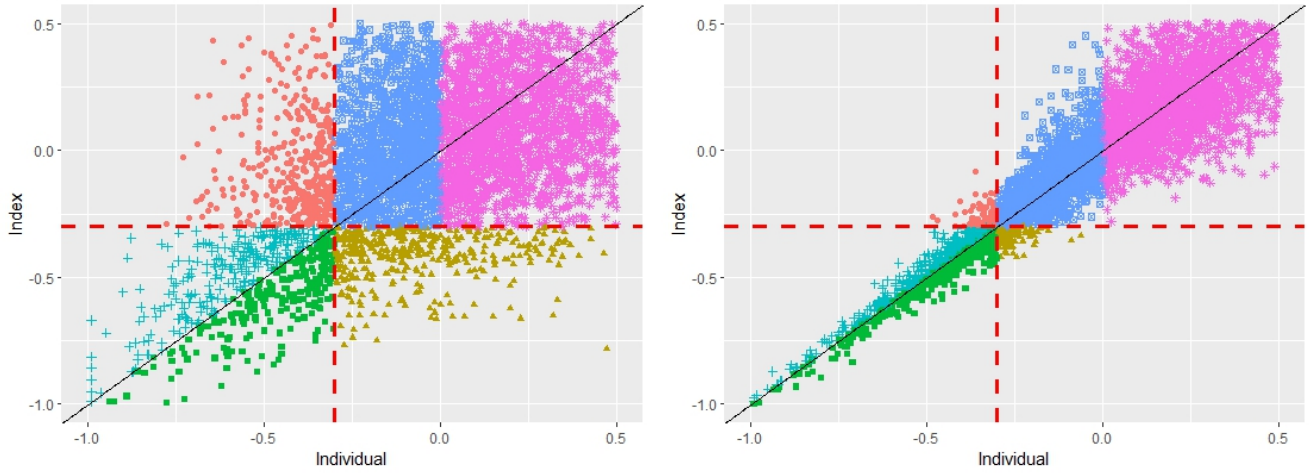
3.3 Exploring the Relationship between Lower Tail Dependence and Basis Risk

To more firmly fix the intuition on the relationship between basis risk, LTD and correlation, this section simulates time series data on a loss index x and an individual loss outcome y, manipulating both the correlation and LTD between the two time series.We begin by assuming that both x and y are marginally distributed as $\mathcal{N}(0, 0.33)$ random variables symmetrically truncated at $[-1, 1]$. We use a threshold lower tail value of -0.3, which is at the 18^{th} percentile of the truncated distribution. We next specify a data generation process that generates bivariate truncated normal variables with a fixed and known correlation coefficient. Using this data generation process, we then randomly generate sets of time series of length 5,000. While each time series will exhibit (approximately) the fixed level of correlation, the generated LTD will vary randomly across the replicated time series. By running a large number of replications, we can pick out for comparison time series with a fixed correlation but with higher and lower levels of LTD.[12]

For purposes of illustration, we consider a high correlation case ($\rho = 0.89$) and an intermediate correlation case ($\rho = 0.49$). For each case, we select two time series that exhibit a higher versus a lower coefficient of LTD (calculated using equation (6) above). To visually illustrate the impact of LTD versus correlation on basis risk, we deploy the graphical framework suggested by Jensen et al. (2016) to study basis risk (see their Figures 1 and 6). The graphs in Figure 1 display the individual outcome on the horizontal axis and the loss index on the vertical. With the critical loss thresholds set at -0.3, a false negative basis risk event occurs whenever $x > -0.3$ and the individual outcome $y < -0.3$. False positives occur whenever $y > -0.3$ and yet $x < -0.3$. The red, dashed horizontal and vertical lines in the graphs mark the insurance loss threshold or trigger values of -0.3. All data points above the 45-degree line are instances where the index under-predicts the losses experienced by the individual. False negatives are in the northwest quadrant defined by the threshold or trigger value lines, whereas false positives are in the southeast quadrant.

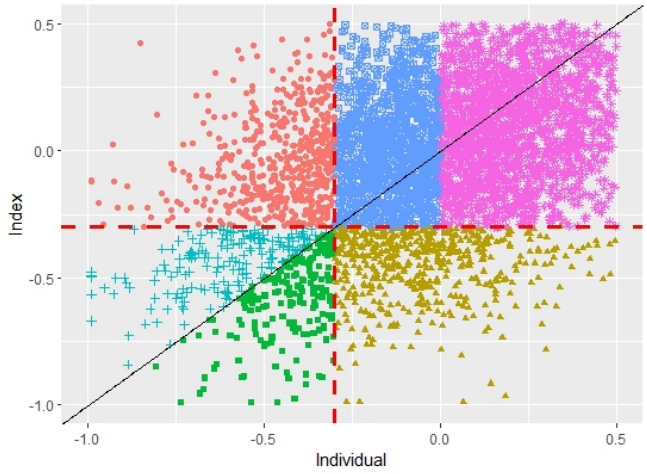
[12]It is also possible to fix the LTD coefficient and let the correlation randomly emerge for any simulated sequence, as we do in the next section.

Figure 1: Basis Risk under Different Levels of Correlation (ρ) and Lower Tail Dependence (λ)

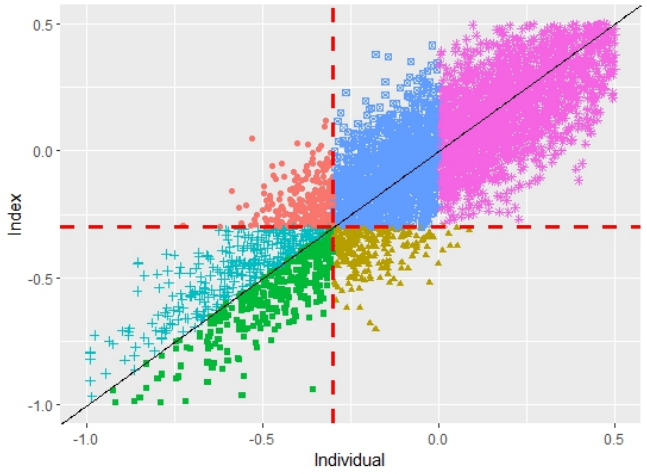


(a) $\rho = 0.49$, $\lambda = 0.50$

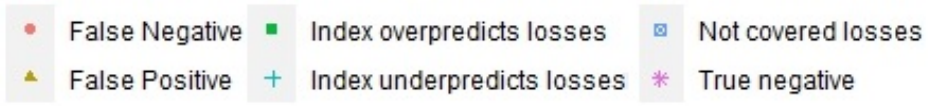
(b) $\rho = 0.89$, $\lambda = 0.90$



(c) $\rho = 0.49$, $\lambda = 0.09$



(d) $\rho = 0.89$, $\lambda = 0.53$



Notes: The dashed red vertical and horizontal lines represent the trigger level. Individual (Index) losses are measured in the horizontal (vertical) axis

The first column of Figure 1 (panels (a) and (c)) illustrate data with correlation held constant (at

$\rho = 0.49$) and with higher versus lower levels of LTD ($\lambda = 0.50$ in (a) and 0.09 in (c)). The second column repeats the same exercise, but with a higher level of correlation ($\rho = 0.89$) but variable levels of λ. As can appreciated visually, conditional on the overall level of correlation, higher λ reduces the number and spread of false negatives in both columns. But, as can be also be visually appreciated, for a fixed ρ, an increase in λ implies a lower correlation in other parts of the distribution. As discussed in Section 2.2, while both false negatives and false positives will reduce index insurance quality, false negatives are especially damaging (see footnote 9). While Figure 1 illustrates the intuition that LTD may be more important than correlation for the design of index insurance, we turn now to more explicitly consider the impact of LTD versus correlation on the economic value of index insurance.

3.4 Lower Tail Dependence and Index Insurance Quality

To more rigorously explore the impact of correlation versus lower tail dependenc on index insurance quality, Figure 2 maps certainty equivalent-based RIB measure of index insurance quality for time series with different degrees of correlation and LTD. Following the procedure outlined in Section 3.3, we first fix the value of correlation between individual losses and an insurance index to be $\rho = 0.01$. For that value of ρ, we randomly generate a bivariate time series of individual losses, $y_{\ell zt}$ and the corresponding index value, x_{zt}, where both random variables vary between 0 (no loss) and -1 (100% loss).[13] To analyze insurance quality, we assume that the individual begins every period with income or wealth level of $k^0 = 1000$ which is then subjected to random shocks. Following the randomly generated loss, individual wealth in time period t is:

$$k_{\ell zt} = k^0(1 + y_{\ell zt}).$$

Based on this time series of uninsured outcomes, we can approximate individual expected utility as shown in section 2.2. We assume a constant relative risk aversion utility function with risk aversion parameter of 1.5.

[13] Each time series is of length 5000.

Following the expression (1), we assume full insurance and write the index insurance linear indemnity function as:

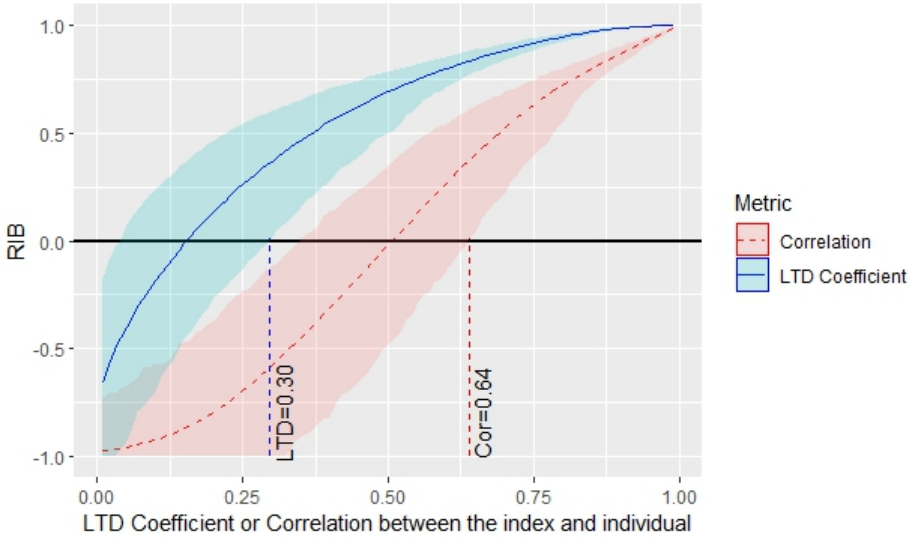
$$I(x_{z_st}) = k^0 \times max[0, -x_{z_st} - d]$$

where we set the deductible at 0.3, meaning the insurance does not cover the first 30% of losses. The insurance premium is simply calculated as the expected of $I(x_{zt})$, marked up by 10%. Again following the procedures laid out in section 2.2, we calculate expected utility when the individual is fully insured and from there use the certainty equivalents of the uninsured and insured expected utility values to calculate the RIB quality measure.

For that given value of ρ, we then replicate the procedure just outlined 999 times, creating a set of 1000 RIB measures, one for each bivariate time series history. Because the time series are randomly generated, this procedure results in a distribution of realized RIB values. We then increase ρ in steps of 0.01 and repeat the RIB analysis for each history and ρ value.

Figure 2 graphs the results of this analysis. The vertical axis in Figure 2 measures the RIB insurance quality indicator. As discussed, $RIB < 0$ implies that the insurance contract makes the insured worse off. The red, dashed line in Figure 2, represents the mean RIB for the corresponding correlation coefficient displayed on the horizontal axis. The shaded area around that mean value encapsulates the 95% of the certainty equivalent values that occurred across the 1,000 histories generated for that value of ρ. As can be seen, it is not until $\rho \geq 0.5$ that we would expect index insurance to make the client better off. Indeed, only at the higher correlation level of $\rho = 0.64$ can we be 95% certain that index insurance will be welfare-improving.

Figure 2: RIB according to different LTD and Correlation between the index and a individual



Notes: Shaded space shows 95% confidence interval.

To analyze the impact of lower tail dependence on insurance quality, we repeat the procedure above, except we generate time series for a fixed coefficient of lower tail dependence, λ, rather than for a fixed the correlation coefficient. To do this, we create an alternative data generation process in which the index x and the individual outcome y have the same marginal distributions as before but the two random variables are linked not by a correlation coefficient but by a parametric copula function with known coefficient of LTD. In particular, we utilize the Clayton copula function that is often used to model joint distributions when LTD is of interest[14]**.** Mimicking the process with a fixed correlation coefficient, we replicate 1,000 time series of length 5,000 for each value of λ between 0 and 1.

Figure 2 also graphs the RIB analysis for different values of λ. The solid, blue line in the figure represents the mean RIB value for the corresponding value of λ. As can be seen in the figure, when $\lambda \geq 0.15$ an index insurance contract would be expected to be welfare improving with the expected $RIB \geq 0$. For values of $\lambda \geq 0.3$, we can be 95% certain that an index insurance contract will be welfare-improving based on the RIB metric.

[14]The Clayton copula is defined with the following equation: $C(u_x, u_y) = (u_x^{-\theta} + u_y^{-\theta} - 1)^{-1/\theta}$ with $\theta \in (0, \infty)$, and its LTD Coefficient is $\lambda_{x,y} = 2^{-1/\theta}$

Stepping back, Figure 2 allows us to see the critical importance of LTD to the design of quality index insurance contracts. For a given level of correlation between individual losses and a loss index, insurance value is always increased if the LTD between those two objects is higher. As can also be discerned, the confidence bands around the LTD-based measures are also thinner. For example for $\rho = 0.5$, the RIB will 95% of the time be between -0.5 and 0.3. At $\lambda = 0.5$,the band runs from only 0.5 to 0.75. Even at the lower value $\lambda = 0.3$, the RIB interval runs only from 0 to 0.6.

While these observations are useful and suggest that lower tail dependence is a more guide to building reliable index insurance, ultimately the behavior of the loss index is endogenous to the individual units that comprise the insurance zone. We turn now to see how this intuition on the importance of LTD can be applied to the design of insurance zones in which multiple individuals must be selectively combined into a single index insurance unit.

4 Generating Optimal Insurance Areas

Section 3 demonstrated that when the lower tail dependence between an individual unit and the index is high, basis risk is low and the quality of index insurance is high. Ultimately, the zone assignment scheme determines the extent of LTD in any zone. This section develops a method to cluster individual units into insurance zones to maximize the LTD between the area index and the individual units assigned to it. We first discuss and compare the use of LTD and correlation as the maximization criteria to create homogeneous insurance zones. Then, we illustrate the basic method using simulated time series data for 1,000 units, which are grouped into 10 insurance zones. For time series characterized by different degrees of correlation and LTD, we assess and compare the RIB when areas are formed to maximize correlation within them versus those formed to maximize LTD. This section closes by developing a zone assignment algorithm that can be applied to actual remote sensing data.

4.1 The Relative Insurance Benefit of Zone Assignment based on Correlation versus Lower Tail Dependence

The analysis to this point has studied how properties of the joint distribution between a single individual (farmer or pixel) and a given index influences the potential quality of index insurance. We now discuss the more realistic situation where the index is not given and we must instead group multiple individual units into a zone, where the average outcome of the individual units assigned to the zone forms the insurance index. The analysis so far suggests that zones should be created in order to maximize LTD amongst the units assigned to the zone. We explore this LTD-based zone assignment scheme and contrast it with an alternative scheme that creates insurance zones in order to maximize correlation between the constituent individual units.

For this exercise, we consider regions comprised of 1000 individual units. The goal of the zone assignment scheme is to group the individual units into 10 insurance zones. We consider 9 different types of regions, distinguished by the *average* correlation in outcomes between the all 1000 units in the region. Region Type 1 has an average correlation of 0.10, Region Type 2 has an average correlation of 0.20, etc. For each region type, we generate 500 histories, where a history is defined by a distinct correlation matrix between each of the 1000 individuals, denoted $P_H^{\bar{\rho}}$, where the superscript $\bar{\rho}$ is the average regional correlation and the subscript $H = [1, 500]$ indexes the history.[15] For each history, 499 years of loss data are randomly generated for each of the 1000 individuals, where the data generation process for each year respects the specific correlation matrix for that history.[16]

For each history, the challenge is to assign each of the 1000 individuals to 10 insurance zones in order to optimize a statistical criterion, either maximizing lower tail dependence within the zone or maximizing linear correlation. For the case of maximizing lower tail dependence to form zones, the zone assignment algorithm proceeds according to the following 4 steps:

1. Form the similarity matrix, $\Lambda_H^{\bar{\rho}}$, where element i, j in the matrix is the realized lower tail dependence between individual i and j for that history, calculated non-parametrically as given

[15] The elements $\rho_{i,j}$ of the correlation matrix are generated following $\rho_{i,j} \backsim \mathcal{N}(\mu, \frac{min(|1-\mu|,|0-\mu|}{3}) \in [-1, 1]$, where μ varies between 0.1 and 0.9 in ranges of 0.1.

[16] As in section 3.3, the marginal distribution of losses is modeled as $y_i \backsim \mathcal{N}(0, 0.33)$ truncated at $[-1, 1]$

in expression (6).

2. Form a dissimilarity matrix, $\Delta_H^{\bar{\rho}}$, where element i, j is defined as $\sqrt{1 - \Lambda_{i,j}}$.[17]

3. Convert the dissimilarity matrix into Euclidean coordinate distance matrix, $D_H^{\bar{\rho}}$ as explained in Appendix A below.[18]

4. Based on $D_H^{\bar{\rho}}$, we perform a k-means clustering algorithm to assign each of the 1000 individuals to one of 10 insurance zones. This procedure creates a single set of zones for each 499-year history. In later analysis based on real NDVI data, we will replace simple k-means clustering with an algorithm that allows imposition of contiguity and zone area constraints.

An identical procedure is followed to define zones based on maximizing correlation, except that the elements of the similarity matrix are based on the realized correlation between each individual over the course of history H.

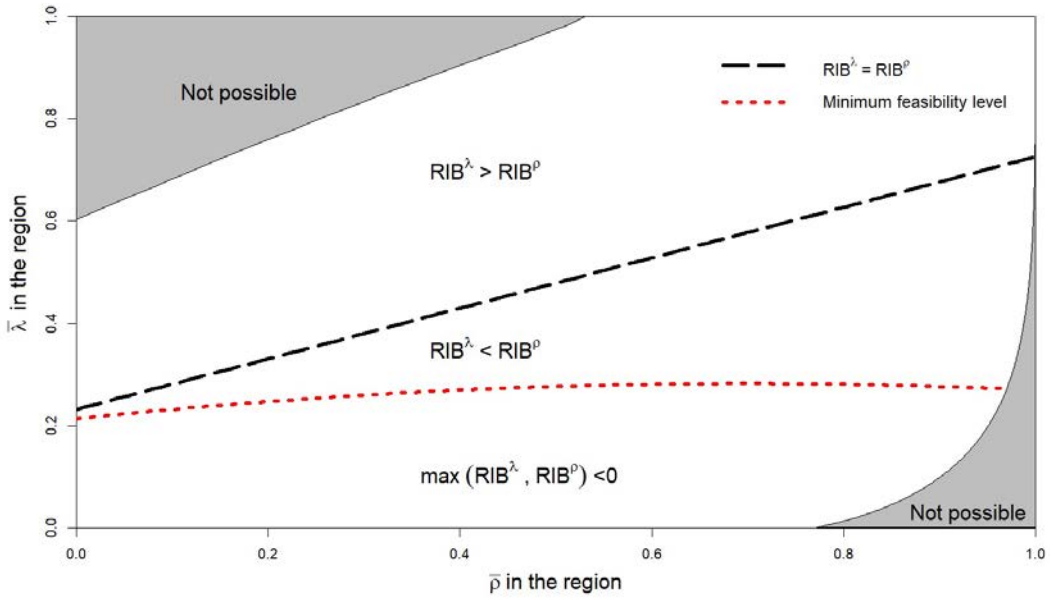
With zones defined, the insurance index for each year is defined as the average losses for the individual units assigned to the zone. As in section 3.3, each individual is assumed to begin each year with 1000 units of income or wealth. Following the procedures outlined in that section a time series of utility values is calculated for each individual (with and without insurance). The RIB for that history and zone assignment scheme is then calculated based on the average certainty equivalent gain across the 1000 individuals. We denote as $RIB^{\bar{\rho}}$ the relative insurance benefit that results from using a zone assignment scheme that maximizes linear correlation.

Figure 3 displays the results of this analysis. The horizontal axis represents the average correlation, $\bar{\rho}$. As mentioned, we generated 500 separate histories for each value of $\bar{\rho}$, where histories are distinguished by the variation in LTD coefficient between individuals as well as by the particular realizations of the random draws of losses for each individual. As mentioned before, each of these histories will randomly result in a realized level of average lower tail dependence between the 1000 individuals, which we denote as $\bar{\lambda}$. Each history can thus be placed in the space defined by average

[17] We move from the similarity matrix to the dissimilarity one because the multidimensional scaling algorithm used in step 4 needs that input.

[18] The procedure outlined here follows the multidimensional scaling algorithm (MDS) developed by Kruskal (1964)

Figure 3: RIB comparison for the LTD- and Correlation-based clusters



Notes: The black and red curves are generated following $RIB = a + b \cdot Cor + c \cdot Cor^2 + d \cdot LTD + f \cdot LTD^2$ for both RIB values obtained from the correlation and LTD-based clusters. In both cases the $R^2 > 0.95$. Minimum feasibility level stands for the curve where $max(RIB^{\lambda}, RIB^{\rho}) = 0$. The gray areas in the bottom right and top left represent the space in which the combination between the LTD and Correlation is not possible (e.g. It is not possible to have a large LTD without having certain amount of correlation).

correlation $\bar{\rho}$ and average lower tail dependence, $\bar{\lambda}$. The shadowed regions in the space demarcate non-feasible combinations of ρ and λ, (*e.g.*, it is not possible to have lower tail dependence higher than 0.7 if correlation is as low as 0.2).

Using the values of RIB^{λ} and RIB^{ρ}, we can partition the space in Figure 3 into distinct zones based on which of the two zone assignment schemes is superior. Several insights emerge from the diagram. First, for any values of $\bar{\lambda}$ and $\bar{\rho}$ below the dotted, red line, index insurance is infeasible as neither zone strategy will yield a positive RIB. Above the dashed, black line, a zone scheme based on maximizing LTD is superior as $RIB^{\lambda} > RIB^{\rho}$. Finally between the dotted and dashed lines, where correlation is relatively high and LTD is relatively low, zones based on linear correlation result in better quality insurance. This finding results because when lower tail dependence is low, false negatives will be unavoidably high. Correlation-based zones at least do a better job at reducing false positives, which also damage index insurance quality as discussed earlier. More generally, if

we were to partition the $\bar{\lambda}, \bar{\rho}$ space with a 45-degree line, we would see that assigning zones based on LTD generates a better index insurance contract for those portions of the space where any index insurance is feasible. In the end, which zone assignment scheme is better is an empirical question, one that Section 5 answers for the case of livestock insurance in Northern Kenya.

4.2 A Zone Assignment Algorithm for High Resolution, High Frequency Remote Sensing Data

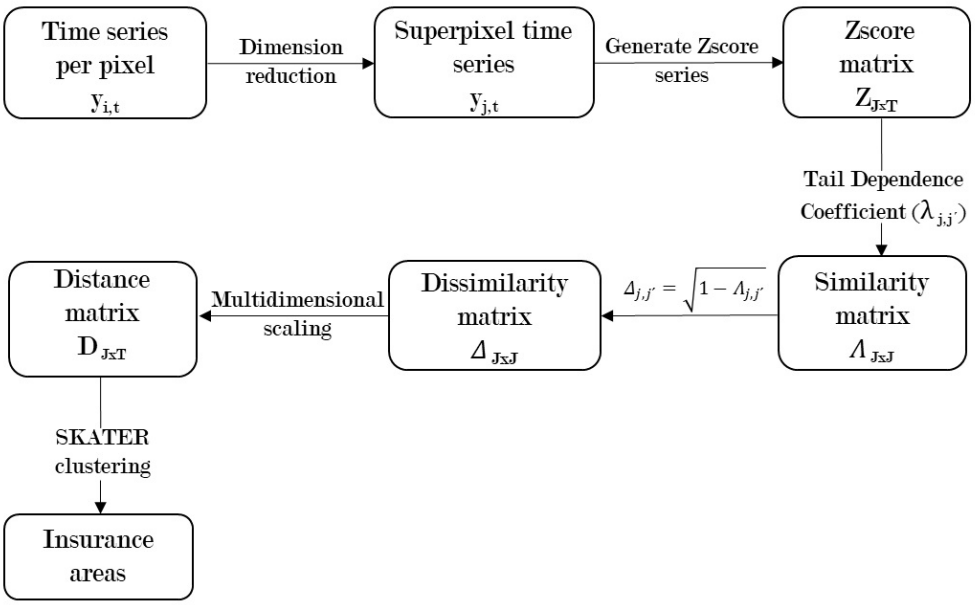
Before turning to the analysis of ground data from Kenya, we need to extend the methodology just developed to account for two additional complexities that confront the creation of index insurance zones in the real world. The first is a big data issue which poses a significant challenge in terms of data management, given that index insurance programs for pastoralist livestock herders cover large regions and resolution of remote sensing data (both spatial and temporal) is getting higher. The second is that the algorithm has to be compact and contiguous insurance zones.

To address the big data challenge, we first reduce the dimensionality of the spatio-temporal database following the "superpixel" segmentation approach described by Estefania-Salazar and Iglesias (2024). As explained by Ren and Malik (2003), superpixels are small patches of pixels that share common attributes. They were first developed for the computer vision domain and restricted to RGB images, but currently they are also being used for remote sensing hyperspectral studies (Subudhi et al., 2021). The algorithm is especially suited to reduce high dimensionality in large spatio-temporal data series with minimal information loss. In essence, this algorithm groups pixels minimizing the "intra-superpixel" time series Euclidean distance. Hence, we move from a pixel level i to a superpixel level j one, reducing the spatial dimensionality of the database. For each superpixel, we take the time series average for all the individual pixels assigned to it. We then convert the superpixel time series into z-scores based on the long-term average for each superpixel.[19]

On the base of these superpixels, we move to define insurance zones that maximize (subject to constraints) lower tail dependence within the zones. Using the non-parametric estimation procedure

[19]Even though most index insurance programs use the accumulated index for the period to be insured, our methodology considers 8-days time series data, considering that different temporal patterns may lead to distinct losses

Figure 4: Methodological approach to define homogeneous areas



in section 3.1, we calculated the pairwise coefficients of lower tail dependence for each set of superpixels.[20] These superpixel LTD coefficients were then assembled into a similarity matrix (Λ), and we then defined the dissimilarity matrix, Δ, and the distances matrix, D, as defineed in section 4.1 In order to then group superpixels into constrained LTD-maximizing insurance zones, we employed the SKATER (Spatial Klustering Analysis by Tree Edge Removal) algorithm (Assuncao et al., 2006), which, in contrast to k-means clustering, allows the imposition of contiguity constraints.

The SKATER algorithm's input is the distance matrix (D), where superpixels experiencing dry spells simultaneously exhibit high LTD and are close to each other. This algorithm consists of three main steps. First, it creates a connectivity graph (tree) indicating whether superpixels are contiguous, assigning a connection value of 1 if contiguous and 0 otherwise. Second, it assigns a weight to each connection based on the Euclidean distance in matrix D, where high weights indicate close proximity. Finally, it removes connections with low weights until the desired number of sub-trees (clusters) is achieved. Figure 4 shows a diagram of the entire process. Further details are given in Appendix

[20]The LTD coefficients were calculated using a threshold value of -0.7, indicating that lower tail dependence was estimated for the left tale tailed defined by the lowest 24% of observations.

Appendix B.

5 The Quality Gain from Optimizing Zones for Index-based Livestock Insurance in Kenya

Index insurance schemes most frequently rely on administrative boundaries to define insurance zones. While this approach is simple and convenient, this paper has so far argued that a statistically rigorous approach based on maximizing LTD between units incorporated into a single zone is likely to result in a higher quality contract that improves the expected economic well-being on the insured. To both illustrate the method and to judge the well-being gains from using it, this section turns to the well-studied Index-based Livestock Insurance contract (IBLI) introduced in Marsabit county in northern Kenya in 2010. Jensen et al. (2024) provide an overview of the evolution of IBLI, including its use as a foundation for livestock insurance programs throughout sub-Saharan Africa, including the government of Kenya's Livestock Insurance Program (KLIP), which was launched in 2015.[21]

Section 5.1 first discusses the current 14 insurance zones used in Marsabit, Kenya, by the on-going livestock insurance program in that region. We then employ the optimization algorithm proposed in Section 4.2 to redraw the zone map in order to maximize LTD within each of 14 new optimized, contiguous insurance zones. Section 5.2 then uses 7 years of data from nearly 1000 households in Marsabit to calculate the gain in insurance quality (using the RIB measure presented in Section 2.2 above) from breaking boundaries and forming statistically optimal insurance zones. For comparison, we compare the gains from relying on maximizing LTD to the gains that could be obtained if zones were formed based on correlation. Finally, section 5.3 explores the magnitude of additional gains that could be achieved by increasing the number of zones beyond 14.

[21]The KLIP program has been replaced by a multi country World Bank program called DRIVE. Within Kenya, DRIVE continues to utilize the administratively-based insurance zones deployed by KLIP.

5.1 Breaking Boundaries

The dark black lines in Figure 5 show the 14 insurance zones currently utilized by the government of Kenya as the basis for its livestock insurance program program.[22] These 14 zones, locally denoted as UAIs (or unit areas of insurance) are based on ward boundaries within Marsabit and we will refer to them as ward-based or administrative UAIs.

In the original IBLI design (see Chantarat et al., 2013), Marsabit was initially divided into 5 zones based on loosely constructed information about agro-ecological zones, species composition of herds and ethnic boundaries. Complaints that the initial zones were too large and heterogenous led to a process of zone expansion. As described in Chelenga et al. (2017), new zones were formed using local administrative units (wards) as the first building block, with community focus groups providing further feedback on the wisdom of combining or not administrative units based on weather, agro-ecological and migratory grazing patterns. There are a total of 20 wards in Marsabit County, and while we were unable to trace the exact origin of the current zones, they follow ward boundaries, with several wards having been combined to create the total of 14 UAIs shown in the figure.

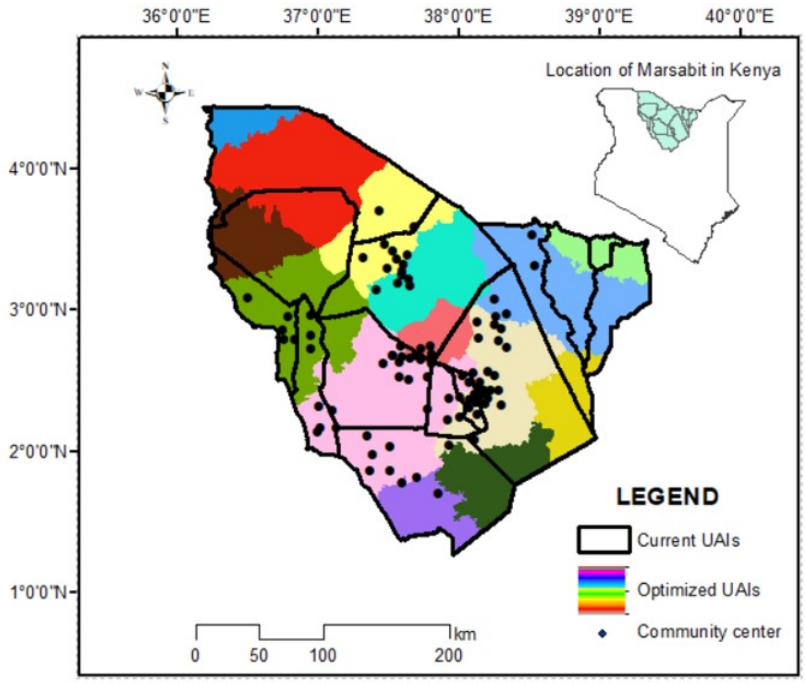
The shaded regions in Figure 5 show the regions that emerged from the LTD maximization algorithm applied to the 2003-2023 NDVI time series (observations once every 8 days) data for the 1,048,592 250x250 m^2 pixels in Marsabit.[23] In the first stage, the 1,048,592 pixels were grouped into 8,971 superpixels. Subject to being contiguous and a constraint that only 14 optimized insurance zones be formed, the optimal zones emerged as shown the different shaded regions on the map. We will refer to these reformulated UAIs as lower tail dependence-maximizing zones, or LTDM UAIs.

Visual inspection of the map of optimized insurance zones reveals that superpixels within most of the administrative units used by the existing livestock insurance have been split and reassigned to 3 or 4 new optimized zones that span ward boundaries. For example, pixels in the southern

[22] In 2022 KLIP was replaced with a larger regional livestock insurance program under the World Bank's DRIVE initiative. DRIVE continues to use these same insurance zones in Marsabit and more generally relies on administrative boundaries to define insurance zones in the other countries in which it operates.

[23] The algorithm also designated a small fraction of the county as non-grazing land as it appeared to be barren, with no detectable vegetative growth. For simplicity purposes, these non-grazing areas have been incorporated into the UAI that contains them.

Figure 5: Current UAIs, Optimized Insurance Areas and distribution of farmers' communities in Marsabit



most administrative unit are reassigned to 4 new optimized zones. The northern part of that is reassigned with a new zone to the north (which spreads across 5 administrative units). Similarly, the southeastern portion of the administrative zone is joined with the superpixels on the western side of the neighboring administratively defined UAIs. These reassignments all reflect the fact that superpixels in the north exhibit exhibit weak LTD with pixels to their south in their ward-based zone. We turn to gauge the importance of these reassignments by comparing the quality of the insurance based on the ward-based UAIs with the proposed downside basis risk-minimizing LTDM zones.

5.2 The Economic Gain from Breaking Boundaries

To gauge the economic gain from breaking boundaries using a rigorous statistical methodology, we take advantage of a long-term research project launched in 2009 to provide evidence on the impact of the IBLI pilot launched in 2010 (see Jensen et al. (2024)) for more detail, including evidence from analyses that relied on the encouragement design incorporated into the study). A total of 903

households from 90 community clusters were included in the original survey, with these households surveyed annually from 2009-2013. Households were then resurveyed in 2015 (and again in 2022). For the analysis here, we rely on survey reports of herd mortality from the 2008-2013 and 2015 surveys. Reflecting the bimodal rainfall distribution of the region, mortality data were collected each year for both seasons (the long rain-long dry season that runs from February-September and the short rain-short dry that runs from October-January). In total, mortality data are available for 7 years and 14 seasons.

Across these 14 seasons, there was substantial variability in terms of mortality. In 2011, the Horn of Africa experienced a severe dry spell. As a result, in the LRLD of 2011, on average, the communities suffered 30.9% of livestock mortality. This contrasts with the lowest mortality rate experienced in the LRLD of 2013 (7.3%). Dealing with intra-seasonal disparity, the average coefficient of variation is 94.0%, a large number which makes the correct clustering of insureds a key element in the design of an index insurance program.

The black dots in Figure 5 indicate the locations of the 90 community clusters. As can be seen, the community clusters do not cover all areas of Marsabit. Four of the ward-based UAIs are not represented in the data, nor are 6 of the new LTDM UAI's. But importantly, for our purposes, the insurance zones are significantly reshaped under the new LTDM UAIs.

For the insurance quality analysis, we treat each of the 90 communities as a single individual and examine the ability of the different livestock insurance structures to properly indemnify average losses in each community. While there is ample evidence that many communities indeed insure against locally insurable idiosyncratic risk (see Townsend, 1994 and the vast literature it inspired), the evidence indicates that such mutual insurance is partial. Our results on insurance quality are thus likely to overstate the expected utility gain any individual (with partial mutual community insurance) would obtain.[24]

To test the impact of these new zones on insurance quality at the community level, we follow the original IBLI design procedure (see Chantarat et al. (2013)) and its subsequent elaboration by

[24]Results obtainable from the author show the value of insurance under the polar opposite assumption that there is no community risk-sharing. Those results show lower values of insurance quality but a similar pattern of increase under the LTDM UAI's.

Kenduiywo et al. (2021). As detailed in Appendix C, for each UAI structure, we estimated mortality response functions in which reported community level mortality data were regressed on cumulative NDVI using segmented regression methods. To prevent overfitting, cross-validation was used and the reported regression coefficients in Appendix C are the average coefficients from the cross-validation exercise.

Using these regression coefficients, we create insurance indexes based on predicted livestock mortality under each UAI scheme. Again following the original IBLI design, payouts begin only when predicted mortality (or equivalently NDVI) crosses a critical threshold. Beyond that threshold, insured communities are indemnified for all marginal animals predicted to have died as shown in equation 1 above. Using this indemnity function and the probability distribution for NDVI in each UAI, we are able to straightforwardly calculate the actuarially fair price of the insurance. After marking this fair price up by 10%, we then follow the procedures outlined in Section 2.2 to calculate the RIB measure of insurance quality.

Table 1 reports the results of this exercise. The columns of the Table 1 present different trigger levels, expressed in terms of z-score levels. A z-score trigger level of -0.7 implies that the contract will trigger 24.2% of the time, or approximately one year in 4. This value is approximately the trigger value currently used in the livestock insurance programs in Kenya. A lower trigger level of -1.0 will trigger 15.9% of the time, while a catastrophic trigger level of -1.5 will trigger only once in about every 20 years.

For each trigger level, Table 1 presents the RIB measure of insurance quality for the current UAI schemes in which insurance zones are defined by administrative units (wards). Focussing first on the -0.7 trigger level, we see that using UAI's based on administrative units yields a community level RIB of 24%, meaning that the index contract yields 24% of the increase in expected utility that a perfect indemnity insurance would generate. Using the LTD coefficient to form statistically optimal insurance zones leads to a 60% jump in insurance quality as the contract now captures 38% of the potential insurance benefit. At a z-score strike point of -1.0 (meaning that the insurance will pay off once every 6-7 years), the LTDM scheme triples the value of index insurance as the RIB rises from 12% to 36%. As the strike point is further lowered (meaning that the insurance only covers the

most catastrophic events), the gain in insurance quality using the LTDM zones versus administrative zones increases monotonically, rising from 100 to over 200%. The absolute insurance benefit relative to perfect insurance does diminish under both zone structures as the strike point falls.[25]

Table 1: RIB under Current versus Optimized UAI's

		z-score Trigger Level (Indemnity Probability)						
UAI Scheme		-0.5 (0.309)	-0.7 (0.242)	-0.9 (0.184)	-1.0 (0.159)	-1.1 (0.136)	-1.3 (0.097)	-1.5 (0.067)
Ward-based	RIB	0.240	0.241	0.188	0.116	0.092	0.047	0.005
LTDM	RIB	0.383	0.384	0.383	0.358	0.287	0.160	0.098
	% Gain	59.6	59.3	103.7	208.6	212	240.4	1860.0
Correlation-based	RIB	0.345	0.344	0.343	0.340	0.16	0.041	0.000
	% Gain	43.8	42.7	82.4	193.1	182.6	-12.8	-100.0

Further insight into these gains from LTDM-based insurance zones can be garnered by examination of the underlying regression coefficients for the mortality regression model reported in Appendix C. For low values of NDVI, the slope coefficient relating NDVI to livestock mortality is steeper under the LTDM zones than under administrative zones (-0.4 versus -0.2). This pattern reflects the fact that under *ad hoc* zones based on administrative units, zone-level NDVI is not very well related to livestock mortality. In the extreme case in which a zone was so heterogenous that average NDVI never predicted losses in the zone, this coefficient would be zero and the index insurance would function like a lottery ticket. Clarke et al. (2012) give a real world example in which a rainfall-based insurance index in India bore almost no relationship whatsoever to actual farmer losses.

While specific to livestock insurance for the northern Kenya rangelands, these results indicate that the economic value of breaking boundaries and forming statistically optimal insurance zones can be enormous. But how much of this gain simply comes from using a statistical criteria to form zones as opposed to specifically relying on our proposed criteria of maximizing LTD? To provide insight into this question, we also reformulated zones under a scheme that maximizes linear correlation within the zone as opposed to LTD. The bottom rows of Table 1 allow us to gauge the performance of index

[25]Results for lower trigger values should be treated cautiously as the limited data available for this analysis has few of these catastrophic realizations.

insurance based on this correlation approach. At the reference strike point of -0.7, the correlation-maximization scheme offers a 44% improvement over administrative districts. While large, this gain is less than the 60% gain offered by the LTDM scheme. Interestingly, subject to the caveat mentioned in footnote 25, we see that for lower strike levels appropriate for catastrophic insurance, the correlation-based zones offer no gain relative to administrative districts.[26] This result is not surprising given that infrequent tail end will have modest impact on the overall level of correlation, whereas they are of course central to the maximization of LTD.

5.3 Are There Further Gains from Increasing the Number of Insurance Zones?

A final potential advantage of defining insurance zones statistically is that the number of zones is not limited to the number of administrative units. As a final exercise, we evaluate the gains from increasing the number of insurance zones in the case of livestock insurance in Kenya. To this end we incrementally increase the number of UAI's from 14 to 104 and calculate the RIB for each increment. Figure 6 presents the RIB that results from increasing the number of UAI's for 4 different trigger levels. Because the individual RIB numbers jump around a bit, we plot smoothed RIB as a function of the number of zones.[27]

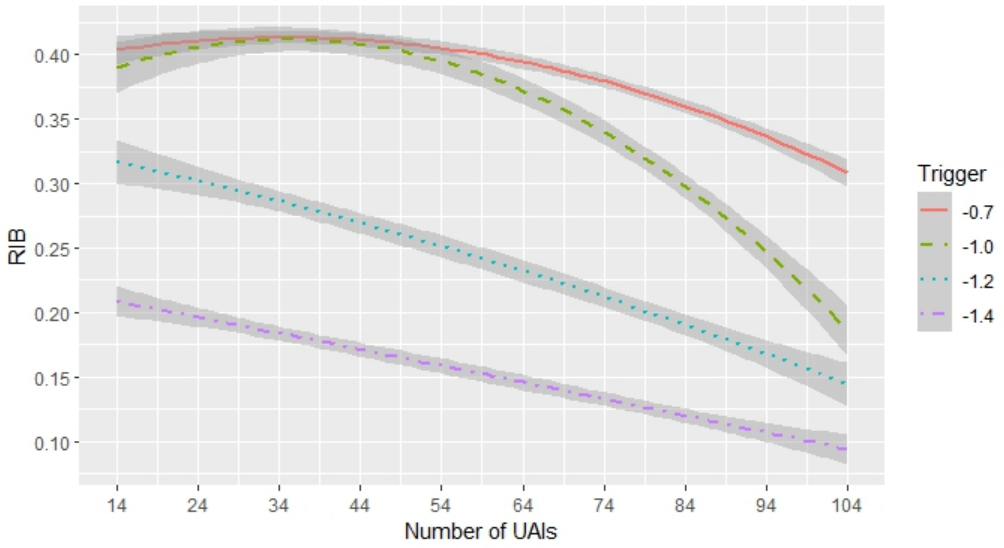
For the higher trigger levels of -0.7 and -1.0 (less catastrophic insurance), RIB modestly increases as the number of insurance triples from 14 to about 40. Beyond that point, RIB declines as the number of zones further increase.

While data limitations (see footnote 27) suggest we should not read too much into this pattern, the results are consistent with the notion that zones can become too small to adequately capture the forage conditions for pastoralist. Unlike sedentary farmers, pastoralists in northern Kenya trek great distances in search of forage for their herds. Insurance based on too small zones may thus

[26] A lower, more catastrophic insurance threshold means that only increasingly rare events trigger payments. Because these rare events will have little aggregate impact on the level of linear correlation, it makes sense that the LTD-based zones perform much better than correlation based zones for more catastrophic insurance coverage.

[27] RIB was smoother using a simple second order polynomial. The lack of smoothness reflects the limited number of observations, a problem that becomes more severe as we increase the number of zones.

Figure 6: RIB obtained varying the number of UAIs



Notes: Lines are smoothed using a second order polynomial $RIB = a + b \cdot UAIs + c \cdot UAIs^2$ ($R^2 > 0.75$). Shaded space shows 95% confidence interval

misrepresent the agro-ecological conditions that are relevant for herd health and mortality. The relationship between number of zones and RIB is likely to be quite different for sedentary farmers. For farmers, more numerous and localized zones would seem likely to increase the value of index insurance, but additional work would be needed to investigate this claim.[28]

6 Conclusions

Despite its promise as a tool to help households manage climate risk and make agricultural systems more resilient, index insurance remains hampered by contracts that may fail to issue payments when losses occur and may also make payments when losses do not occur. This problem of "basis risk" under index insurance contracts has spawned a number of efforts to improve the quality of loss assessment and reduce its cost. While the rapid advance of remote sensing technologies makes these efforts promising, this paper has focusses on the other part of the basis risk problem–the formation of the insurance zones within which all individuals are covered by the same loss index–that is relatively

[28]In the sedentary agricultural setting, decreasing zone size will ultimately create potential moral hazard problems as zones would become small enough that any individual farmer could take actions that would influence the probability of an insurance payoff.

understudied.

In nearly all existing index insurance contracts, insurance zones are formed based on administrative boundaries. While convenient, this *ad hoc* approach may result in zones that are highly heterogenous such that the average losses in the zone (however detected) poorly represent the losses of most individuals. While there have been some calls to reduce the size of zones in order to hopefully reduce this problem, more zones can increase costs and have an uncertain statistical benefit.

This paper has stepped back and asked how best to design index insurance zones in a statistically rigorous way that will improve insurance quality by reducing basis risk. Specifically, the paper shows how to create index insurance zones that maximize the coefficient of lower tail dependence between units included in a zone. Analyzing simulation data from known data generation processes, we show that in most cases maximizing lower tail dependence (which is statistically identical to minimizing downside basis risk) will result in the highest economic value of insurance for a given loss-detection or index technology.

After using the simulation data to develop the machine learning algorithms needed to create lower tail dependence maximizing zones, the paper puts its ideas to the test by analyzing the well-known case of index-based livestock insurance (IBLI) in Marsabit County, Kenya. In addition to its longevity (IBLI began in 2010), IBLI has supported the long-term collection of household data which allows us to assess the impact the insurance would have had under statistically rigorous zones (Jensen et al. (2024) summarizes IBLI and the research that has surrounded it). Using the relative insurance benefit metric proposed by Kenduiywo et al. (2021), we show that replacing the current 14 insurance zones, based on administrative units, with 14 zones designed to maximize lower tail dependence more than doubles the value of IBLI. Under the current zone structure, IBLI only achieves 12% of the expected utility gain that a perfect, but unattainable, zero basis risk contract would obtain. In contrast, replacing the existing *ad hoc* zones with lower tail dependence maximizing zones triples the value of index insurance to 36%.[29] Increasing the number of zones beyond 14 allows for modest

[29]These figures are for the case when the insurance contract pays off 16% of the time. For less catastrophic coverage, which pays off 1 year in 4, insurance quality is 24% using administrative zones and 38% for lower tail dependence maximizing zones, or a 50% quality gain. For more catastrophic policies, the proportional gain for improved zones is higher, but the absolute value of the quality measure is lower.

additional gains. We also show that using a different statistical criterion to define zones (maximizing linear correlation within the units assigned to a zone) also results in gains relative to administrative zones, but smaller than those achieved with the lower tail dependence maximand.

While we have so far only applied our method to one real world case, the striking results from this study, especially when combined with the statistical logic and analysis carried out in this paper with simulated data, suggests that there are large gains to be had from breaking administrative boundaries and defining insurance zones following the rigorous statistical methodology laid out here. The method used here is easily scaled up. Authors of this study have already applied this method to define optimal insurance zones for the entire 13-country, West African pastoralist region. Work with Takaful Insurance Africa in all 5 counties that constitute northern Kenya is already using lower tail dependence maximizing zones as the basis for a new insurance product.

The main drawback from breaking administrative boundaries is that while most people know the name of the country, sub-county or ward in which they live, they will need to learn their zone of residence when zones are defined by statistical criteria. Fortunately, index insurance is increasingly sold with smart phone apps, making it simple to geo-locate the insured and assign them to their correct insurance zone. Future work will also need to explore the use of this methodology for sedentary farmers, although in principal that should be a simpler problem to solve.

Acknowledgements

Enrique Estefania-Salazar was supported by the Programa Propio de I+D+i 2021 of the Universidad Politécnica de Madrid (Grant # PREDOC-21-KB36HA-52-2L7YEQ).

References

Arias, O. V., Garrido, A., Villeta, M., and Tarquis, A. M. (2018). Homogenisation of a soil properties map by principal component analysis to define index agricultural insurance policies. *Geoderma*, 311:149–158.

Assuncao, R. M., Neves, M. C., Camara, G., and Da Costa Freitas, C. (2006). Efficient regionalization techniques for socio-economic geographical units using minimum spanning trees. *International Journal of Geographical Information Science*, 20(7):797–811.

Benami, E. and Carter, M. R. (2021). Can digital technologies reshape rural microfinance? Implications for savings, credit, & insurance. *Applied Economic Perspectives and Policy*, 43(4):1196–1220.

Benami, E., Jin, Z., Carter, M. R., Ghosh, A., Hijmans, R. J., Hobbs, A., Kenduiywo, B., and Lobell, D. B. (2021). Uniting remote sensing, crop modelling and economics for agricultural risk management. *Nature Reviews Earth & Environment*, 2(2):140–159.

Bokusheva, R. (2018). Using copulas for rating weather index insurance contracts. *Journal of Applied Statistics*, 45(13):2328–2356.

Bucheli, J., Dalhaus, T., and Finger, R. (2021). The optimal drought index for designing weather index insurance. *European Review of Agricultural Economics*, 48(3):573–597.

Caillault, C. and Guegan, D. (2005). Empirical estimation of tail dependence using copulas: application to Asian markets. *Quantitative Finance*, 5(5):489–501.

Carter, M. and Chiu, T. (2018). Quality standards for agricultural index insurance: An agenda for action. *State of Microinsurance*, (4).

Carter, M. R., Cheng, L., and Sarris, A. (2016). Where and how index insurance can boost the adoption of improved agricultural technologies. *Journal of Development Economics*, 118:59–71.

Ceballos, F. and Robles, M. (2020). Demand heterogeneity for index-based insurance: The case for flexible products. *Journal of Development Economics*, 146:102515.

Chantarat, S., Mude, A. G., Barrett, C. B., and Carter, M. R. (2013). Designing Index-Based Livestock Insurance for Managing Asset Risk in Northern Kenya. *Journal of Risk and Insurance*, 80(1):205–237.

Chelenga, P., Khalia, D., Fava, F., and Mude, A. G. (2017). Determining insurable units for index-based livestock insurance in northern Kenya and southern Ethiopia. ILRI Resaerch Brief 83, International Livestock Research Institute.

Clarke, D., Mahul, O., Rao, K., and Verma, N. (2012). Weather Based Crop Insurance in India. Policy Research Working Paper 5985, World Bank, Washington D.C.

Clarke, D. J. (2016). A Theory of Rational Demand for Index Insurance. *American Economic Journal: Microeconomics*, 8(1):283–306.

Cole, S. A. and Xiong, W. (2017). Agricultural Insurance and Economic Development. *Annual Review of Economics*, 9(1):235–262.

Conradt, S., Finger, R., and Bokusheva, R. (2015). Tailored to the extremes: Quantile regression for index-based insurance contract design. *Agricultural Economics*, 46(4):537–547.

De Luca, G. and Zuccolotto, P. (2017). Dynamic tail dependence clustering of financial time series. *Statistical Papers*, 58(3):641–657.

De Oto, L., Vrieling, A., Fava, F., and de Bie, K. C. A. J. M. . (2019). Exploring improvements to the design of an operational seasonal forage scarcity index from NDVI time series for livestock insurance in East Africa. *International Journal of Applied Earth Observation and Geoinformation*, 82:101885.

Duque, J. C., Anselin, L., and Rey, S. J. (2012). The Max-P-Regions Problem. *Journal of Regional Science*, 52(3):397–419.

Elabed, G., Bellemare, M. F., Carter, M. R., and Guirkinger, C. (2013). Managing basis risk with multiscale index insurance. *Agricultural Economics*, 44(4-5):419–431.

Elabed, G. and Carter, M. R. (2015). Compound Risk Aversion, Ambiguity and the Willingness to Pay for Microinsurance. *Journal of Economic Behavior and Organization*, 118:150–166.

Engle, R. F. (2023). Compound Tail Risk.

Farrin, K. and Miranda, M. J. (2015). A heterogeneous agent model of credit-linked index insurance and farm technology adoption. *Journal of Development Economics*, 116:199–211.

Hartmann, P., Straetmans, S., and Vries, C. G. d. (2004). Asset Market Linkages in Crisis Periods. *The Review of Economics and Statistics*, 86(1):313–326.

Jensen, N., Stoeffler, Q., Fava, F., Vrieling, A., Atzberger, C., Meroni, M., Mude, A., and Carter, M. (2019). Does the design matter? Comparing satellite-based indices for insuring pastoralists against drought. *Ecological Economics*, 162:59–73.

Jensen, N. D., Barrett, C. B., and Mude, A. G. (2016). Index Insurance Quality and Basis Risk: Evidence from Northern Kenya. *American Journal of Agricultural Economics*, 98(5):1450–1469.

Jensen, N. D., Fava, F., Mude, A. G., Barrett, C. B., Wandera-Gache, B., Vrieling, A., Taye, M., Takahashi, K., Lung, F., Ikegami, M., Ericksen, P., Chelenga, P., Chantarat, S., Carter, M. R., Bashir, H., and Banerjee, R. (2024). *Escaping Poverty Traps and Unlocking Prosperity in the Face of Climate Risk: Lessons from Index-Based Livestock Insurance*. Cambridge University Press.

Jensen, N. D., Mude, A. G., and Barrett, C. B. (2018). How basis risk and spatiotemporal adverse selection influence demand for index insurance: Evidence from northern Kenya. *Food Policy*, 74:172–198.

Jungnickel, D. (2013). Spanning Trees. In Jungnickel, D., editor, *Graphs, Networks and Algorithms*, Algorithms and Computation in Mathematics, pages 103–134. Springer, Berlin, Heidelberg.

Karlan, D., Osei, R., Osei-Akoto, I., and Udry, C. (2014). Agricultural Decisions after Relaxing Credit and Risk Constraints. *The Quarterly Journal of Economics*, 129(2):597–652.

Karra, K., Kontgis, C., Statman-Weil, Z., Mazzariello, J. C., Mathis, M., and Brumby, S. P. (2021).

Global land use / land cover with Sentinel 2 and deep learning. In *2021 IEEE International Geoscience and Remote Sensing Symposium IGARSS*, pages 4704–4707.

Kenduiywo, B. K., Carter, M. R., Ghosh, A., and Hijmans, R. J. (2021). Evaluating the quality of remote sensing products for agricultural index insurance. *PLOS ONE*, 16(10):e0258215.

Kruskal, J. B. (1964). Multidimensional scaling by optimizing goodness of fit to a nonmetric hypothesis. *Psychometrika*, 29(1):1–27.

Lichtenberg, E. and Iglesias, E. (2022). Index insurance and basis risk: A reconsideration. *Journal of Development Economics*, 158:102883.

Miranda, M. J. and Farrin, K. (2012). Index Insurance for Developing Countries. *Applied Economic Perspectives and Policy*, 34(3):391–427.

Ren and Malik (2003). Learning a classification model for segmentation. In *Proceedings Ninth IEEE International Conference on Computer Vision*, pages 10–17 vol.1.

Rigo, R., Santos, P., and Frontuto, V. (2022). Landscape heterogeneity, basis risk and the feasibility of index insurance: An analysis of rice in upland regions of Southeast Asia. *Food Policy*, 108:102237.

Sarris, A. (2013). Weather index insurance for agricultural development: introduction and overview. *Agricultural Economics*, 44(4-5):381–384.

Sklar, M. (1959). Fonctions de rapartition a N dimensions et leurs marges. *Annales de l'ISUP*, VIII(3):229–231.

Stigler, M. and Lobell, D. (2023). Optimal index insurance and basis risk decomposition: an application to Kenya. *American Journal of Agricultural Economics*, n/a(n/a):1–24.

Subudhi, S., Patro, R. N., Biswal, P. K., and Dell'Acqua, F. (2021). A Survey on Superpixel Segmentation as a Preprocessing Step in Hyperspectral Image Analysis. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 14:5015–5035. Conference Name: IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing.

Tang, Y., Yang, Y., Ge, J., and Chen, J. (2019). The impact of weather index insurance on agricultural technology adoption evidence from field economic experiment in China. *China Agricultural Economic Review*, 11(4):622–641.

Townsend, R. M. (1994). Risk and Insurance in Village India. *Econometrica*, 62(3):539–591.

Vrieling, A., Meroni, M., Shee, A., Mude, A. G., Woodard, J., de Bie, C. A. J. M. K., and Rembold, F. (2014). Historical extension of operational NDVI products for livestock insurance in Kenya. *International Journal of Applied Earth Observation and Geoinformation*, 28:238–251.

Zimmer, D. M. (2012). The Role of Copulas in the Housing Crisis. *The Review of Economics and Statistics*, 94(2):607–620.

Appendices For Online Publication

Appendix A Zone Formulation Algorithm

The methodology we propose needs as input a set of time series. In the case of Section 4 these time series correspond to simulated losses and in the case study (Section 5 to the z-score of NDVI values. In this appendix we will refer to what explained in Section 4 to describe the core of the methodology. We will focus on the methodology which maximizes the lower tail dependence.

First, the similarity matrix Λ is formed based on the LTD. Each element i,j of the matrix shows the LTD coefficient between individual i and j. The LTD coefficient is estimated non-parametrically setting the threshold value at -0.3, which corresponds with the trigger level of the fictitious index insurance program. Then, Λ is transformed into a dissimilarity matrix Δ following following $\Delta_{i,j} = \sqrt{1 - \Lambda_{i,j}}$, where $\Delta_{i,j}$ represents an element of the dissimilarity matrix and $\Lambda_{i,j}$the LTD coefficient (similarity) between individuals i and j. This step is required for performing the non-metric multidimensional scaling (MDS) algorithm, which converts a dissimilarity matrix to a Euclidean coordinate matrix with m dimensions. We need to move from a symmetric dissimilarity matrix to a D coordinates matrix, because the clustering algorithm requires that input.

The MDS starts with a randomly generated configuration in a m dimension space (We set m = 499 =Length of the time series). Then, distances between individuals are calculated and these values are regressed against the original distances from Δ to predict distances for each pair of individuals. How good the m dimension space represents the distances is measured by the stress function (See Equation 7), which has to be minimized. The proposed configuration is improved by changing the positions in the direction in which the stress descends more quickly until the results converge.

$$S = \sqrt{\frac{\sum_{i,j}(\Delta_{i,j} - \hat{\Delta}_{i,j})^2}{\sum_{i,j}\Delta_{i,j}^2}} \tag{7}$$

Where:

$\Delta_{i,j}$:Original distance from Δ between individuals i and j

$\hat{\Delta_{i,j}}$:Predicted distance from the regression between individuals i and j

After this step, dissimilar individuals (individuals that do not experience losses together) will be far in the newly generated D coordinates map. Finally, the k-means clustering is performed with D as input. In the real world, where there is a contiguity constraint, we use a spatially-constraint clustering algorithm (See Appendix B)

Appendix B Details of the methodological approach applied to the case study

This appendix explains the methodological approach focusing on the case study. Appendix B.1 describes the source of the remote sensing data and the preprocessing steps needed to obtain the NDVI time series for each grazing land pixel. Appendix B.2 elaborates on the superpixel approach used to reduce the dimensionality of the data that allows to move from a pixel level to a multipixel level. Finally, once the superpixels are generated Appendix B.3 shows how the superpixels are grouped to form the insurance areas based on what described in Section 4 and further explained in Appendix A.

Appendix B.1 Data acquisition and preprocessing

Data is obtained from 2 sources. The NDVI time series is constructed based on the data from the satellite Terra, sensor MOD09Q1.061 (MODIS). The satellite information is downloaded from the software AppEEARS (Application for Extracting and Exploring Analysis Reading Samples). The bands we use to construct the NDVI are bands 1 and 2 which correspond to Red and Near Infrared (NIR) bands respectively. The spatial resolution of this sensor is 250m and observations are made with an 8-day frequency (46 observations per year). The first observation was done in February 4, 2002 but, for the region we are interested in, observations are done regularly after March 30,

2002. Therefore, the time series starts that day and finishes in February 28, 2023, which makes 962 data points in a region of 1,330,510 250x250 m^2 pixels. As we are interested in grazing lands we download the 2021 land cover map generated by the satellite Sentinel-2 (Karra et al., 2021). The spatial resolution for this map is 10m.

First, we calculate the NDVI, as $NDVI = \frac{NIR-RED}{NIR+RED}$. Then, we remove all values out of the range [-1,1] (NDVI can only lie within that range) by treating them like missing values. Next, we fill in the missing values using a ± 3 window exponential weighted moving average. We perform a Savitzky-Golay filtering to smooth the time series by eliminating the noise. There are products from which this information can be directly downloaded and these steps are not necessary. However, we proceed in this way to provide the reader with a global overview of the process.

Dealing with the land cover data, we convert the resolution from 10 m to 250 m. We do that because the resolution for the NDVI maps is 250 m. Since the target land use are grazing lands, we assign the value of grazing land to the lower resolution pixel if more than a half of the 10m pixels that form the 250 m pixel belong to that class. Then, we select (mask) only the pixels that correspond to rangelands, defined as open zones covered by grass and no tall vegetation. It includes savannas and zones covered by bushes, shrubs and tufts of grass. In total 1,048,592 pixels are analyzed. Once the smoothed time series for all the rangeland pixels is generated, we perform a z-score . This way, we obtain a time series not taking into account the absolute NDVI value, but its value relative to the average value of the period.

Appendix B.2 Dimension reduction at pixel level (Superpixel approach)

After the preprocessing, we have 1,048,592 pixels i and for each of them 962 z-score values t. First of all, the high-resolution, high-frequency remote sensing data must be simplified in order to make the data manageable. This step is especially significant in view of higher data frequencies and resolutions in the near future. Its objective is to simplify the satellite image by moving from a pixel level to a multi-pixel level one. To do so, we use superpixel segmentation. Introduced by Ren and Malik (2003), superpixels group similar pixels according to their color or other properties making images

computationally feasible for subsequent algorithms. In this case, the objective is to group pixels that have similar z-score values throughout the 20 years.

The superpixel algorithm starts with a Principal Component Analysis (PCA) in order to reduce the t periods to t' and make the subsequent steps computationally more efficient. The principal components $t(')$ are a linear combination of the t z-score values which maximizes the variance explained, using the Lagrangian method. We use the first 200 components (projection axis of the biggest 200 eigenvalues) in such a way that more than 95% of the variance is explained.

After this step, we continue with the same number of pixels but with t'=200 variables or linear combinations of the original t periods. In order to reduce the spatial dimension, the goal is to use the Max-p spatially constrained clustering algorithm (Duque et al., 2012)[30]. This algorithm groups similar features according to the Euclidean distances of their variables, and its main particularity is that it is not necessary to determine the number of groups desired, but it tries to maximize the number of groups with the condition that each of them contains a minimum number of observations (pixels in this case). However, since it is a highly time-consuming algorithm, we firstly perform a k-means clustering which does not require to fulfill the spatially contiguity condition. We set the number of clusters equal to the number of pixels the region has divided by 80 (13,107 clusters), since we want to create superpixels of approximately 5km^2, in order to have a trade-off between the accuracy of the algorithm and the computing time.

After performing the k-means, pixels belonging to the same cluster and which are contiguous will be treated as a single isolated unit. These multi-pixel islands are then joined using the Max-p algorithm until each clump of islands has at least 80 pixels. The result of this step is the generation of 8,971 superpixels j that will be the base for the subsequent steps. Once the superpixels are generated based on the z-scores at the pixel level, we calculate their counterpart at superpixel level. To do so, we take as the NDVI of the superpixel the average NDVI of all pixels forming the superpixel for each period of the entire time series. Then we calculate the z-score of the superpixel. This way, we generate a z-score matrix Z with 8,971 rows (superpixels) and 962 columns (time series)

[30]The SKATER algorithm could also be used but in this case we should set a number of superpixels, and we want to set only a minimum size for the superpixels

Appendix B.3 Generating Insurance Areas

This subsection shows what discussed in Section 4 and Appendix A applied to the case study. In this case, the input to generate the similarity matrix Λ is the z-score matrix Z. The z-score threshold for the estimating the LTD coefficient is -0.7 (the trigger value for IBLI program). Then, we calculate the dissimilarity Δ and the distance matrix D using the non-metric multidimensional scaling with 962 dimension.

Finally, contrary to Section 4 and Appendix A, we do not use the k-means algorithm but a spatially-constraint clustering algorithm. For this purpose we use the SKATER (Spatial "K"lustering Analysis by Tree Edge Removal) algorithm (Assuncao et al., 2006). This algorithm creates a connectivity graph between elements and assigns to each connection a similarity value based on the Euclidean distance of the D matrix coordinates. In order to make the process more efficient, the graph is transformed into a Minimum Spanning Tree (MST) based on Prim's algorithm (Jungnickel, 2013). This tree is a subset of the original one that contains all the vertices (elements) and has the minimum cost. This way, connections between very different superpixels are removed. Then, the MST is iteratively pruned into k (Number of desired clusters), cutting the edges that minimize the sum of intracluster square deviation. In addition, a minimum cluster size can be set, which is especially valuable to ensure that the insurance areas are large enough to avoid moral hazard. In the case study, we set a minimum size of 380km^2, which is the size of the current smallest UAI.

Appendix C Segmented Regression Approach for Creating Livestock Loss Index

The NDVI of each area is calculated as the average value of the NDVI of all the grazing land pixels within it. As a second step, we aggregate the NDVI of all the periods, so that we have a cumulative NDVI for the LRLD and SRSD season every year. Then, we compute the z-score of each season as done in Kenduiywo et al. (2021). Finally, we fit the following segmented regression model to predict

livestock mortality based on the z-score of the cumulative NDVI:

$$M_{ls} = \beta_0 + \beta_1 z_{ls} + \beta_2 (z_{ls} - \psi) I(z_{ls} > \psi) + \varepsilon_{ls}$$

where M_{ls} is livestock mortality in UAI l in season s; z_{ls} is the z-score of the cumulative NDVI in UAI l in season s; ψ is the breakpoint where the slope of the regression changes; and, $I(z_{ls} > \psi)$ is an indicator function which takes on the value 1 one when the inequality holds. The logic behind this segmented model is that in non-drought conditions, when cumulative NDVI is high, we would expect there to be little relationship between mortality and NDVI. In contrast, when NDVI is low, signaling drought conditions, we would expect a more robust (negative) relationship between NDVI and mortality. Under this parameterization, β_1 is the key factor. For a highly heterogeneous zone, we would expect β_1 to approach zero, whereas in a better formed zone we would expect β_1 to be higher in absolute value. As can be seen in Table A1 for the full sample results, the critical slope coefficient for the LTDM optimized zones is nearly double that of the same coefficient under administrative zones. The R^2 statistic is also higher under the LTDM zones, meaning that these zones are better able to capture mortality losses.

To combat overfitting of the mortality function, we follow Kenduiywo et al. (2021) and employ five-fold cross-validation. Table A1 also presents the average value of the coefficients obtained from the cross-validation exercise. These average coefficients are those that are used in the analysis reported in the main body of the paper.

Table A1: Mortality Segmented Regression Estimates

	Average Cross-validated Estimates			**Full Sample Estimates**		
	Ward-based	*LTDM*	*Correlation*	*Ward-based*	*LTDM*	*Correlation*
ψ	-0.58	-0.91	-0.77	-0.69	-0.90	-0.82
β_0	-0.03	-0.25	-0.12	-0.06	-0.25***	-0.14*
	–	–	–	(0.04)	(0.05)	(0.06)
β_1	-0.22	-0.40	-0.29	-0.24***	-0.40***	-0.31***
	–	–	–	(0.04)	(0.04)	(0.05)
β_2	0.21	0.38	0.27	0.22***	0.38***	0.29***
	–	–	–	(0.04)	(0.04)	(0.05)
R^2	0.23	0.30	0.23	0.23	0.30	0.23