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

Matthieu Stigler, David Lobell
Stanford University,
Center for Food Security and the Environment

UC Davis, ARE Ag Seminar
May 13 2021
Outline

1 Introduction

2 Theory

3 Data

4 Results
   Zonal risk
   Design risk
   Expected utility metrics

5 Summary and further steps
Motivation

Risk in agriculture:

- Farmers face considerable yield and price risk
- Risk has several negative consequences on production

Instruments to reduce risk:

- **Farm based**: indemnity if own field has loss
- **Index based**: payout if index indicates low harvest
  - area-based: index is zone average yield
  - weather-based: index is precipitation, temperature, etc.
  - NDVI, drought index, soil moisture, etc.
Index insurance

Why is index insurance preferred?

- **Pros**: no moral hazard, less adverse selection, lower monitoring costs
- **Cons**: farmer can have a loss yet no payout $\Rightarrow$ basis risk

Basis risk definition(s):

- Probability of not receiving payout while having a loss: $P(\text{no indemnity}|\text{loss})$
- Lack of correlation with the index: $\text{cor}(\text{index}, \text{yield}_i)$. Elabed et al (2013): basis risk $= 1 - R^2$
Sources of basis risk

Decompose the risk into two sources:

- **Zonal risk**: $\Leftarrow$ aggregate index used instead of individual yields.
- **Design risk**: $\Leftarrow$ aggregate co-movement not well summarized by the index.

This decomposition is useful as indicates different solutions:

- **Zonal risk**: can only be improved by changing the zone delimitation.
- **Design risk**: can be improved by searching for a better index.
Sources of basis risk

Decompose the risk into two sources:

- **Zonal risk**: $\Leftarrow$ aggregate index used instead of individual yields.
- **Design risk**: $\Leftarrow$ aggregate co-movement not well summarized by the index.

This decomposition is useful as indicates different solutions:

- **Zonal risk**: can only be improved by changing the zone delimitation.
- **Design risk**: can be improved by searching for a better index.

Current limitations and contributions

Limitations in the literature:
- 95% of current literature has only focused on design risk:
  - Elusive quest of the index: predicting (mean) yields at 100% gives 100% accurate index
- For the other 5%, confusion on risk decomposition:
  - Holds at individual level?
  - What is optimal index?

This paper: basis risk decomposition in Kenya

Contribution:
- Theoretical: formalize decomposition, derive optimal index
- Empirical: quantify the decomposition, investigate effect of changing area vs. changing index.
Outline

1. Introduction

2. Theory

3. Data

4. Results
   - Zonal risk
   - Design risk
   - Expected utility metrics

5. Summary and further steps
We start from the simple regression of individual yields \((y_{it})\) on the index \((f_t)\):

\[
y_{1t} = \alpha_1 + \beta_1 f_{t} + \varepsilon_{1t} \\
\ldots = \ldots \\
y_{Nt} = \alpha_N + \beta_N f_{t} + \varepsilon_{Nt}
\]

We seek here to maximise the average: \(\overline{R^2} = \frac{1}{N} \sum R_i^2\)
What is the optimal index $f_t$? Is it the county average yield, $\bar{y}_t = \frac{1}{N} \sum y_{it}$ as suggested in many papers?

**Counter example**

- Simulate a zone with three fields, depending on weather $W$:
  - $y_{it} = \alpha_i + \beta_i W_{it} + \epsilon_{it}$ \quad $i \in [A, B, C]$  
  - *Lone* unit A: $\text{cor}(A, B) = \text{cor}(A, C) = 0.1$
  - *Similar* units B and C: $\text{cor}(B, C) = 0.9$

- Resulting $R^2$ with weather/mean/submean index:

<table>
<thead>
<tr>
<th>Unit</th>
<th>$R^2$ Weather</th>
<th>$R^2$ Mean</th>
<th>$R^2$ $w=[0,1,1]$</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>0.04</td>
<td>0.19</td>
<td>0.01</td>
</tr>
<tr>
<td>B</td>
<td>0.52</td>
<td>0.80</td>
<td>0.90</td>
</tr>
<tr>
<td>C</td>
<td>0.47</td>
<td>0.79</td>
<td>0.89</td>
</tr>
</tbody>
</table>

| Mean $R^2$: | 0.35 | 0.59 | 0.60 |
Look for optimal output-based index, \( f = \sum w_i y_{it} \), where \( w_i \) is the weight for each field \( i \).

**Theorem**

The average \( \overline{R(w)^2} \) is a function of the covariance between fields \( \Sigma \): \[
\overline{R(w)^2} = \text{trace} \left( D^{-1/2} \Sigma w (w' \Sigma w)^{-1} w' \Sigma D^{-1/2} \right) / N
\]

1. The optimal \( w^* \) is a function of the first principal component of the correlation matrix \( C = D^{-1/2} \Sigma D^{-1/2} \).

2. At the optimum \( w^* \), \( \overline{R(w^*)^2} \) is equal to the share of the first eigenvalue of \( C \), \( \lambda_1^C / \sum_i \lambda_i^C \).
The theorem formalizes the basis risk decomposition:

- **the zonal risk**: is given by the first eigenvalue $\bar{R}(w^*)^2 = \lambda_1^C / \sum \lambda_i^C$.

- **the design risk**: the difference between $\bar{R}(w^*)^2$ and $\bar{R}(f)^2$.

<table>
<thead>
<tr>
<th>Unit</th>
<th>$R^2$ Weather</th>
<th>$R^2$ Mean</th>
<th>$R^2$ $w=[0,1,1]$</th>
<th>$R^2$ Optim</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>0.04</td>
<td>0.19</td>
<td>0.01</td>
<td>0.05</td>
</tr>
<tr>
<td>B</td>
<td>0.52</td>
<td>0.80</td>
<td>0.90</td>
<td>0.89</td>
</tr>
<tr>
<td>C</td>
<td>0.47</td>
<td>0.79</td>
<td>0.89</td>
<td>0.89</td>
</tr>
<tr>
<td>Mean</td>
<td>$R^2$:</td>
<td></td>
<td>0.60</td>
<td>0.61</td>
</tr>
</tbody>
</table>
The theorem formalizes the basis risk decomposition:

- **the zonal risk**: is given by the first eigenvalue $\frac{R(w^*)^2}{\sum \lambda_i} = \frac{\lambda^C_1}{\sum \lambda_i^C}$.

- **the design risk**: the difference between $R(w^*)^2$ and $R(f)^2$.

<table>
<thead>
<tr>
<th>Unit</th>
<th>$R^2$ Weather</th>
<th>$R^2$ Mean</th>
<th>$R^2$ w=[0,1,1]</th>
<th>$R^2$ Optim</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>0.04</td>
<td>0.19</td>
<td>0.01</td>
<td>0.05</td>
</tr>
<tr>
<td>B</td>
<td>0.52</td>
<td>0.80</td>
<td>0.90</td>
<td>0.89</td>
</tr>
<tr>
<td>C</td>
<td>0.47</td>
<td>0.79</td>
<td>0.89</td>
<td>0.89</td>
</tr>
<tr>
<td>Mean</td>
<td>$R^2$:</td>
<td>0.35</td>
<td>0.59</td>
<td>0.60</td>
</tr>
</tbody>
</table>

The zonal risk is highlighted in the table.
What if instead we were to use the 1st eigenvalue of the covariance matrix $\Sigma$?

**Theorem II: optimal index for Total $R^2$**

Let $\overline{R^2} \equiv 1 - \sum_i SSR_i / \sum_i SST_i$ be the *Total SSR*

1. The optimal $w^*$ maximizing $\overline{R(w)^2}$ is the first PC of $\Sigma$.
2. At the optimum $w^*$, $\overline{R(w^*)^2}$ is equal to the share of the first eigenvalue of $\Sigma$, $\lambda_{1,\Sigma}^2 / \sum_i \lambda_i^\Sigma$.

Which criterion is more relevant?

- $\overline{R^2}$ is a variance-weighted average of individual $R_i^2 \Rightarrow$ More weight to more “risky” fields
What if instead we were to use the 1st eigenvalue of the covariance matrix $\Sigma$?

**Theorem II: optimal index for Total $R^2$**

Let $\overline{R^2} \equiv 1 - \sum_i SSR_i / \sum_i SST_i$ be the Total SSR

1. The optimal $w^*$ maximizing $\overline{R(w)^2}$ is the first PC of $\Sigma$.

2. At the optimum $w^*$, $\overline{R(w^*)^2}$ is equal to the share of the first eigenvalue of $\Sigma$, $\frac{\lambda_1^\Sigma}{\sum_i \lambda_i^\Sigma}$

Which criterion is more relevant?

- $\overline{R^2}$ is a variance-weighted average of individual $R_i^2 \Rightarrow$ More weight to more “risky” fields
Limitations:

• Is the (average) correlation of index-yields the right measure?
  • Improvement is not uniformly (Pareto) enhancing!?
  • Index should be particularly correlated with losses, not necessarily gains?
  • Correlation is not causality!

• Estimation of $\lambda$ with $N \gg T$ very hard! Sample eigenvalue over-estimates true $\lambda$. 

Limitations:

- Is the (average) correlation of index-yields the right measure?
  - Improvement is not uniformly (Pareto) enhancing!?
  - Index should be particularly correlated with losses, not necessarily gains?
  - Correlation is not causality utility!
- Estimation of $\lambda$ with $N \gg T$ very hard! Sample eigenvalue over-estimates true $\lambda$. 
Use data in Kenya to quantify:

- **Zonal risk:** effect of varying zone size?
- **Design risk:** using weather index versus zone average yields?
Outline

1 Introduction

2 Theory

3 Data

4 Results
   Zonal risk
   Design risk
   Expected utility metrics

5 Summary and further steps

- Maize yield
- 4 years
- Crop map: created by Atlas AI
- We sample 200 "fields" in 453 divisions based on contiguous maize pixels
A few limitations:

- Only four years of data
- Accuracy of SCYM with crop cuts is \( \sim \) medium
- Using pixels instead of fields \( \Rightarrow \) over-estimate variance, hence under-estimate \( R^2 \) measures?

Still, this is currently the most exhaustive yield estimates data available!
Outline

1. Introduction
2. Theory
3. Data
4. Results
   - Zonal risk
   - Design risk
   - Expected utility metrics
5. Summary and further steps
If one were to use a single insurance zone for whole Kenya, the zonal risk is 
\[ 1 - R(w^*)^2 = 100 - 38 = 62\% . \]

- This is rather high, as using a feasible index will further increase the overall risk.
- Compare this to the US, where we find 100%-55%=45%

How does change by using smaller zone definitions?
Zonal risk

If one were to use a single insurance zone for whole Kenya, the zonal risk is $1 - R(w^*)^2 = 100 - 38 = 62\%$.

- This is rather high, as using a feasible index will further increase the overall risk.
- Compare this to the US, where we find $100\%-55\% = 45\%$

How does change by using smaller zone definitions?

<table>
<thead>
<tr>
<th>Level</th>
<th>Id</th>
<th>N units</th>
<th>N fields (ave)</th>
<th>Zonal common risk</th>
</tr>
</thead>
<tbody>
<tr>
<td>Country</td>
<td>L0</td>
<td>1</td>
<td>90600</td>
<td>61.6%</td>
</tr>
<tr>
<td>County</td>
<td>L1</td>
<td>24</td>
<td>3775</td>
<td>52.4%</td>
</tr>
<tr>
<td>Sub-county</td>
<td>L2</td>
<td>115</td>
<td>788</td>
<td>53.4%</td>
</tr>
<tr>
<td>Ward</td>
<td>L3</td>
<td>453</td>
<td>200</td>
<td>52.4%</td>
</tr>
</tbody>
</table>

Comparing this again to the US, at the finest level, the risk shrinks to 23\%!
Zonal risk II

Comparing with the US:

Figure: $R(w^*)^2$, by zone level
Is the effect of choosing finer zones uniformly positive?

Figure: $R(w^*)^2$, by zone level
Zonal risk III

What if we restrict even more, just using neighboring fields?

Figure: $R(w^*)^2$ on local zones
What is the distribution of the $R(w^*)^2$ within each of the wards (L3 zones)?

**Figure:** Distribution of $R(w^*)^2$ at L3 level
Summary for zonal common risk:

1. Decreases from $1 - R(w^*)^2 = 61\%$ at country level (L0) to 52\% at Ward (L3) level

2. Large differences between L3 zones, from 20\% to 70\%

3. When taking very small zones (20 neighbors), still relatively high
We investigate now the impact of using zone-average yields or weather instead of the unfeasible optimal index.

1. Zone mean versus optimal index?
2. Weather index versus zone mean?
Design risk II

Compare correlation with optimal index \( R(w^*)^2 \) versus correlation with zone-average index:

Figure: \( R(1/N)^2 \) versus \( R(w^*)^2 \)
Look now at weather/VI indices:

- **MODIS**:
  - NDVI
  - EVI

- **CHIRPS**
  - Precipitation

- **TerraClimate**
  - Temp min
  - Temp max
  - Precipitation
  - Vapour Pressure Deficit (VPD)
Design risk IV

Compare $R(f)^2$ for all wards (L3):

**Figure:** Weather indices
What is the average $R^2$ for each measure?

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Variable</th>
<th>$R^2$</th>
<th>Cor with opt index</th>
</tr>
</thead>
<tbody>
<tr>
<td>Yields</td>
<td>Optimal Index</td>
<td>47.56</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Zone mean</td>
<td>45.70</td>
<td>0.96</td>
</tr>
<tr>
<td>MODIS</td>
<td>EVI</td>
<td>9.16</td>
<td>0.60</td>
</tr>
<tr>
<td></td>
<td>NDVI</td>
<td>8.96</td>
<td>0.53</td>
</tr>
<tr>
<td>CHIRPS</td>
<td>Precip</td>
<td>8.40</td>
<td>0.35</td>
</tr>
<tr>
<td>TerraClimate</td>
<td>Precip</td>
<td>8.94</td>
<td>0.27</td>
</tr>
<tr>
<td></td>
<td>Temp min</td>
<td>9.77</td>
<td>0.43</td>
</tr>
<tr>
<td></td>
<td>Temp max</td>
<td>9.31</td>
<td>0.33</td>
</tr>
<tr>
<td></td>
<td>VPD</td>
<td>10.25</td>
<td>0.40</td>
</tr>
</tbody>
</table>
Summary of design risk:

- Index with Zone average yields does not too bad compared to optimal index
- Weather indices and VIs have really low $R(f)^2$
Comparison with expected utility

Expected utility (EU) is considered a more comprehensive criterion. How do our zonal risk measures compare to EU?

Figure: Expected Utility versus Zonal Risk
Outline

1 Introduction
2 Theory
3 Data
4 Results
   Zonal risk
   Design risk
   Expected utility metrics
5 Summary and further steps
Contributions of the paper:

- Formalization of the basis risk decomposition: upper bound on the $R(f)^2$ any index can reach
- Zonal risk: using more local zones increases the $R(w^*)^2$, yet it remains relatively low
- Design risk: Using weather or VI indices leads to a large deterioration in the $R(f)^2$ compared to $R(w^*)^2$
Further steps

Current work:

• Compute expected (EU) utility measures of the various zones/indices. Do those correlate well with the $R(f)^2$ ones?

Maybe:

• Using fields instead of sampled points?
• Correcting for the bias of the sample eigenvalue?
Simulation details

The model in vector form is:

\[ y_t = BW_t + \varepsilon_t \]

Which gives the covariance matrix:

\[ \Sigma_Y = B\Sigma_W B + \Sigma_\varepsilon \]

The simulation uses following values:

\[
\begin{bmatrix}
1.04 & . & . \\
0.14 & 2.10 & . \\
0.14 & 1.60 & 1.90
\end{bmatrix}
= \text{Diag} \begin{pmatrix} 0.20 \\ 1.05 \\ 0.95 \end{pmatrix}
\begin{bmatrix}
1.00 & . & . \\
0.20 & 1.00 & . \\
0.20 & 0.75 & 1.00
\end{bmatrix}
\text{Diag} \begin{pmatrix} 0.20 \\ 1.05 \\ 0.95 \end{pmatrix}
+ \begin{bmatrix}
1.00 & . & . \\
0.10 & 1.00 & . \\
0.10 & 0.85 & 1.00
\end{bmatrix}
\]