Optimal index insurance and basis risk decomposition: an application to Kenya

Matthieu Stigler, David Lobell

Center on Food Security and the Environment, Stanford University

November 1 2021

Abstract

Index insurance is a promising tool to reduce the risk faced by farmers, but high basis risk, which arises from imperfect correlation between the index and individual farm yields, has limited its adoption to date. Basis risk arises from two fundamental sources: the intrinsic heterogeneity within an insurance zone (*zonal risk*), and the lack of predictive accuracy of the index (*design risk*). Whereas previous work has focused almost exclusively on design risk, a theoretical and empirical understanding of the role of zonal risk is still lacking.

Here we investigate the relative roles of zonal and design risk, using the case of maize yields in Kenya. Our first contribution is to derive a formal decomposition of basis risk, providing a simple upper bound on the insurable basis risk that *any* index can reach within a given zone. Our second contribution is to provide the first large-scale empirical analysis of the extent of zonal versus design risk. To do so, we use satellite estimates of yields at 10m resolution across Kenya, and investigate the effect of using smaller zones versus using different indices. Our results show a strong local heterogeneity in yields, underscoring the challenge of implementing index insurance in smallholder systems, and the potential benefits of low-cost yield measurement approaches that can enable more local definitions of insurance zones.

Keywords: index insurance, agricultural risk management basis risk, risk decomposition, satellite data

JEL codes: G22, O16, Q14

1 Introduction

Risk is ubiquitous in agriculture and can have detrimental effects on producers, particularly in smallholder agriculture where access to formal credit and savings is limited. Index insurance, where payouts are triggered based on an external index, has been advocated as a promising tool to reduce risk compared to traditional individual-based insurance. By de-linking payments from individual loss, index insurance removes moral hazard and reduces

adverse selection, the two main issues that plague individual-based insurance. By relying on measuring an index at an aggregate level instead of measuring each individual field, it furthermore reduces transaction costs.

Despite its theoretical benefits, success of index insurance has been rather limited, with a low take-up for most products (Cole et al., 2013, 2017). A core reason for this low success lies in the concept itself of index insurance: by de-linking payment from individual losses, index insurance introduces *basis risk*, the possibility that a farmer suffers a harvest loss and yet does not receive any payment. As pointed out by Clarke (2016), this eventuality suggests that a highly risk-averse farmer would not want to purchase index insurance. Unfortunately, the basis risk appears to be high in practice. Clarke et al. (2012) report correlations of only 0.14 between payments and yields in India, predicting a probability as high as 1/3 of not receiving any indemnity even in the catastrophic case of yields falling to zero. Jensen et al. (2016) likewise find substantial basis risk in a livestock insurance in Kenya.

The presence of basis risk has triggered a large literature, extending now well beyond the economic realm, on a quest for the "perfect" index. Indices considered range from input-based indices such as precipitation, temperatures, soil moisture (Vroege et al., 2021) or satellite-derived vegetation indices (Chantarat et al., 2013) to output-based indices using average yields, either from official statistics (Skees et al., 1997), sampling surveys (World Bank, 2015) or from satellite estimates (Flatnes et al., 2018); see Benami et al. (2021) for a survey. With the proliferation of these studies, there is some confusion with the notion of basis risk. In some studies, finding the perfect index is sometimes conflated with the task of finding the best predictor of individual or aggregate yields. Implicitly assumed is the idea that if one were able to perfectly predict yields, one would obtain a perfect index without any basis risk. This neglects, however, the important fact that index insurance is still using a regional-based index. Unless yields are perfectly correlated, any regional index will be an imperfect predictor of individual farm-level yields and hence will introduce basis risk. However, there is no clear understanding in the literature on the extent of this *irreducible* basis risk that simply arises from using a regional index. As a result, it is currently impossible to distinguish index quality from zone quality: a finding of high total basis risk can correspond to both a "good index within a bad zone" or to a "bad index within a good zone".

In this paper, we seek to refocus the debate on basis risk by examining the role of irreducible zonal risk. To do so, we formalize the decomposition of basis risk advocated by Elabed et al. (2013). According to these authors, basis risk arises from two distinct sources, 1) the *zonal* risk[1] due to the use of a regional index to predict individual yields, and 2) the *design* risk due to lack of accuracy of the regional index at summarizing the aggregate covariant movement in yields. Intuitively, zonal risk is intimately linked to the heterogeneity within a zone. The lower the correlation between fields, the higher the zonal risk. Conversely, the design risk stems from the use of a sub-optimal index that does not summarize well the covariant risk. Figure 1, inspired by Benami et al. (2021), illustrates the decomposition of

---

[1]Elabed et al. (2013) originally used the term of *idiosyncratic* risk to denote what we call here *zonal* risk. We prefer the term *zonal* as we feel that it conveys better the fact that this risk only stems from using the same index for a zone. Using a zone with only one field would effectively bring this risk to zero.

Total Basis Risk		
Covariant risk (potentially insurable)		Irreducible basis risk: Zonal risk
Covariant risk (actually insured)	Total basis risk: Design risk	Zonal risk

Figure 1: Decomposition of basis risk. Source: adapted from Benami et al. (2021).

risk. While a valuable and intuitive decomposition, Elabed et al. (2013) did not specify what the optimal index is. Specifying the optimal index is, however, a necessary condition for the decomposition to be valid: absent this specification, the distinction between zonal and design risk becomes arbitrary. Some authors assume that the optimal index is the zone-average yield (Flatnes et al., 2018) or an abstract weather-based index (Conradt et al., 2015; Yu et al., 2019). As we will show, this is not correct and can lead to nonsensical situations where an "imperfect" index performs better than a supposedly "perfect" one.

Our first contribution is to derive the optimal index according to two correlation-based metrics of basis risk, both linked to the R_i^2 from the regression between each individual field i and the index. Our first zone-specific metric $\overline{R^2}$ is the average of the R_i^2, while the second $\overline{\overline{R^2}}$ corresponds to a weighted-average of the same R_i^2. We derive the optimal index for each metric, which provides a formal definition of zonal risk, giving an upper bound on the $\overline{R^2}$ *any* index could reach for a given zone. Beyond this value, improvements in the $\overline{R^2}$ can only be achieved by re-designing the zone itself. Equipped with this upper bound, a practitioner can now distinguish index quality from zone quality and set priorities accordingly: if the upper bound is high, the zone is relatively homogeneous and it is worth embarking on the quest for a better index. On the other hand, if that maximum $\overline{R^2}$ turns out to be low, the quest for the best index is rather futile, and efforts should be rather spent on finding more homogeneous (and likely smaller) zones.

To compare our optimality theory from previous results in the literature, it is useful to emphasize the conditions under which our optimality result is obtained. We define here as index insurance a scheme that uses the same index and indemnity function over all fields, excluding insurance schemes with individual-specific indices or indemnities. Most of the optimality results in the literature take the opposite route, seeking to derive the optimal indemnity function for a given index. Mahul (1999) derives the optimal field-specific indemnity function assuming that the index is the zone-average yield, and Zhang et al. (2019) extend it to a weather-based index. Bourgeon and Chambers (2003) extend Mahul (1999)'s results to include asymmetric information. Mahul and Wright (2003) consider the case of revenue insurance, and Mahul (2001) of a weather-based index. Unlike these studies, we focus on

the optimal index regardless of the indemnity function, instead of the optimal indemnity for a given index. It is likely nonetheless that most of the results from this strand of literature hold with the optimal index we derive.

A second way in which our approach differs from the literature is in the choice of the loss function and the focus on basis risk measures. Broadly speaking, two types of measures are used for index insurance: 1) basis risk metrics, which capture the relationship between individual yields and the index, and 2) insurance benefits metrics, which evaluate the expected utility of yields with and without insurance. We focus here on a simpler basis risk measure based on the linear correlation between yields and the index. This is made mainly for convenience, as considering instead the correlation between "post-indemnity" yields and the index, or considering an expected utility loss function, makes the problem highly non-linear and hence hard to solve analytically. On the other hand, our metric has the advantage of being free of assumptions on the specific forms of the utility and indemnity functions. Furthermore, as shown in Stigler and Lobell (2021), the simple metric based on the average R^2 between yields and the index is a good predictor of the expected utility of index insurance for maize and soybeans in the United States. We observe a similar result in our robustness check section, finding a high correlation between our linear measure of basis risk and more sophisticated metrics.

Putting theory into practice, our second contribution is to showcase how satellite data can be used to measure the extent of zonal and design risk. We focus on Kenya, where the government has been investigating options for providing large-scale insurance to farmers, and where the World Bank (2015) recommends an index insurance scheme based on zone-average yields (also called area-yield insurance). Yet, the question of which and how many zones to use for the insurance remains open. Furthermore, at the time of the report, the only available data were district-level averages, which are not informative about the critical statistic for index insurance, within-district heterogeneity. To overcome these limitations, we use a dataset of yield estimates at a very fine (10m) resolution based on satellite data spanning the entire area cropped with maize in Kenya over 2016-2019 (Lobell et al., 2015; Burke and Lobell, 2017; Jin et al., 2017b, 2019). Using such a high-resolution dataset of yields allows us to go beyond the previous literature that used satellite data to build indices but relied on ground-collected data to measure basis risk and insurance efficacy (see Chantarat et al., 2013; Jensen et al., 2016, 2019). Unlike these studies, we use the satellite data to measure basis risk, providing a detailed picture of within-district heterogeneity over many districts spread over multiple agro-ecological zones. Facing no restriction due to a ground-collected sample, we can further investigate the effect of varying the size of the insurance zone using various administrative boundaries, providing guidance for the implementation of the government's large-scale program.

The paper is organized as follows: Section 2 develops the theory of the optimal index, Section 3 describes the data used, Section 4 the main results, Section 5 robustness checks, while Section 6 concludes.

2 Theoretical results

In this section, we give a formal definition of the basis risk decomposition, define a measure of basis risk and derive the optimal index relative to the selected basis risk metric. Following the intuition in Elabed et al. (2013), we construct the basis risk decomposition as:

Definition 1. *Let $BR(f)$ be a measure of basis risk associated with index f, with higher values $BR(f') > BR(f)$ indicating that index f' has more basis risk than index f. Define as $f^\star$ the optimal index that minimizes $BR(\cdot)$, i.e. $f^\star \equiv \arg\min_f BR(f)$. Then we define:*

- *Total basis risk: $BR(f)$*

- *Zonal risk: $BR(f^\star)$.*

- *Design risk: $BR(f) - BR(f^\star)$*

The *zonal* risk $BR(f^\star)$ corresponds to the minimum irreducible basis risk attainable within a zone. It arises due to the fact that we are using a regional index to predict individual yields. The *design* risk is then defined as the difference between the optimal and actual index $BR(f) - BR(f^\star)$ and arises due to the use of a sub-optimal index. Together, these two components form the *total* basis risk $BR(f)$, which is the value measurable by a practitioner when evaluating the basis risk of a given index.

To be valid, the basis risk decomposition requires that $BR(f) \geq BR(f^\star) \; \forall f$. If this is not the case, design risk can become negative, leading to a nonsensical situation where design risk reduces total basis risk. This implies that the decomposition cannot hold for individual-level basis risk $BR_i(\cdot)$: for a given index f, one can always generate a negative design risk on a single field by trivially setting the index to that field's own yield, $f' = y_i$. In other words, there is no index $f^\star$ that satisfies $BR_i(f) \geq BR_i(f^\star)$ simultaneously $\forall f$ and $\forall i$. This makes it clear that the basis risk decomposition and the concept of design risk can only hold at the zone level, for a zone-specific basis risk measure $BR = f(BR_i)$.

We now introduce some notation to define our zone-wide basis risk metric BR. Let us denote by y_{it} the yields for field $i = 1 \ldots N$ at time $t = 1 \ldots T$, and by f_t the index. The index can be any input- or output-based measure, with the index by definition taking a constant value across space within the same zone, i.e $f_{it} = f_t \; \forall i$. We distinguish two types of indices, the output-based ones using a weighted average of observed yields, $f_t^Y \equiv \sum_{i=1}^{N} w_i y_{it}$ and the input-based ones using a variable x, $f_t^X \equiv \sum_{i=1}^{N} w_i x_{it}$. Typical output-based indices are the zone-average yields, $f_t^Y = \bar{y}_t$, and typical input-based indices rely on precipitation or other weather variables. In matrix notation, denoting by Y the $T \times N$ matrix of yields, an output-based index can be written as $f^Y = Yw$, where w is a $N \times 1$ vector of weights.

Miranda (1991)'s seminal paper on the benefits of index insurance assumed an index based on zone-average yields:

$$y_{it} = \alpha_i + \beta_i \bar{y}_t + \epsilon_{it} \tag{1}$$

Interpreting this model in a general way, we assume that yields follow an arbitrary distribution with covariance matrix Σ of dimension $N \times N$. A diagonal element of Σ represents

field's i variance, σ_i^2, while a typical off-diagonal element σ_{ij} represents the covariance between field i and j. Under this interpretation, β_i is the least-square projection of y_{it} on $\bar{y}_t$ from (1). Denoting by $\mathbf{1}$ a vector of ones of dimension $N \times 1$, the vector of β from regression (1) using the zone average $f = \bar{y}$ as index can be written in compact form as:[2]

$$\beta(\bar{y}) = N \frac{\Sigma \mathbf{1}}{\mathbf{1}'\Sigma\mathbf{1}} \tag{2}$$

As shown by Miranda (1991), the β_i parameter plays an important role to assess the variance-reduction effects of index insurance for field i. A drawback of the β_i parameter is that it depends on the field's own variance and hence is unbounded.[3] This makes it challenging to build a basis risk metric based on β_i, and complicates further a comparison between fields or between insurance zones. Following Elabed et al. (2013), we define rather basis risk as $BR_i(f) = 1 - R_i^2(y_i, f)$, the proportion of variance of field i not explained by the index. Conversely, R_i^2 represents the insurable *covariant risk*, i.e. the proportion of risk explained and insurable by the index. The R_i^2 is bounded between 0 and 1 and thus can be compared across fields. Denoting by $\mathcal{R}^2$ a $N \times N$ matrix whose diagonal elements are the individual R_i^2, the R_i^2 using the zone average as index, $f = \bar{y}$, can be written compactly as:[4]

$$\mathcal{R}^2(\bar{y}) = D^{-1/2}\Sigma\mathbf{1}\left(\mathbf{1}'\Sigma\mathbf{1}\right)^{-1}\mathbf{1}'\Sigma D^{-1/2} \tag{3}$$

with D a diagonal matrix containing the σ_i^2 from Σ. While the above expressions are for the population β and $\mathcal{R}^2$, their sample counterparts $\hat{\beta}$ and $\hat{\mathcal{R}}^2$ can simply be obtained by replacing Σ with its sample counterpart, S. This provides a compact way to obtain the $\hat{\mathcal{R}}^2$, and is numerically equivalent to obtaining the individual $\hat{\beta}_i$ and $\hat{R}_i^2$ by running N regressions separately.

As noted above, the basis risk decomposition holds at the zone level, not the individual level. Hence, we need to choose an aggregation function for the zone-specific basis risk BR. We take a simple average as the aggregation function, $BR = 1/N \sum BR_i = 1/N \sum(1 - R_i^2)$, and define the average covariant risk $\overline{R^2}$ as the average R_i^2 over the fields within the same insurance zone:

$$\overline{R^2(\bar{y})} \equiv \frac{1}{N}\sum_i R_i^2 = \text{Trace}\left(\mathcal{R}^2(\bar{y})\right)/N \tag{4}$$

We can now ask which is the index f maximizing the covariant risk $\overline{R^2(f)}$. Several studies have used the basis risk decomposition, assuming either that the optimal index was an input-based one (Conradt et al., 2015; Yu et al., 2019) or an output-based one such as the zone-average yield (Flatnes et al., 2018). Intuitively, it is clear that an output-based index

---

[2]To gain some intuition, this comes from the fact that $cov(y_i, \bar{y}) = cov(y_i, 1/N \sum y_j) = 1/N \cdot (var(y_i) + \sum_{j\neq i} cov(y_i, y_j))$, which in vector notation is equivalent to the row-sum of Σ, i.e. $\Sigma\mathbf{1}/N$. Furthermore, the variance of $\bar{y}$ is given by: $\mathbf{1}'\Sigma\mathbf{1}/N^2$. Taken together, we see that $cov(y_i, \bar{y})/var(\bar{y}) = N\Sigma\mathbf{1}/\mathbf{1}'\Sigma\mathbf{1}$

[3]β_i can be written as $\beta_i = \rho\sigma_i/\sigma_f$, where σ_f denotes the variance of the index, and ρ is the correlation coefficient between field's i yields and the index.

[4]The off-diagonal elements of $\mathcal{R}^2$ have no direct interpretation and are not used in our analysis.
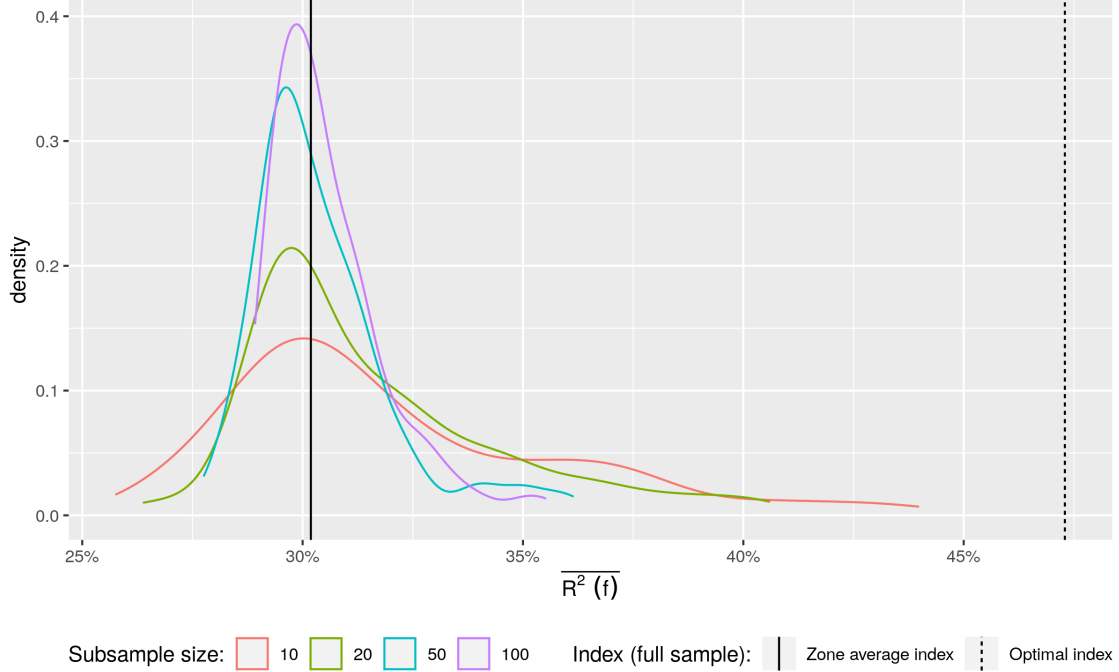
Figure 2: Comparing $\overline{R^2(f)}$ using indices based on full and subsampled averages. The solid and dashed lines represent the $\overline{R^2(f)}$ evaluated a respectively the zone-average and the optimal index. The density represent the $\overline{R^2(f)}$ using subsampled means as index, based on 200 replications for each subsample size.

should lead to a lower basis risk compared to an input-based index, given that the output-based index relies on a richer set of information encompassing both weather components and yield shocks. Does this mean that the zone-average index $f = \bar{y}$ is the optimal index, that is the index leading to the highest $\overline{R^2(f)}$?

To verify whether the zone average is indeed the optimal index, Figure 2 shows a small experiment based on the yield data introduced below. In a nutshell, we consider an insurance zone with 200 fields, and measure the $\overline{R^2(f)}$ for various indices. We consider as index the zone-average $f = \bar{y}$, as well as "improper" averages $f = \tilde{y}$ based on random subsamples of size 10, 20, 50 and 100. We draw multiple random samples for each subsample size and show their density in Figure 2. Contrary to intuition, the $\overline{R^2(f)}$ based on the improper averages $f = \tilde{y}$ are often found to be higher than the one based on the full-sample average, $\overline{R^2(\bar{y})}$. About half of the indices based on averaging just 10 observations do better than the full-sample mean. This is an important concern for the validity of the basis risk decomposition: assuming that the zone average $\bar{y}$ is the optimal index can lead to an ill-posed negative design risk!

What is then the index maximizing $\overline{R^2(f)}$ if it is not the zone-average $\bar{y}$? To find this

7

out, let us consider more general output-based indices $f = Yw$ using any weight w_i for field i, instead of focusing on the zone-average $\bar{y}$ restricting the weights to $w_i = 1/N$. Looking at vectors with different weights is motivated by the experiment above, which showed that using a mean based on a subsample (i.e., using a weight vector where some entries are set to 0) can perform better than the full-sample mean. $\mathcal{R}^2(f{=}Yw)$, the matrix containing the R^2 based now on the generalized index $f = Yw$, can be written as:

$$\mathcal{R}^2(f{=}Yw) = D^{-1/2}\Sigma w \left(w'\Sigma w\right)^{-1} w'\Sigma D^{-1/2} \tag{5}$$

Our statistic of interest is now $\overline{R^2(f{=}Yw)}$, and we search for the optimal $w^\star$ maximizing $\overline{R^2(f{=}Yw)}$, i.e. minimizing the average basis risk $1 - \overline{R^2(f{=}Yw)}$. Proposition (1) derives this optimal $w^\star$:

Proposition 1. Let $w^* \equiv \arg\max_w \text{Trace}\left(\mathcal{R}^2(f{=}Yw)\right)/N$ be the optimal vector of weights maximizing $\overline{R^2(f{=}Yw)}$, $f^* \equiv Yw^\star$ the corresponding optimal index and C the correlation matrix $C \equiv D^{-1/2}\Sigma D^{-1/2}$. Then:

1.1 $w^\star = D^{-1/2}eigenvector(C)$. That is, the optimal weight vector is the first eigenvector of the correlation matrix C rescaled by a function of D, the diagonal matrix of variances from Σ.

1.2 $f^\star = YD^{-1/2}eigenvector(C)$.

1.3 $\overline{R^2(f^\star)} = \lambda_1^C/\sum_i \lambda_i^C$, where λ_i^C is the ith largest eigenvalue of C. That is, the objective function $\overline{R^2(f)}$ evaluated at the optimal predictor $f^\star$ is the (relative share) of the largest eigenvalue of the correlation matrix between fields .

Proof: see proposition A.1 in the appendix.

Proposition (1) derives the optimal index $f^\star$ and quantifies the extent of the irreducible *zonal* risk. We now have a well-defined decomposition of basis risk, where the zonal risk is defined as $1 - \overline{R^2(f^\star)}$, i.e. the lower bound on the $1 - \overline{R^2(f)}$ that any index will ever be able to reach for a given zone. Beyond that point, there is no possible improvement in the $\overline{R^2(f)}$ (or reduction in $1 - \overline{R^2(f)}$), unless one starts to re-define the zone itself. Accordingly, design risk for any index f can now be defined as $\overline{R^2(f^\star)} - \overline{R^2(f)}$ and is now guaranteed to be positive. Furthermore, (1.3) shows that the value of the irreducible zonal risk is easily computed, given the relationship $\overline{R^2(f^*)} = \lambda_1^C/\sum_i \lambda_i^C$: all that is needed is to build the correlation matrix between fields and compute the share of its first eigenvalue. [5] To illustrate this, the optimal value $\overline{R^2(f^\star)}$ is shown as a dashed black line in the previous example comparing the $\overline{R^2(f)}$ using the full- or sub-samples average, see Figure 2. It is

---

[5]Computation of the share of the first eigenvalue can furthermore be done very quickly, noting that XX' and $X'X$ share the same first $k = \min(N, T)$ eigenvalues and the same trace. This means that computing the first eigenvalue of a very large dataset containing thousands or even millions of fields can be done in just milliseconds. We provide R routines to compute this efficiently at XXX (`removed to preserve the anonymity of the reviewing process`).

clearly visible that while many indices have a higher $\overline{R^2}$ than the zone-average $\overline{R^2(\bar{y})}$, they all lie below the optimal index $\overline{R^2(f^\star)}$.

Our result can be extended to an alternative objective function for basis risk, following the argument in Meredith and Millsap (1985) and Jong and Kotz (1999). Let us assume now that one wants to minimize the total sum of squared residuals $TSSR(f) \equiv \sum_i SSR_i(f)$, instead of maximizing the average $\overline{R^2}$. It is easily seen that minimizing the total sum of squared residuals is equivalent to maximizing what we call the *total* R^2, $\overline{\overline{R^2}} \equiv 1 - \sum_i SSR_i / \sum SST_i$, where SST_i is the sum of squares total for field i.[6] One can show (see Jong and Kotz, 1999):

Proposition 2. *Let* $w^{\star\star} \equiv \arg\min_w \sum_i SSR_i(w) = \arg\max_w \overline{\overline{R^2(w)}}$. *Then:*

2.1 $w^{\star\star} = eigenvector(\Sigma)$. *The optimal weight vector is the first eigenvector of the covariance matrix* Σ.

2.2 $f^{\star\star} = Y \cdot eigenvector(\Sigma)$. *The optimal index is the first principal component of* Σ.

2.3 $\overline{\overline{R^2(f^{\star\star})}} = \lambda_1^\Sigma / \sum_i \lambda_i^\Sigma$

Proposition (2) leads to very similar conclusions to Proposition (1). In particular, (2.3) shows that the covariant risk as defined by $\overline{\overline{R^2}}$ can never be higher than the (share of the) first eigenvalue of the covariance matrix Σ. The total $\overline{\overline{R^2}}$ is closely related to the average $\overline{R^2}$, noting that the former is a variance-weighted average of the individual R_i^2, while the latter uses equal weights.[7] In practice, we found that the two criteria were giving very similar answers, and focus henceforth on the $\overline{R^2}$.

Theorems (1) and (2) make two similar contributions: (1) they each derive the optimal index according to a specific criterion, and (2) they each indicate how to measure zonal risk as a simple function of eigenvalues. It is important to note that we view the second contribution as the most important, and do not necessarily advocate using the optimal index in practice. Indeed, attributing distinct weights to every field defeats the purpose itself of index insurance, which is to provide a simple and cost-effective way to cover risk, without the need to measure yields on every field. Furthermore, the weights of the optimal indices do not necessarily sum to one and can even be negative. This implies that some fields might be disadvantaged when the optimal index is used. On the other hand, we view the second result showing how to compute the zonal risk as the main contribution of theorems (1) and (2). Even if one does not intend to use the optimal index, having a clear and computationally simple indication of how the best index would perform can be very informative, indicating how far the elusive quest of the best index can go. This sets the basis to shift the focus from index quality to zone quality, which has been largely ignored in the literature. Moreover, the result formalizes the basis risk decomposition intuited by Elabed et al. (2013). As we illustrated above, absent this theoretical contribution, nonsensical situations can arise where "imperfect" indices perform better than the supposedly "perfect" one, leading to an ill-defined

---

[6]The sum of squares total is simply the quantity $\sum_i (y_i - \bar{y})^2$.

[7]One can write $\overline{\overline{R^2}} = \sum_i \alpha_i R_i^2$, with $\alpha_i = SST_i / \sum_j SST_j = var(y_i) / \sum var(y_j)$.

negative design risk. Finally, the two theorems reveal a connection between index insurance and principal component analysis that was, to the best of our knowledge, not known. This allows to rephrase the premise of index insurance in statistical terms: in essence, one is assuming that the data can be adequately represented by a one-factor model. This is, in fact, the implicit assumption between most of the models on index insurance (Miranda, 1991; Mahul, 1999; Elabed et al., 2013; Bucheli et al., 2020). Establishing this link opens the door to enrich further our understanding of index insurance by drawing on the rich literature on factor models (Bai, 2003; Bai and Li, 2012).

3 Data

The main dataset we use consists of satellite-based maize yield estimates from 2016 to 2019 in Kenya. The data span all regions in Kenya that are cropped with maize. Yields were predicted with the SCYM model, which uses a crop model to estimate the relationship between satellite measures of greenness and final crop yield (Lobell et al., 2015; Burke and Lobell, 2017; Jin et al., 2017b, 2019). The estimated relationship is then used to predict yields throughout the country using the satellite observations as input. When assessed against ground-truth measures in Kenya, Burke and Lobell (2017) report a prediction R^2 between .15 and 0.4, Jin et al. (2017a) report an R^2 of 0.4 against average yields at the county level and 0.3 against yields at the sub-county level. In a third iteration of the model, Jin et al. (2019) report an accuracy of 0.5 against average yields at the county level. While there is room for progress, these numbers are encouraging, as predicting yields in smallholder systems with high heterogeneity of practices and very small field size is particularly challenging, and errors in ground-truth will lead to underestimation of satellite performance (Lobell et al., 2021). Here we use estimates from a fourth iteration of the model, which were provided by Atlas AI. Figure 3 shows the average yields over the area cropped with maize, as well as annual yields for a representative location. The SCYM model captures both large regional discrepancies as well as very localized differences in yields over the years. One potential concern with using satellite-based yield estimates is that measurement errors could influence results. For this reason, we also utilize crop-cut yields available for a subset of our study region to evaluate measurement error and demonstrate that our main conclusions are robust to plausible estimates of measurement error (see 5.1).

Given the very high number of pixels covering the whole area, we opt for a field-level analysis by randomly selecting 200 pseudo fields in each of Kenya's lowest administrative divisions, the *ward*. To define a pseudo-field, we consider only "interior" pixels whose adjacent pixels are also classified as maize, and randomly select 200 interior pixels from the pool of all interior pixels in a ward. This results in a dataset of 159 600 units in 798 wards observed over four years.

To investigate the impact of using smaller insurance zones, we make use of the very detailed administrative zones in Kenya. Kenya is divided into three administrative levels, the
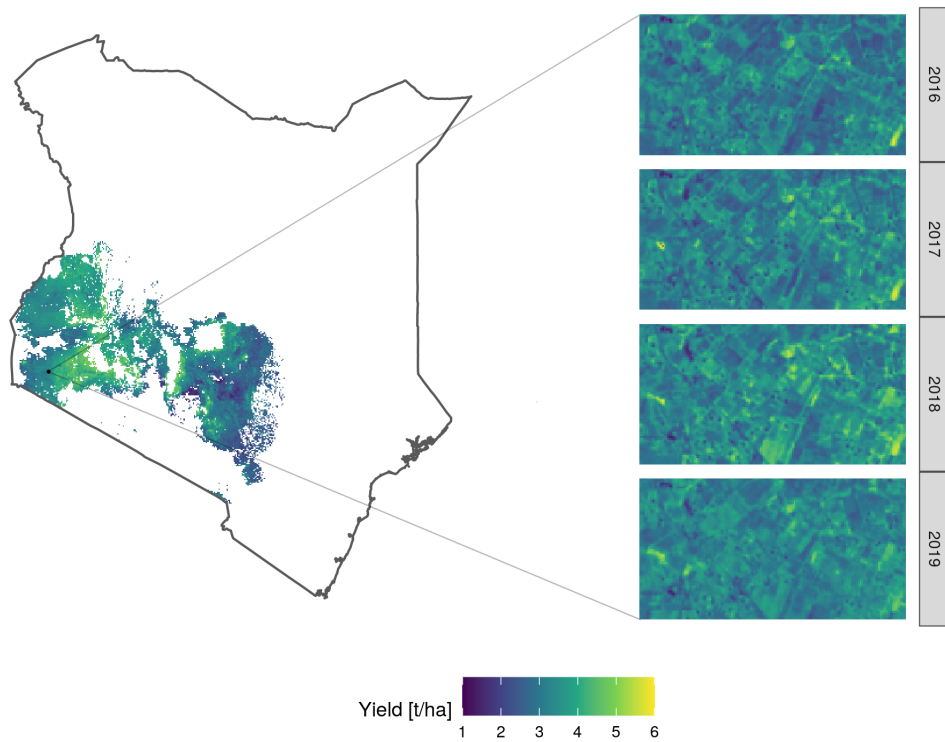
Figure 3: Yields in Kenya as predicted by the SCYM model. The left panel shows the SCYM yield predictions for the whole country, averaged over the four years. The right panel shows annual yields in a randomly selected region.
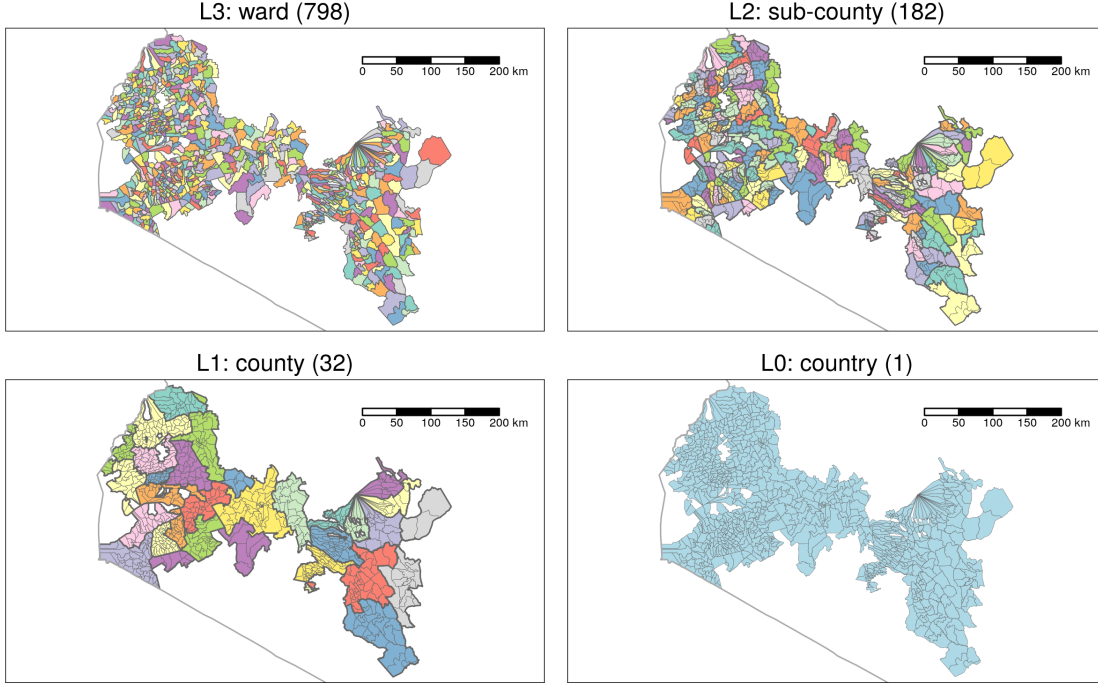
Figure 4: Kenya administrative units. The numbers in parentheses show the total number of units for each administrative level.

county, the sub-county and the ward, which we will abbreviate as L1, L2 and L3. We sampled 200 fields within each of the 798 wards where we have available data, to yield a sample of 159 600 fields. These 798 wards are part of 182 sub-counties (L2) and 32 counties (L1). Figure 4 shows the administrative boundaries for each level. The sub-county (L2) and wards (L3) are relatively small administrative units, with an average area of 97 km^2 for the wards and of 427 km^2 for the sub-counties (see Table 1).

We also construct several weather-based indices to analyze design risk. These indices include minimum and maximum temperature (*tmmn* and *tmmx*), precipitation (*precip*), and indicators of soil moisture (vapour pressure deficit, *vpd* and soil moisture *ssm*). We also include two popular vegetation indices, the normalized difference index (NDVI) and the related enhanced vegetation index (EVI). Strictly speaking, these last two indices are targeting the yields directly so they could be considered as output-based indices in our classification. However, given that they are only proxies, we consider those as input-based indices here. The data comes from four different sources: 1) the vegetation indices come from the MODIS satellite, 2) temperature, vapour pressure deficit and precipitation are extracted from TerraClimate (Abatzoglou et al., 2018) and 3) an alternative precipitation variable comes from the CHIRPS dataset (Funk et al., 2015). 4) soil moisture is derived from SMAP (O'Neill et al., 2016).

Table 1: Zonal risk measures

Level	Id	N units	N fields (ave)	Area $[km^2]$	$\overline{R^2(f^\star)}$	Zonal risk $\overline{1 - R^2(f^\star)}$
Country	L0	1	159600	77739	41.1%	58.9%
County	L1	32	4988	2429	47.9%	52.1%
Sub-county	L2	182	877	427	49.7%	50.3%
Ward	L3	798	200	97	52.2%	47.8%

4 Results

4.1 Zonal risk

To start, we present our estimates of the irreducible zonal risk $\overline{1 - R^2(f^*)}$, that is the lowest basis risk *any* index can reach in a given zone. For ease of exposition, we focus on the complement of zonal risk, the covariant risk $\overline{R^2(f^*)}$. Recall that this is given by the relative share of the first eigenvalue of the correlation matrix for the first metric $\overline{R^2(f^*)} = \lambda_1^C / \sum_i \lambda_i^C$, and by that of the covariance matrix for the second metric $\overline{\overline{R^2(f^*)}} = \lambda_1^\Sigma / \sum_i \lambda_i^\Sigma$. In practice we found that the two metrics were very close and consequently report only the values for $\overline{R^2(f^*)}$.

Table 1 shows the covariant risk for various definitions of the insurance zone. We start by considering a national insurance zone, i.e. use a single index for the whole country. We find a $\overline{R^2(f^*)}$ of 41.1%, implying an irreducible zonal risk $\overline{1 - R^2(f^*)}$ of 58.9%. Finding that barely 40% of the total variation in yields is explained by the index seems rather small, but this could be due to the use of a very large insurance zone covering heterogeneous regions. To gain more insights into this heterogeneity, we vary the definition of the insurance zone using either the county, sub-county or ward level. To do so, we compute the $\overline{R^2(f^*)}$ for each administrative unit and take their average over the whole country. Looking again at Table 1, we see that using instead county (L1) boundaries for the insurance zones, the $\overline{R^2(f^*)}$ increases from 41% to 48%, reducing the zonal risk from 59% to 52%. Using sub-county (L2) or wards (L3) as insurance zones further reduces the zonal risk, yet to a smaller extent. At the finest level (L3 wards), the zonal risk averaged over the 798 wards units is still 48%, meaning that the best local index is only capturing half of the local variations in yields. To put this number in context, Stigler and Lobell (2021) report that at the lowest administrative level in the United States—the county—the average zonal risk is close to 25% for maize. Noting that a US county is about 17 times bigger on average than a ward in Kenya, this comparison suggests a substantial local heterogeneity in Kenya.

The analysis so far used administrative boundaries to delineate insurance zones. However, administrative boundaries tend to be arbitrary, and do not necessarily reflect natural homogeneous zones in terms of agronomic conditions. This is particularly true in the Eastern part of our sample, where wards are in a "wheel and spoke" configuration, where the center is Mount Kenya. Given the high density of fields we have in the sample, we can investigate the effect of using alternate definitions of the zones, as well as using even smaller zones.

To create smaller insurance zones, we draw a circle around each field and vary the size of the circle's radius from 50 meters to 50000m. Using these zone definitions is perhaps an unfeasible scenario but nonetheless provides a useful benchmark.[8] To ensure that there are enough fields in each zone, we extend our definition of pseudo fields to encompass all maize pixels within a given circle instead of only considering the 200 pseudo fields initially sampled. We then compute our measure of zonal risk among all pseudo fields found within each circle. To ensure that we do not include pixels that belong to the central field, we repeat the exercise computing the same measure yet excluding pseudo fields less than 50m from the reference field.

Figure 5 shows the covariant risk averaged among all fields for various radius sizes. Lines in black represent the average "radius" of the administrative zones[9] together with the average zonal risk within the administrative units. For zones with a radius smaller than 1000m, the global $\overline{R^2}$ is substantially higher than values observed using administrative boundaries. When the circle's radius approaches the average radius of the administrative zones, results become very similar. For small zones, excluding the 50m inner circle from the computation leads to lower covariant risk. However, the difference vanishes quickly and becomes imperceptible with radius sizes of 1 km or more.

The analysis so far focused on the mean covariant risk averaged over all insurance zones. However, the average might conceal important variations between the insurance zones. A map of zonal risk in each ward (Figure 6) reveals a large variation, with values of $\overline{R^2(f^*)}$ varying from 35% to 85%. The map also reveals important spatial patterns, indicating that the covariant risk $\overline{R^2(f^*)}$ is higher in the central and eastern part than in the western region. In general, zonal homogeneity seems to be clustered within regions, with clusters of high as well as very low values in the western part.

4.2 Design risk

We investigate now the extent of design risk, that is the additional risk introduced by using a feasible index instead of the optimal one. The feasible index can be either output- or input-based. For output-based indices, we consider the standard zone-average yield (also called area-yield), a scheme implemented in the USA insurance program, and which World Bank (2015) suggested implementing in Kenya. For input-based indices, we consider various indices including weather indices (temperature, precipitation, etc) and satellite vegetation indices.

---

[8]An alternative benchmark could use small non-overlapping zones instead of using a field-specific zone as we do here. Segmenting the whole maize area into contiguous homogeneous zones is, however, a very difficult problem computationally that we leave for future research.

[9]We use as measure of a polygone's average "radius" the average length of the segments linking each edge to the centroid of the polygon.
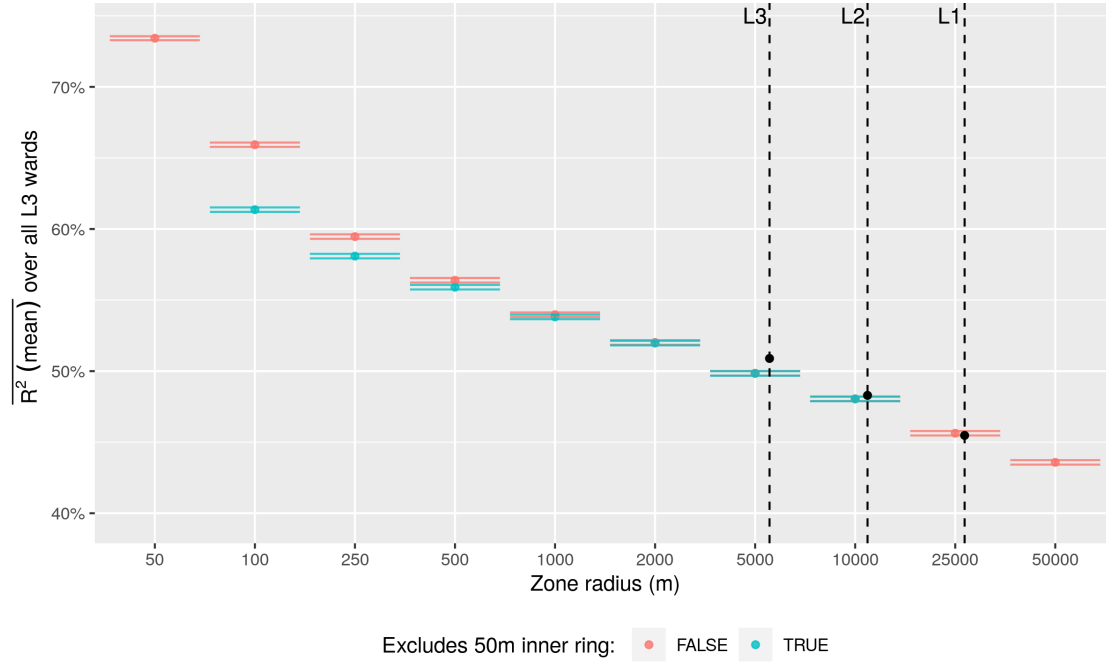
Figure 5: Average R^2 using local zones. The dashed lines represent the average "radius" of the administrative levels. The black dots represent the corresponding average zonal R^2 at each administrative level.
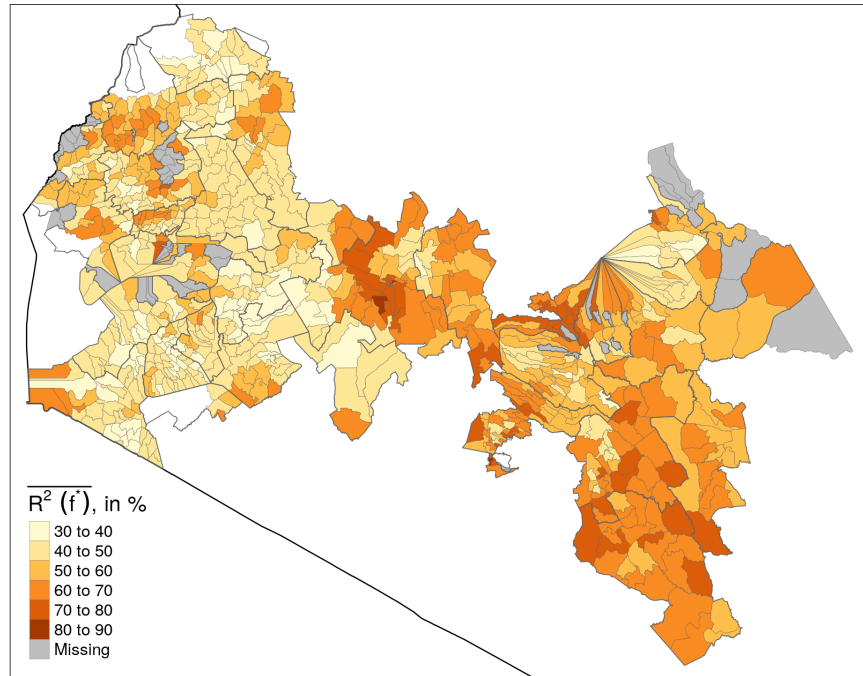

Figure 6: Covariant risk $\overline{R^2(f^*)}$ using wards (L3) administrative boundaries
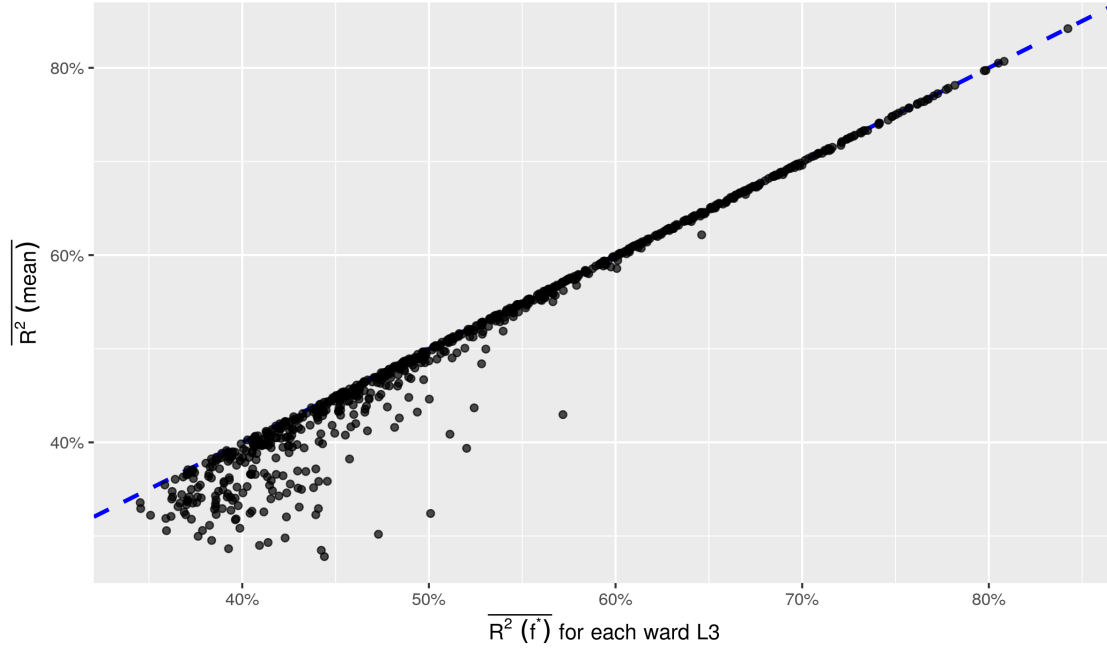
Figure 7: Design risk: Comparing $\overline{R^2}$ from optimal and zone-average index

Starting with an index based on the average yield, Figure 7 compares the basis risk measures using either the optimal index $\overline{R^2(f^*)}$ or average yields $\overline{R^2(\bar{y})}$ for each of the 798 wards. As predicted by the theory developed above, the $\overline{R^2}$ based on the optimal index is always greater than or equal to the $\overline{R^2}$ based on the zone-average index. Interestingly, for high values of the underlying optimal index, the two measures are very close. For lower values of the optimal index, $\overline{R^2(\bar{y})}$ can be substantially lower than the $\overline{R^2(f^*)}$. This result implies that in zones that already have a high zonal risk (a low covariant risk), using a zone-average index adds a supplementary amount of design risk, thus increasing total basis risk. For example, in some wards with a zonal basis risk of 60%, the total basis risk goes up to 70% using an average-yield index rather than the optimal index. Conversely, the potentially insurable covariant risk, initially of 40%, would be reduced to an actually insured covariant risk of 30%.

Table 2 shows the average $\overline{R^2(f)}$ measures at the ward level (L3) for various indices f together with the $\overline{R^2(\bar{y})}$ obtained using the zone-average $f = \bar{y}$ and the optimal index $f^* = \mathbf{Y}w^*$. The column *N wards* indicates the number of wards for which we have full data. The full data is not always available due to missing observations in the weather products. Turning to the main column $\overline{R^2(f)}$, the two first rows under the *Yields* category report the numbers reported above, of $\overline{R^2(f^\star)} = 52\%$ for the optimal index and $\overline{R^2(\bar{y})} = 51\%$ for the index using zone averages. We move now to input-based indices, which include various simple measures of weather (precipitation, min and max temperature), soil moisture, vapour

Table 2: Design risk: $\overline{R^2(f)}$ from weather

Dataset	Variable	Resolution	N wards	$\overline{R^2(f)}$	Cor with opt index
Yields	Optimal Index	10m	798	52.2	
	Zone mean	10m	798	50.9	0.98
MODIS	EVI	500m	733	11.7	0.55
	NDVI	500m	733	11.6	0.56
CHIRPS	Precip	~5000m	798	10.9	0.50
TerraClimate	Precip	~4000m	798	10.0	0.38
	Temp min	~4000m	798	11.7	0.48
	Temp max	~4000m	798	11.0	0.43
	VPD	~4000m	798	11.9	0.47
SMAP	Surface soil moisture	~12000m	784	11.1	0.57
	Subsurface soil moisture	~12000m	784	11.1	0.56

pressure deficit (VPD) or vegetation conditions (EVI and NDVI). The indices are built by using a growing-season temporal average for each pseudo field and are then spatially averaged over the pseudo fields within each ward.

We observe a very sharp drop in the $\overline{R^2(f)}$ For the input-based indices, with all indices averaging around 11%, This implies a total basis risk close to 90%. With such a high basis risk for input-based indices, implementation of a scheme with a weather index would likely encounter limited success. Surprisingly, we see little differences in the indices themselves, all leading to an average $\overline{R^2(f)}$ around 11%. Where we see some differences is in the correlation between the $\overline{R_w^2(f)}$ and the $\overline{R_w^2(f^*)}$ at the ward level w, shown in the column *Cor with opt index*. The temperature and precipitation indices have the lowest correlation with the index, while measures based on vegetation (NDVI, EVI) or soil moisture (SMAP) have higher correlations, close to 50%. This suggests that the *zone quality* (the R^2 from the optimal index) is weakly linked to the total basis risk (the R^2 from the optimal index).

Several factors can contribute to the high estimated values for design risk. First, we used relatively simple weather indices, based on average values over the growing-season months. Additional effort on index design would likely result in better indices. A second explanation could lie in the resolution of the data. While the resolution of the yield data is at $10m \times 10m$ (from the Sentinel 2 satellite), the resolution of the other products ranges from $500m^2$ to $10km^2$. With broader resolutions, the probability that a pixel includes information from other fields as well as from non-agricultural elements (roads, forests, etc) inevitably increases. This "pixel contamination effect" could partly account for the high design risk observed here. A final explanation could come from measurement error in the SCYM yield data, which would attenuate the correlation between predicted yields and the weather indices, hence inflating design risk. While carefully disentangling each explanation lies beyond the scope of this paper, the next section addresses measurement error in the SCYM data.

Table 3: SCYM measurement error: comparing SCYM against crop cuts from 1AF

Regression Type	Cor	Coefficient	Gamma	$P(\gamma = 1)$
Pooled	0.39	γ	1.20	0.06
Temporal	0.50	$\overline{\gamma_i}$	1.05	0.77
Spatial	0.37	$\overline{\gamma_t}$	1.14	0.19

5 Robustness checks

5.1 Impact of measurement error

While using satellite data provides us with both a very detailed and very broad dataset compared to alternative data sources, it has the potential drawback of being less precise than traditional ground truth measurements. The analysis so far was carried assuming that the satellite data was exempt of measurement error. To investigate the impact of measurement error, we leverage a dataset of crop cuts collected by the One Acre Fund Foundation (1AF). This is a particularly rich crop cuts dataset containing 14 000 samples collected over the same four years, yet it comes with its own limitations. The sampled area changed from year to year and geo-referencing of the fields is too imprecise to attribute a crop cut to a specific field. This leads us to construct a panel dataset aggregating crop cuts at the ward (L3) level instead of using the field as unit of analysis. Using only wards that have four years of observations and at least three crop cuts each year, we end up with a dataset of 107 units. The SCYM dataset is constructed in the same way, aggregating our pseudo fields at the ward level for the same 107 units.

To assess the measurement error structure, we estimate the following regression:

$$y_{it}^{SCYM} = \alpha + \gamma y_{it}^{cCut} + \varepsilon_{it} \tag{6}$$

We report two statistics from regression (6), the correlation between the SCYM and the crop cuts ρ, as well as the slope coefficient γ. The slope parameter γ is of particular interest, since it informs about the conditional bias of the predictions. A value close to 1 indicates that the predictions are "conditionally unbiased". On the other hand, the intercept parameter *alpha* is of less interest, since presence of an unconditional bias *alpha* $\neq 0$ does not impact estimates of variance, covariance or correlation. The first row of table 3 shows that for the pooled regression (6), the correlation between the predicted and ground-truth value is 0.39. This is somewhat smaller than the accuracy reported in the original papers (Jin et al., 2017a, 2019), which were based on data from 2015 to 2017. The slope coefficient is 1.2, and the hypothesis that it is different from 1 can be rejected at the 10% level, yet not at 5%.

The slope coefficient in the measurement error regression (6) corresponds to a pooled panel, and is hence a mixture of cross-sectional and temporal regressions (Wooldridge, 2010). As our basis risk measures are essentially based on regressions over time, we might

Table 4: SCYM measurement error: covariant risk estimated on crop cuts versus SCYM data

		$\overline{R^2(f^\star)}$		
Analysis level	N. units	Crop Cut	SCYM: pseudo fields	SCYM: all pixels
L1	14	0.52	0.45	0.50
L2	49	0.52	0.48	0.46
L3	103	0.39	0.50	0.51

rather be interested in $\overline{\rho_i}$ and $\overline{\gamma_i}$, the average coefficients from a regression over time, i.e. $y_{it}^{SCYM} = \alpha_i + \gamma_i y_{it}^{cCut} + \varepsilon_{it}$. We hence compute each regression separately, and report the averages $\overline{\rho_i}$ and $\overline{\gamma_i}$ in the second row of Table 3.[10] The average temporal correlation $\overline{\rho_i}$ is 0.5, higher than the pooled ρ which was 0.39. The $\overline{\gamma_i}$ parameter is also closer to 1, and now statistically not distinguishable from 1. The same exercise can be repeated for the cross-sectional variation, tracking how SCYM correlates on average with the crop cuts over space. The corresponding parameter $\overline{\rho_t}$ and $\overline{\gamma_t}$ are shown in the third row. The average correlation with the ground truth data $\overline{\rho_t}$ is 0.37, closer to the 0.39 we found for the pooled correlation ρ. The slope parameter $\overline{\gamma_t}$ is not statistically different from 1 either. Taken together, these findings indicate that while still leaving room for progress, SCYM's accuracy is higher for temporal than for cross-sectional variations, and has no conditional bias. These two features are particularly important in our context, where we focus on estimating basis risk from regressions over time.

The ground truth dataset can also be used to assess the sensitivity of our results to measurement error. To do so, we compute our main statistic of interest, $\overline{R^2(f^\star)}$, on both the ground truth and the corresponding SCYM dataset. Recall that both are constructed by considering a ward as a unit of analysis, and hence do not directly correspond to the values reported elsewhere in the paper. We repeat the same exercise considering the sub-counties (L2) or counties (L1) as units of analysis instead. Table 4 reports the values computed at the three levels of analysis. Comparing the $\overline{R^2(f^\star)}$ estimated on the crop cut or SCYM pseudo-field data, we notice that SCYM under-estimates the covariant risk in two cases (for the analysis at L2 and L1), and over-estimates it at the L3 level. In all these cases, we see that the bias is moderate and never higher than 0.1. Were this bias extrapolated to our analysis at the field level, it seems unlikely that our general finding of a low covariant risk would radically change without measurement error. As a last robustness check, we include an alternative version of the SCYM data, where the units are constructed by averaging all the pixels within a ward instead of restricting it to our 200 randomly sampled pseudo fields. Column *SCYM: all pixels* shows that these values are relatively close to the ones obtained with the pseudo fields.

---

[10]To do inference on $\overline{\gamma_i}$, we exploit the fact that $\overline{\gamma_i}$ is closely related to the γ^{FE_i} obtained using time fixed effects (see Gibbons et al. 2019), and hence use a fixed-effect estimator to test $\overline{\gamma_i} = 1$

5.2 Relation to other measures

The linear correlation measures we use so far have limitations in the context of index insurance. Arguably, it is more important for an index to accurately predict yield losses than to predict good harvests. Various approaches have been suggested to take this into account, ranging from quantile regression (Conradt et al., 2015) to more sophisticated left-tail dependence indices (Bokusheva, 2018). Another approach seeks to directly assess the benefits of index insurance using an expected utility (EU) framework (Carter and Chiu, 2018). Under the standard assumption of risk-aversion and concave risk-utility function, an index that accurately predicts yield losses will be preferred to one that accurately predicts good harvests. While more theoretically consistent, deriving the optimal index according to a quantile loss function or an EU one is very challenging, and we leave this for future work. However, we can at least examine how our correlation-based measures compare with more refined ones. We opt for two measures: a quantile-based correlation measure similar to our linear correlation measure and an expected utility measure.

Computing the quantile measures of basis risk and the EU measures of insurance benefits requires several adjustments and assumptions. For the quantile measure, the previous literature has mainly used quantile regression to determine the indemnity function (Conradt et al., 2015; Bucheli et al., 2020), yet we are not aware of previous work using quantile regression to simply measure basis risk. To do so, we derive a quantile version of our total R^2 measure based on the quantile pseudo R^2 developed by Koenker and Machado (1999). Koenker and Machado (1999) suggest a pseudo $R^2(\tau) = 1 - V(f,\tau)/V(1,\tau)$ at quantile τ, where $V(f,\tau)$ is the quantile analogue to the sum of squared residuals (SSR) in a linear regression with variable f. Likewise, $V(1,\tau)$ is the quantile analogue to the total sum of squares from a regression including only a constant. In a similar way to our average $\overline{R^2(f)}$, we define the average quantile pseudo R^2 as $\overline{R_q^2(f)} = 1 - \sum V_i(f,\tau)/V_i(1,\tau)$, and use the value of $\tau = 0.3$ following previous literature (Bucheli et al., 2020).

Unlike the correlation-based measures, computing the expected utility measures of index insurance requires multiple assumptions. One needs to define an indemnity function, assume a specific form for the utility function and decide which benchmark to use as comparison. For the indemnity function, we use the function in Miranda (1991), $I_t = \max(\lambda \bar{y}_{..} - \bar{y}_{.t}, 0)$, where $\bar{y}_{.t}$ corresponds to the zone-average yield at time t, and $\bar{y}_{..}$ to its value averaged over time. λ is the trigger level, that is the threshold (in percentage) below which an indemnity is paid. We set this level at a relatively high value of 90%, to ensure that payouts occur in our sample. We also use a fair premium, i.e. we set it equal to the expected indemnity, $\pi = E[I_t]$. Yields with insurance $\tilde{y}$ are then given by $\tilde{y}_{it} = y_{it} + I_t - \pi$. For the utility function, we follow previous literature (Wang et al., 1998; Deng et al., 2008; Flatnes et al., 2018) and use a constant relative risk aversion (CRRA) iso-elastic utility function, with a parameter of 1.5. Using this function, we compute the expected utility of the post-insurance yields $\tilde{y}_{it}$, $E[u(\tilde{y}_{it})]$, and derive the corresponding certainty equivalent (CE).[11] The last ingredient is the choice of the benchmark, i.e. whether we are comparing the CE of index insurance

---

[11]The certainty equivalent of a lottery y is the sure amount c such that: $u(c) = E[U(y)]$ holds. It has the advantage of being expressed in the same unit as y, whereas EU measures are expressed in "utility units".

to the CE of yields without insurance or to the CE of yields with individual insurance. As showed by Stigler and Lobell (2021), a comparison of index insurance to no insurance will be particularly sensitive to the variance of yields within that zone. That is, a field with a high variance yet little correlation with the index might perform better in that metric than a field with high correlation yet low variance. Conversely, comparing index insurance to individual insurance amounts to comparing two different variance reduction tools. Therefore, it is less sensitive to the raw variance of a given field. Given that our R^2 measure of basis risk is variance-free, the second approach seems more natural in our context. We hence adopt the *farm-equivalent risk coverage* metric proposed by Stigler and Lobell (2021), which seeks the highest level of individual insurance compatible with an index insurance at 90% level.

A final issue with the post-insurance EU measures is that they critically depend on whether an indemnity is paid. This in turn depends on whether the minimum value is above or below 90%, that is, whether the lowest observed (zone-average) yield is below 90% of its mean. If not, there is no insurance payment at all, making the evaluation of the benefit of insurance impossible. Unfortunately, with just four years of data, this happens relatively often: 66% of the L3 wards never experience a minimum value below 90% of their mean. To address this problem, we proceed to a data simulation. We first estimate regression (1) relating field's i to the ward average, extracting the $\hat{\alpha}_i$, $\hat{\beta}_i$ and $\hat{\sigma}_i$. We then simulate 30 years of ward-average yields $\hat{y}_{jt}$, based on a normal distribution with mean and variance parameters estimated from the yield data. We finally simulate individual yields by plugging in the field-level estimates and the simulated means, that is $\hat{y}_{ijt} \sim \mathcal{N}\left(\hat{\alpha}_i + \hat{\beta}_i\hat{y}_{jt}, \hat{\sigma}_i\right)$. As a result of this simulation, we now only have 30% of wards for which the relative minimum is never below 90%.

Figure (8) compares the simple linear correlation $\overline{R^2(\bar{y})}$ to the quantile correlation $\overline{R_q^2(\bar{y})}$ and to the EU-based *farm-equivalent risk coverage.* The two more sophisticated measures seem to agree remarkably well with the simple linear correlation measure used throughout. The correlation with the quantile pseudo R^2 is 95%, while the correlation with the EU-based measure is 84%. These results are very encouraging, as they suggest that our focus on simple correlation to describe basis risk is justified, since this measure seems to be a strong proxy for more comprehensive measures. Hence, the bound on basis risk we obtained under the linear correlation criterion may be a good approximation for the bound on these other metrics.

6 Conclusions

Basis risk is often recognized as one of the main impediments to the adoption of index insurance. This has led to an extensive literature on the best approach to lowering basis risk. Largely absent in this literature is a recognition that basis risk is comprised of both zonal and design risk, and even the best possible index can only reduce design risk but not zonal risk. Intuitively, in a region with very heterogeneous fields showing little correlation
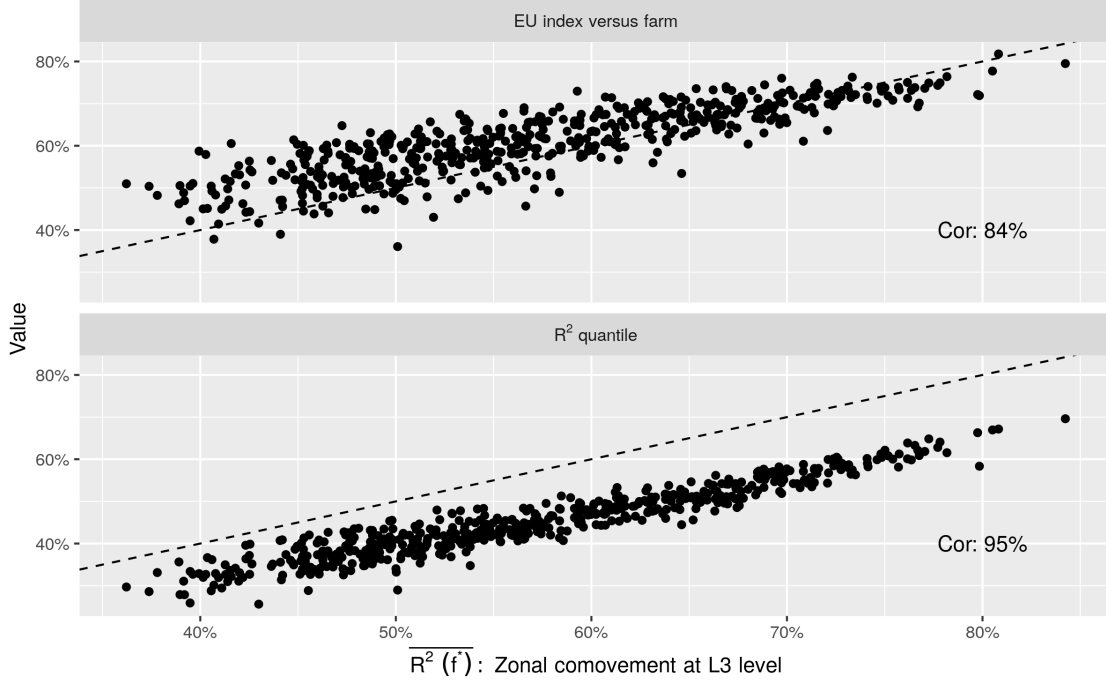
Figure 8: Comparing $\overline{R^2}$ with quantile $\overline{R_q^2}$ and EU measures

over time, a good index can do little to reduce basis risk. Without a proper understanding of zonal risk, it is impossible to distinguish a "good index in a bad zone" from a "bad index in a good zone". This distinction is, however, crucial since it indicates whether research efforts should be spent on designing better indices or refining insurance zone delineations.

This paper provides a theoretical decomposition of basis risk and illustrates its application to maize yield data in Kenya. We formalize the basis risk decomposition of Elabed et al. (2013), according to which basis risk can be decomposed within a given insurance zone into zonal and design risk. Ensuring that this decomposition is well-defined requires a notion of the optimal index for a given insurance zone, a question Elabed et al. (2013) left open. We fill this gap by deriving the optimal index according to two linear correlation criteria. This provides a useful upper bound on basis risk reduction for a given zone, beyond which basis risk can only be reduced by redefining the zone itself.

Our second contribution is to illustrate how this decomposition can be used to understand the sources of basis risk for maize in Kenya. Using satellite-derived maps of maize yields for four years, we investigate the extent of the irreducible zonal risk and of the design risk arising from using a feasible index. We find the irreducible zonal risk to be rather high, about 48% using the smallest administrative divisions available in Kenya. By definition, this zonal risk can only be reduced by considering refined zones and we investigate the benefits of using satellite data to define smaller zones. The results are rather encouraging, indicating that zonal risk can be reduced to about 40% using local indices based on 250m neighborhoods around each field. We document also strong regional patterns, with the zonal

risk being much higher in the western part of Kenya. Turning to the additional design risk, we find that total basis risk is only slightly increased by using zone-average yields instead of the optimal index, but exacerbated by using weather-based indices such as weather or soil moisture. As a direct consequence, this paper confirms the recommendations by World Bank (2015) to implement an area-yield index scheme in Kenya. Yet, at the same time, our results on the spatial distribution of zonal risk indicate that the success of such a scheme is likely to be limited in some regions. Our findings also have important implications for the design of measurement surveys to inform a future area-yield scheme. The strong local heterogeneity we document suggests that accurately measuring average yields will require collecting data on many fields, which could be very costly with traditional approaches. With further advances in remote sensing techniques, the question will soon arise whether using many cheap imperfect measures from satellite data can be more useful than using a few costly ground measures, even if the accuracy of ground measures is higher (for a related discussion, see Browning and Crossley, 2009).

This work can be extended in several ways. At the theoretical level, we focused on linear correlation-based measures of basis risk, which allowed us to derive analytical formulas for the irreducible zonal risk. Arguably, non-linear measures of basis risk based on quantile regression, or even expected utility measures of yields covered by index insurance, are more appropriate to assess the benefits of index insurance. Extending our theoretical results to a risk decomposition based on these criteria will be an beneficial endeavor for future research. Due to the non-linearity of these measures, deriving the optimal index will likely be a challenging task, without a guarantee that an analytical solution for the optimal index exists. Most likely, methods based on numerical approximation will be required, which might become cumbersome for large datasets.

On the empirical side, our analysis could be extended in several ways, notably by using fields instead of pixels and by using longer time series with improved accuracy. Doing an analysis at the field level instead of using pixels would reveal important insights into the variability of yields within- and between fields, yet it could be very costly if the data collection were done in the traditional way by manually geo-referencing fields on the ground. Methods based on satellite imagery, either using manual digitization based on crowdsourcing (Estes et al., 2016) or using image segmentation based on deep learning (Yan and Roy, 2016; Waldner and Diakogiannis, 2020; Waldner et al., 2021) seem very promising. Likewise, the empirical results of this paper were obtained using a relatively short dataset of yields covering just four years. While we believe that this provides essential insights into basis risk in Kenya, reconsideration of our results when longer time series become available will be needed. Given the rapid progress in remote sensing and yield prediction techniques, improved dataset spanning longer time periods should be available in the near future to refine our estimates. This will open the way for an in-depth analysis of our proposal to design local indices with satellite data, which we show here to be potentially very effective at reducing zonal risk. Of particular interest will be to analyze the trade-off in the optimal zone size between the usefulness of local indices and the risk of moral hazard.

References

ABATZOGLOU, J. T., S. Z. DOBROWSKI, S. A. PARKS, AND K. C. HEGEWISCH (2018): "TerraClimate, a high-resolution global dataset of monthly climate and climatic water balance from 1958–2015," *Scientific Data*, 5, 170191.

BAI, J. (2003): "Inferential theory for factor models of large dimensions," *Econometrica*, 71, 135–171.

BAI, J. AND K. LI (2012): "Statistical Analysis of Factor Models of High Dimension," *The Annals of Statistics*, 40, 436–465.

BENAMI, E., Z. JIN, M. R. CARTER, A. GHOSH, R. J. HIJMANS, A. HOBBS, B. KENDUIYWO, AND D. B. LOBELL (2021): "Uniting Remote Sensing, Crop Modeling, and Economics for Agricultural Risk Management," *Nature Reviews Earth and Environment*, 21:2, 140–159.

BOKUSHEVA, R. (2018): "Using copulas for rating weather index insurancecontracts, Journal of Applied Statistics," *Journal of Applied Statistics*, 45:13, 2328–2356.

BOURGEON, J.-M. AND R. G. CHAMBERS (2003): "Optimal Area-Yield Crop Insurance Reconsidered," *American Journal of Agricultural Economics*, 85, 590–604.

BROWNING, M. AND T. CROSSLEY (2009): "Are Two Cheap, Noisy Measures Better than One Expensive, Accurate One?" *The American Economic Review*, 99, 99–103.

BUCHELI, J., T. DALHAUS, AND R. FINGER (2020): "The optimal drought index for designing weather index insurance," *European Review of Agricultural Economics*, jbaa014.

BURKE, M. AND D. LOBELL (2017): "Satellite-based assessment of yield variation and its determinants in smallholder African systems," *PNAS*, 114:9, 2189–2194.

CARTER, M. R. AND T. CHIU (2018): "Quality Standards for Agricultural Index Insurance: An Agenda for Action." Url: https://microinsurancenetwork.org/sites/default/files/SoM_2018_WEB_final.pdf.

CHANTARAT, S., A. G. MUDE, C. B. BARRETT, AND M. R. CARTER (2013): "Designing Index-Based Livestock Insurance for Managing Asset Risk in Northern Kenya," *Journal of Risk and Insurance*, 80(1), 205–237.

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

CLARKE, D. J., O. MAHUL, K. N. RAO, AND N. VERMA (2012): "Weather based crop insurance in India," Policy Research Working Paper Series 5985, The World Bank.

COLE, S., X. GINÉ, J. TOBACMAN, P. TOPALOVA, R. TOWNSEND, AND J. VICKERY (2013): "Barriers to Household Risk Management: Evidence from India," *American Economic Journal: Applied Economics*, 5, 104–35.

COLE, S., X. GINÉ, AND J. VICKERY (2017): "How Does Risk Management Influence Production Decisions? Evidence from a Field Experiment," *The Review of Financial Studies*, 30, 1935–1970.

CONRADT, S., R. FINGER, AND R. BOKUSHEVA (2015): "Tailored to the extremes: Quantile regression for index-based insurance contract design," *Agricultural Economics*, 46, 537–547.

DENG, X., B. J. BARNETT, G. HOOGENBOOM, Y. YU, AND A. G. Y. GARCIA (2008): "Alternative Crop Insurance Indexes," *Journal of Agricultural and Applied Economics*, 40, 223–237.

ELABED, G., M. F. BELLEMARE, M. R. CARTER, AND C. GUIRKINGER (2013): "Managing basis risk with multiscale index insurance," *Agricultural Economics*, 44, 419–431.

ESTES, L., D. MCRITCHIE, J. CHOI, S. DEBATS, T. EVANS, W. GUTHE, D. LUO, G. RAGAZZO, R. ZEMPLENI, AND K. CAYLOR (2016): "A platform for crowdsourcing the creation of representative, accurate landcover maps," *Environmental Modelling & Software*, 80, 41 – 53.

FLATNES, J. E., M. R. CARTER, AND R. MERCOVICH (2018): "Improving the Quality of Index Insurance with a Satellite-based Conditional Audit Contract," Tech. rep., Basis, Feed the Future Innovation Lab for Markets, Risk and Resilience, UC Davis.

FUNK, C., P. PETERSON, M. LANDSFELD, D. PEDREROS, J. VERDIN, S. SHUKLA, G. HUSAK, J. ROWLAND, L. HARRISON, A. HOELL, AND J. MICHAELSEN (2015): "The climate hazards infrared precipitation with stations—a new environmental record for monitoring extremes," *Scientific Data*, 2, 150066.

GIBBONS, C., J. C. SUÁREZ SERRATO, AND M. URBANCIC (2019): "Broken or Fixed Effects?" *Journal of Econometric Methods*, 8, 1–12.

JENSEN, N., Q. STOEFFLER, F. FAVA, A. VRIELING, C. ATZBERGER, M. MERONI, A. MUDE, AND M. CARTER (2019): "Does the design matter? Comparing satellite-based indices for insuring pastoralists against drought," *Ecological Economics*, 162, 59–73.

JENSEN, N. D., C. B. BARRETT, AND A. G. MUDE (2016): "Index Insurance Quality and Basis Risk: Evidence from Northern Kenya," *American Journal of Agricultural Economics*, 98, 1450–1469.

JIN, Z., G. AZZARI, M. BURKE, S. ASTON, AND L. DAVID (2017a): "Mapping Smallholder Yield Heterogeneity at Multiple Scales in Eastern Africa." *Remote Sensing*, 9(9):931.

JIN, Z., G. AZZARI, AND D. B. LOBELL (2017b): "Improving the accuracy of satellite-based high-resolution yield estimation: A test of multiple scalable approaches," *Agricultural and Forest Meteorology*, 247, 207 – 220.

JIN, Z., G. AZZARI, C. YOU, S. DI TOMMASO, S. ASTON, M. BURKE, AND D. B. LOBELL (2019): "Smallholder maize area and yield mapping at national scales with Google Earth Engine," *Remote Sensing of Environment*, 228, 115 – 128.

JOLLIFFE, I. (2002): *Principal Component Analysis, Second Edition*, Springer.

JONG, J.-C. AND S. KOTZ (1999): "On a Relation Between Principal Components and Regression Analysis," *The American Statistician*, 53, 349–351.

KOENKER, R. AND J. A. F. MACHADO (1999): "Goodness of Fit and Related Inference Processes for Quantile Regression," *Journal of the American Statistical Association*, 94, 1296–1310.

LOBELL, D. B., S. DI TOMMASO, M. BURKE, AND T. KILIC (2021): "Twice Is Nice: The Benefits of Two Ground Measures for Evaluating the Accuracy of Satellite-Based Sustainability Estimates," *Remote Sensing*, 13.

LOBELL, D. B., D. THAU, C. SEIFERT, E. ENGLE, AND B. LITTLE (2015): "A scalable satellite-based crop yield mapper," *Remote Sensing of Environment*, 164, 324 – 333.

MAHUL, O. (1999): "Optimum Area Yield Crop Insurance," *American Journal of Agricultural Economics*, 81, 75–82.

——— (2001): "Optimal Insurance Against Climatic Experience," *American Journal of Agricultural Economics*, 83, 593–604.

MAHUL, O. AND B. D. WRIGHT (2003): "Designing Optimal Crop Revenue Insurance," *American Journal of Agricultural Economics*, 85, 580–589.

MEREDITH, W. AND R. E. MILLSAP (1985): "On component analyses," *Psychometrika*, 50, 495–507.

MIRANDA, M. J. (1991): "Area-Yield Crop Insurance Reconsidered," *American Journal of Agricultural Economics*, 73, 233–242.

O'NEILL, P. E., S. CHAN, E. G. NJOKU, T. JACKSON, AND R. BINDLISH (2016): "SMAP L3 Radiometer Global Daily 36 km EASE-Grid Soil Moisture, Version 4. Boulder," Tech. rep., NASA National Snow and Ice Data Center Distributed Active Archive Center.

SKEES, J. R., J. R. BLACK, AND B. J. BARNETT (1997): "Designing and Rating an Area Yield Crop Insurance Contract," *American Journal of Agricultural Economics*, 79, 430–438.

STIGLER, M. AND D. LOBELL (2021): "On the benefits of index insurance in US agriculture: a large-scale analysis using satellite data," Tech. rep., Stanford University, Center for Food Security and the Environment.

VROEGE, W., J. BUCHELI, T. DALHAUS, M. HIRSCHI, AND R. FINGER (2021): "Insuring crops from space: the potential of satellite-retrieved soil moisture to reduce farmers' drought risk exposure," *European Review of Agricultural Economics*, 48, 266–314.

WALDNER, F. AND F. I. DIAKOGIANNIS (2020): "Deep learning on edge: Extracting field boundaries from satellite images with a convolutional neural network," *Remote Sensing of Environment*, 245, 111741.

WALDNER, F., F. I. DIAKOGIANNIS, K. BATCHELOR, M. CICCOTOSTO-CAMP, E. COOPER-WILLIAMS, C. HERRMANN, G. MATA, AND A. TOOVEY (2021): "Detect, Consolidate, Delineate: Scalable Mapping of Field Boundaries Using Satellite Images," *Remote Sensing*, 13.

WANG, H. H., S. D. HANSON, R. J. MYERS, AND J. R. BLACK (1998): "The Effects of Crop Yield Insurance Designs on Farmer Participation and Welfare," *American Journal of Agricultural Economics*, 80, 806–820.

WOOLDRIDGE, J. M. (2010): *Econometric Analysis of Cross Section and Panel Data*, MIT Press, 2nd ed.

WORLD BANK (2015): "Kenya, Toward a National Crop and Livestock Insurance Program, BACKGROUND REPORT," Tech. rep., World Bank.

YAN, L. AND D. ROY (2016): "Conterminous United States crop field size quantification from multi-temporal Landsat data," *Remote Sensing of Environment*, 172, 67 – 86.

YU, J., M. VANDEVEER, J. D. VOLESKY, AND K. HARMONEY (2019): "Estimating the Basis Risk of Rainfall Index Insurance for Pasture, Rangeland, and Forage." *Journal of Agricultural and Resource Economics*, 44, 179 – 193.

ZHANG, J., K. S. TAN, AND C. WENG (2019): "Index Insurance Design," *ASTIN Bulletin*, 49, 491–523.

A   Appendix

A.1   Proofs

Proposition A.1. *We can rewrite (5) with $\tilde{w} = D^{1/2}w$, and inserting $D^{-1/2}D^{1/2}$:*

$$\mathcal{R}^2(w) = D^{-1/2}\Sigma w \left(w'\Sigma w\right)^{-1} w'\Sigma D^{-1/2} \tag{7}$$

$$= D^{-1/2}\Sigma D^{-1/2}D^{1/2}w \left(w'D^{1/2}D^{-1/2}\Sigma D^{-1/2}D^{1/2}w\right)^{-1} w'D^{1/2}D^{-1/2}\Sigma D^{-1/2} \tag{8}$$

$$= C\tilde{w} \left(\tilde{w}'C\tilde{w}\right)^{-1} \tilde{w}'C$$

then Property A.6 (2.3.3) in Jolliffe (2002) shows that the trace of $\mathcal{R}^2(w)$ is maximized using the eigenvector of C, i.e. $\tilde{w}^* = eigenvector(C)$. Then $w^* = D^{-1/2}eigenvector(C) = D^{-1/2}v$ is recovered. Turning to proposition 1.3: by definition of eigenvector, $Cv = \lambda_1 v$, and with normalization $v'v = 1$ this implies $\lambda_1 = v'Cv$. Then $\mathcal{R}^2 = Cv \left(v'Cv\right)^{-1} v'C = \lambda_1 v \left(\lambda_1\right)^{-1} v'\lambda_1 = \lambda_1 vv'$. Now we want $1/N \cdot \text{Tr}(\mathcal{R}^2) = 1/N \cdot \text{Tr}(\lambda_1 vv') = \lambda_1/N$. Because C is a correlation matrix, $\text{Tr}(C) = N = \sum \lambda_j$, hence $1/N \cdot \text{Tr}(\mathcal{R}^2) = \lambda_1/\sum \lambda_j$.