Agricultural Index Insurance: An Optimization Approach

José I. Velarde Morales Linwei Xin

University of Chicago
Booth School of Business

July 1, 2025

Project Overview

- The goal of this project is to improve the design of index insurance contracts. I am particularly interested in the developing country setting.
- The original motivation was to develop a method to simultaneously design contracts for all insured zones, in order to better manage risk. However, upon learning more about the context I realized that the optimization based approach could be an improvement even in the single zone case.
- We develop a method that tries to maximize farmer utility, incorporates different kinds of constraints, and yields interpretable contracts.

Motivation for multi-zone optimization

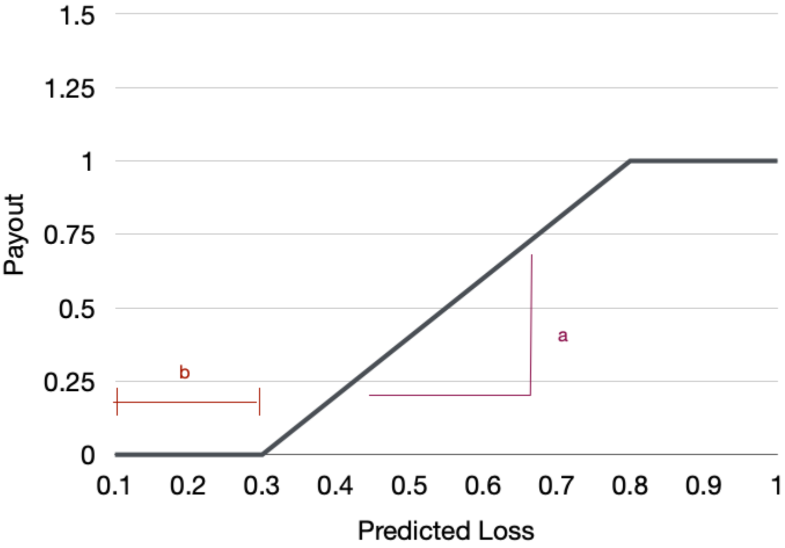
- For the insurer, the risk (and therefor cost) of providing a line of insurance depends on the other lines of insurance in its portfolio.
- If all of its lines of insurance are highly correlated, it will be more expensive due to capital requirements.
- This is similar to the problem of portfolio design, where the correlations between stocks affect optimal portfolio composition.
- Insurers can influence the composition of their portfolio through the contracts they offer, so they should be better able to manage their costs if they take into account the correlation of different lines of insurance when designing contracts.

Content

Index Insurance: Definition and Parameters

- Index insurance uses a signal, θ, to predict agricultural loss, $\hat{\ell}(\theta)$
- Contract form: $I(\theta) = \min\left\{\max\left\{a\hat{\ell}(\theta) - b, 0\right\}, 1\right\}$, a, b are the contract parameters.

Example of Index Insurance Contract



Current methods to design index insurance

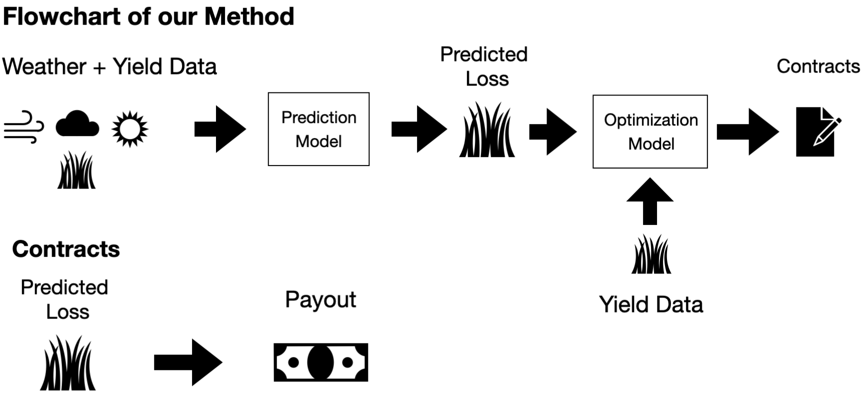
- **Baseline Method:** Predict then design, design contracts to maximize correlation between payouts and losses. Chantarat et al. (2013)
  - **Pros:** simple, interpretable
  - **Cons:** can't incorporate constraints
- **NN Based Method:** End to end, use NN to design contracts. Combine prediction and contract design into single step. Chen et al. (2023)
  - **Pros:** directly maximizes utility, admits price contstraints
  - **Cons:** not interpretable, hard to adjust, requires a lot of data to train.

Content

Overview

- We opt for a "predict-then-optimize" approach.
- We use specialized time-series feature extraction algorithms and traditional ML algorithms (e.g. Random Forest, Gradient Boosting, Support Vector Machines) to build a loss prediction model.
- We then use model predictions and actual losses to design contracts to maximize utility.

Our Method Flowchart



Benefits of Our Approach

- Can admit many common definitions of the premium
- Can incorporate price and payout frequency constraints
- Requires less data to train than NN approach.

Utility Framework/Model

$$
\begin{aligned}
w &= w_0 + 1 - \ell + I(\hat{\ell}) - \pi \\
\pi &= \mathbb{E}[I(\hat{\ell})] + c_\kappa K \\
K &= \text{CVaR}_{1-\epsilon}\left(I(\hat{\ell})\right) - \mathbb{E}[I(\hat{\ell})]
\end{aligned}
$$

CRRA Utility function: $u(w) = \frac{1}{1-\alpha} w^{1-\alpha}$. Premium definition comes from *Risk Modeling for Appraising Named Peril Index Insurance Products: A Guide for Practitioners* (Mapfumo, Groenendaal, and Dugger (2017)). The EU's Solvency requirements are similar.

Idealized Model

$$
\begin{aligned}
\max_{a,b,\pi,K} \quad & \mathbb{E}\left[U\left(w_0 - \ell - \pi + I(\theta)\right)\right] && (1)\\
\text{s.t.} \quad & I(\theta) = \min\left\{\max\left\{0, a\hat{\ell}(\theta) + b\right\}, 1\right\} && (2)\\
& \pi = \mathbb{E}\left[I(\theta)\right] + c_\kappa\left[\mathrm{CVaR}_{1-\epsilon_K}\left(I(\theta)\right) - \mathbb{E}[I(\theta)]\right] && (3)\\
& \pi \leq \overline{\pi}. && (4)
\end{aligned}
$$

Convex Relaxation

$$
\begin{aligned}
\max_{a,b,K,\pi} \quad & \mathbb{E}\left[U\left(w_0 - \ell - \pi + \underline{I(\theta)}\right)\right] && (5)\\
\text{s.t.} \quad & \overline{I(\theta)} = \max\left\{0, a\hat{\ell}(\theta) + b\right\} \\
& \underline{I(\theta)} = \min\left\{a\hat{\ell}(\theta) + b, 1\right\} \\
& \pi = (1 - c_\kappa)\,\mathbb{E}\left[\overline{I(\theta)}\right] + c_\kappa \mathrm{CVaR}_{1-\epsilon_K}\left(\overline{I(\theta)}\right) && (6)\\
& \pi \leq \overline{\pi}.
\end{aligned}
$$

Multiple Zone Case

In the multi-zone case, the problems are coupled through the required capital

$$K = \mathrm{CVaR}_{1-\epsilon}\left(\sum_z s_z I_z(\hat{\ell}_z)\right) - \mathbb{E}\left[\sum_z s_z I_z(\hat{\ell}_z)\right]$$

The first term will be affected by the correlation of losses across zones, so the optimal contracts will vary depending on this correlation.

Multiple Zone Model

$$\max_{a,b,K,\pi} \quad \mathbb{E}\left[\sum_z U\left(w_{0,z} - \ell_z^j - \pi_z + I_z(\theta_z^j)\right)\right] \tag{7}$$

$$\text{s.t.} \quad \pi_z = \mathbb{E}\left[\overline{I_z(\theta_z)}\right] + \frac{c_\kappa}{\sum_z s_z} K \tag{8}$$

$$K = \text{CVaR}_{1-\epsilon_K}\left(\sum_z s_z \overline{I_z(\theta_z)}\right) - \mathbb{E}\left[\sum_z s_z \underline{I_z(\theta_{z'})}\right] \tag{9}$$

$$\overline{I_z(\theta_z)} = \max\left\{0, a_z \hat{\ell}_z(\theta_z) + b_z\right\}$$

$$\underline{I_z(\theta_z)} = \min\left\{a_z \hat{\ell}_z(\theta_z) + b_z, 1\right\}$$

$$\pi_z \leq \overline{\pi_z}.$$

Content

Evaluation: Thai Data

- Index isurance is especially popular in developing countries beacuse of its low cost.
- However, one of the difficulties in studying index insurance is the lack of data in these settings.
- I collaborated with the Bank of Thailand and was given access to a novel dataset of census-tract-level (Tambon) losses from natural disasters.

Thai Data: Administrative Data Sources

- Thailand Department of Agricultural Extension (DOAE): This dataset contains detailed information on the planting activities of over 3.1 million Thai farmers. Information includes plant date, plot size, rice variety, etc
- Thai Governemnt Disaster Relief Data: This data is from the current national disaster relief program. It contains information about the disaster type, losses, and disaster date.

Remote Sensing Data Sources

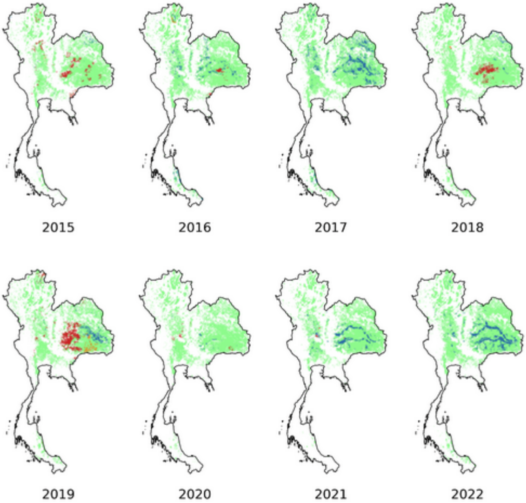
- Climate Hazards Center InfraRed Precipitation with Station data (CHIRPS): daily rainfall data 1981-present
- MOD21A1 Land Surface Temperature: daily temperature data, available 2000-present
- FLDAS Famine Early Warning Systems Network: monthly evapotranspiration data, available 1992-present

Data Summary Statistics

Characteristic	Overall N = 49,052[1]	2015 N = 6,037[1]	2016 N = 6,358[1]	2017 N = 6,285[1]	2018 N = 6,288[1]	2019 N = 6,268[1]	2020 N = 6,079[1]	2021 N = 5,510[1]	2022 N = 6,227[1]
Loss	0.02 (0.08)	0.00 (0.03)	0.02 (0.08)	0.04 (0.11)	0.01 (0.05)	0.01 (0.07)	0.00 (0.03)	0.03 (0.10)	0.03 (0.11)
Zone									
C1	6,052 (12%)	743 (12%)	763 (12%)	757 (12%)	761 (12%)	762 (12%)	766 (13%)	737 (13%)	763 (12%)
C2	3,064 (6.2%)	366 (6.1%)	404 (6.4%)	394 (6.3%)	393 (6.3%)	396 (6.3%)	391 (6.4%)	329 (6.0%)	391 (6.3%)
C3	3,139 (6.4%)	392 (6.5%)	426 (6.7%)	405 (6.4%)	410 (6.5%)	410 (6.5%)	398 (6.5%)	294 (5.3%)	404 (6.5%)
N1	6,312 (13%)	784 (13%)	808 (13%)	805 (13%)	803 (13%)	798 (13%)	798 (13%)	718 (13%)	798 (13%)
N2	3,593 (7.3%)	451 (7.5%)	451 (7.1%)	454 (7.2%)	449 (7.1%)	451 (7.2%)	450 (7.4%)	437 (7.9%)	450 (7.2%)
N3	2,105 (4.3%)	269 (4.5%)	268 (4.2%)	268 (4.3%)	265 (4.2%)	264 (4.2%)	265 (4.4%)	241 (4.4%)	265 (4.3%)
NE1	1,294 (2.6%)	162 (2.7%)	162 (2.5%)	162 (2.6%)	162 (2.6%)	162 (2.6%)	162 (2.7%)	160 (2.9%)	162 (2.6%)
NE2	9,400 (19%)	1,178 (20%)	1,181 (19%)	1,183 (19%)	1,183 (19%)	1,179 (19%)	1,178 (19%)	1,138 (21%)	1,180 (19%)
NE3	10,546 (21%)	1,317 (22%)	1,320 (21%)	1,319 (21%)	1,319 (21%)	1,320 (21%)	1,319 (22%)	1,315 (24%)	1,317 (21%)
S1	1,025 (2.1%)	111 (1.8%)	145 (2.3%)	141 (2.2%)	140 (2.2%)	148 (2.4%)	127 (2.1%)	76 (1.4%)	137 (2.2%)
S2	2,522 (5.1%)	264 (4.4%)	430 (6.8%)	397 (6.3%)	403 (6.4%)	378 (6.0%)	225 (3.7%)	65 (1.2%)	360 (5.8%)

[1] Mean (SD); n (%)

Disaster Activity Map



Evaluation Procedure

- We use leave one year out cross validation.
- In each CV fold, data from year y is used for testing. Data from $y' \neq y$ is used for insurance design (prediction model training and contract design).
- This gives us out of sample predictions for every observation in our data. We then compute utility-based performance metrics across all out of sample predictions.

Evaluation Details

- We use training data to compute premium $\rightarrow$ this reflects practice.
- We modify Chen's method to use the same premium definition as ours.
- We use a copula approach to estimate the CVaR when calculating the premium.
- Same method is used to calculate the premium for all methods.
- We use a CRRA utility function with $\alpha = 1.5$.

Perfromance Metrics

We focus on the Relative Insurance Benefit (RIB) (Kenduiywo et al. (2021)) as a performance metric. It measures the share of all possible welfare gains that are recovered by our insurance.

- Relative Insurance Benefit (RIB): $\frac{CE^J - CE^{NI}}{CE^{PI} - CE^{NI}}$
- Certainty equivalent (CE): $u^{-1}(\mathbb{E}[u(w^I)])$
- Utility: $u(w) = \frac{1}{1-\alpha} w^{\frac{1}{1-\alpha}}$

Results for $c_k = 0.02, \alpha = 1.5$

Here, the contract design and pricing for each zone happens independently of all other zones.

Table: Single Zone Results

Method	**RIB**	**Premium**	**Cost II**	**Cost PI**	**CapShare**
VMX	0.874	0.022	0.028	0.030	0.125
Chantarat	-0.249	0.005	0.003	0.030	0.227
Chen	-8.797	0.096	0.061	0.030	0.257

Single Zone Results Key Points

- Our method achieves 67% of the utility gains of perfect insurance.
- Has a lower percentage of the premium cost going to required capital.

Note on the value of c_κ

- Potential value for farmers is highly dependent on c_κ and risk aversion, α
- For the most commonly used value of α, farmers could only benefit from perfect insurane if $c_\kappa = 0.02$.
- As risk aversion increases, the gains from perfect insurance increase.

Effects of Cost of Capital and Risk Aversion

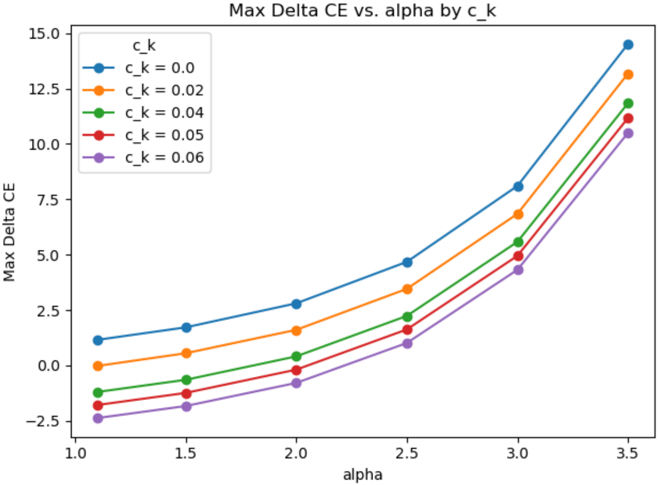


Figure: Perfect Insurance

Multi zone results

- Multi zone model performs worse than just optimizing contract for each zone independently :(

Content

Data

Used two main data sources

- Illinois annual corn yield data from the National Agricultural Statistics Service (NASS). Data is available at the county level from 1925-2022. 84 counties.
- Weather data from the PRISM climate group. Has monthly data on several weather variables (temperature, precipitation, etc). Available 1895-present.

Evaluation Procedure

- We use a 70/15/15 train/val/test split. Data is kept in chronological order. Training data has older years and test data has the newest years.
- We modified Chen's method to use the same definition of the premium as our method.
- We use the training and validation data to design the contracts using both methods, apply the contracts to farmers in the test set, and compute performance metrics.
- We used a data shortening exercise to evaluate how the performance of both methods changed as more data became available.

Overview

- Our method performs similarly or outperforms the Chen model when there is less than 50 years of data available.
- In terms of farmer utility, our method tends to work better with realistic data lengths. Most satellite data starts at 1980 at the earliest.
- When using data from other states, our method consistently outperforms Chen's method, but is not always better than no insurance, at least at the full premium price. This might not be a huge problem, since agricultural insurance tends to be heavily subsidized, both in rich and poor countries. In the US, it averaged 62% of premiums in 2022.

Illinois Short Data (Less than 40 years)

Method	DeltaCE	Premium	Required Capital
Chantarat	4.13	113.96	158.73
Chen uc	4.87	60.69	134.35
Our Method	5.31	62.49	97.91

Illinois Long Data (More than 40 years)

Method	DeltaCE	Premium	Required Capital
Chantarat	2.05	105.53	135.72
Chen uc	4.82	50.06	115.02
Our Method	3.82	41.75	67.80

Overview

- Our multiple zone model adjusts contracts based on the correlation between the insured zones. In this case, it leads to contracts that pay out more frequently, but at a lower rate. This reduces the tail risk for the insurer and reduces the amount of capital needed.
- It outperforms Chen's model and the no insurance case consistently, and has lower costs and required capital than the Chen model.
- We also wanted to compare it to using our single zone model. In the following figures, Our Method: SZ refers to using our single zone method to design the contract of each state individually, but then calculating the premium as if it was a part of the portfolio.

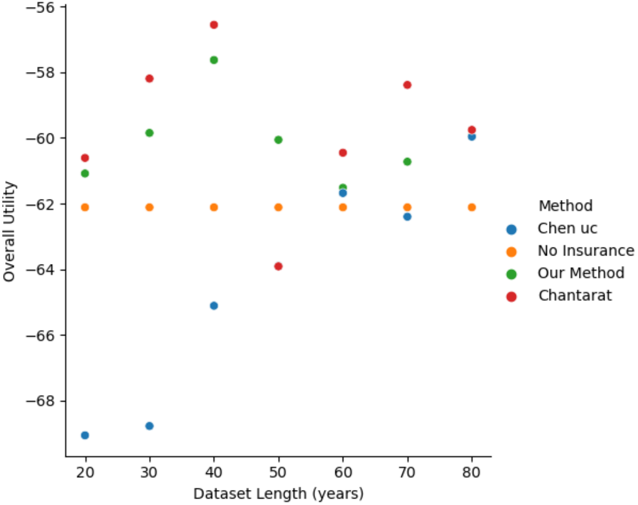
Midwest Short Data Results

Method	DeltaU	Insurer Cost	Required Capital
Chantarat	5.90	397,474.47	17720.74
Chen uc	-8.92	159,928.89	19865.41
Our Method	4.18	150,414.21	9982.12

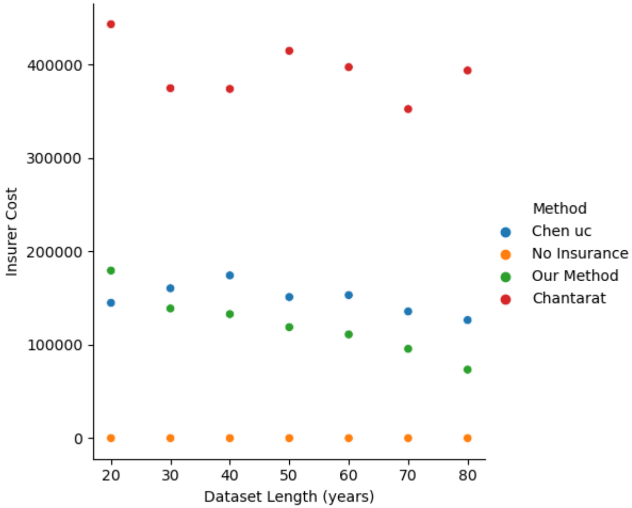
Midwest Long Data Results

Method	DeltaU	Insurer Cost	Required Capital
Chantarat	2.40	389,678.99	16874.15
Chen uc	0.21	141651.90	12142.11
Our Method	2.50	99,769.84	5559.64

Midwest: Overall Utility



Midwest: Insurer Cost



Things I was thinking about adding

- Deep dive as to why my method does better, and when it does better?
- Run the same evaluation but with shuffled data instead of having it orderd.
- I wanted to do more of a deep dive on the benefits of taking a portfolio approach, but not sure how beyond just looking at capital requirements. Maybe using an expected return and variance approach?
- A robustness check on the stability of the solutions at different dataset lengths.

Questions

- What else should I report?
- How can I strengthen this evaluation?
- Any robustness checks I should add?
- I feel like I should do a deep dive as to why the method does

Thai Data

- Plot-level losses caused by natural disasters between 2015-2022.
- I can access the raw data, and they can send me aggregate versions as well. There are around 7000 counties in Thailand, and around 80,000 villages, the data can be aggregated at both of those levels.

Loss Definition

From what I can tell from the replication files, they seem to define loss in every year as:

$$\ell_{st} = R_s^* - R_{st}$$

where R_s^* is the maximum revenue observed in state s across all time periods, and R_{st} is the revenue in state s at time t.

Initial wealth

According to the paper, they set $w_0 = 389$. However, in the replication files, they set it to be $w_0 = 813 - 504 + 389$. According to the comments, 504 is the fixed cost of operating a farm, and there are no comments regarding the 813, I'm assuming it corresponds to R^*.

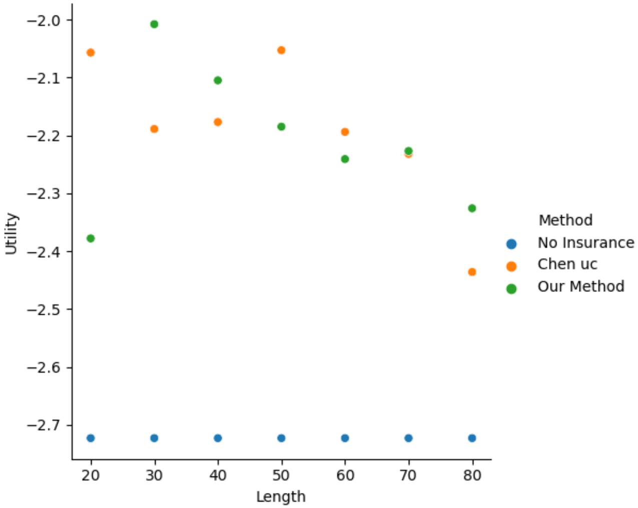
Detrending

- According to the paper, they detrend the county level yield data using a 2nd order polynomial fit with a "robust" regression method, but they don't specify what they use, and it's not in the replication files. They also don't specify if they remove the trend using additive or multiplicative decomposition model. Using an additive decomposition model yielded the most similar losses to what they provide in the replication files.
- There are a couple of papers showing that using locally weighted regression to detrend works better.

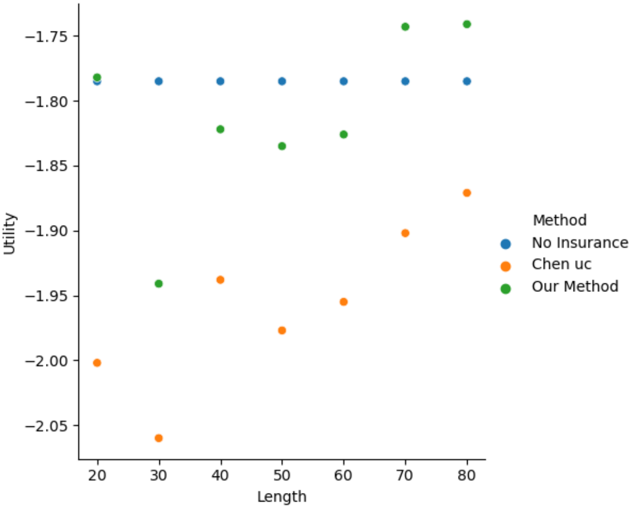
Questions

- Would it make sense to define loss as deviation from historical average? Allowing it to be positive in some years? In other words, we would first adjust all of the yield data to 2020 levels and then calculate the historical average. The loss in each year would be the deviation from this historical average.
- Should I simply follow their lead on detrending? Or should I try to improve on it?
- Do you think it's necessary to show results with both definitions of the premium?
- Do you think subsidy vs lump sum results would be interesting?

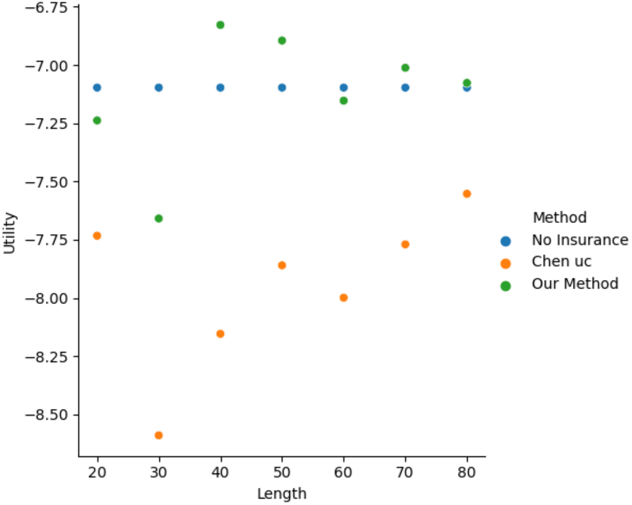
Their Defn of Premium: Illinois Utility



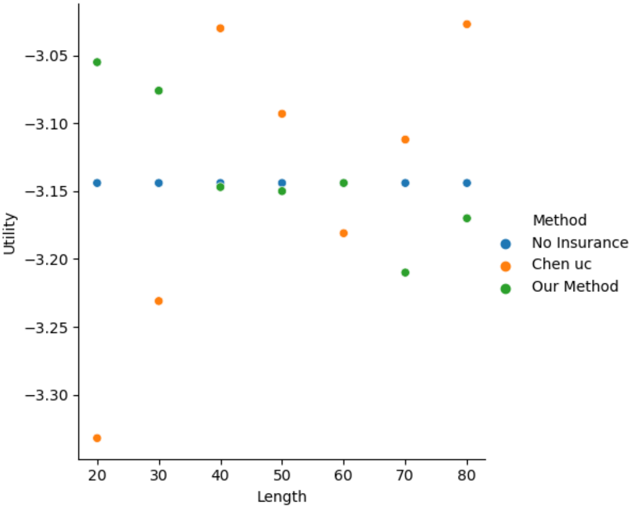
Their Defn of Premium: Iowa Utility



Their Defn of Premium: Missouri Utility



Their Defn of Premium: Indiana Utility



References

Chantarat, Sommarat et al. (2013). "Designing index-based livestock insurance for managing asset risk in northern Kenya". In: *Journal of Risk and Insurance* 80(1), pp. 205–237.

Mapfumo, Shadreck, Huybert Groenendaal, and Chloe Dugger (2017). *Risk modeling for appraising named peril index insurance products: A Guide for practitioners*. World Bank Publications.

Kenduiywo, Benson K et al. (2021). "Evaluating the quality of remote sensing products for agricultural index insurance". In: *PLoS One* 16(10), e0258215.

Chen, Zhanhui et al. (2023). "Managing weather risk with a neural network-based index insurance". In: *Management Science*.

Idealized CVaR Model

- **Objective:** conditional value at risk of the farmers' loss net of insurance.
- **Constraint 1:** piecewise linear structure of the contract.
- **Constraint 2:** budget constraint.
- **Constraint 3:** definition of required capital.

$$\min_{a,b,\pi,K} \quad CVaR_{1-\epsilon}\left(\ell - I(\theta)\right)$$

$$\text{s.t. } I(\theta) = \min\{(a\hat{\ell}(\theta) + b)^{+}, P\} \tag{10}$$

$$\mathbb{E}\left[I(\theta)\right] + c_{\kappa} K \leq B \tag{11}$$

$$K = \left(CVaR_{1-\epsilon}\left(I(\theta)\right) - \mathbb{E}[I(\theta)]\right) \tag{12}$$

The problem is non-convex, so we need convex approximations

We use the following approximations of $I(\theta)$ to make the problem convex:

$$\overline{I(\theta)} \triangleq \max\left\{0, a\hat{\ell}(\theta) + b\right\}$$
$$\underline{I(\theta)} \triangleq \min\{a\hat{\ell}(\theta) + b, K\}$$

- Note that $\overline{I(\theta)} \geq I(\theta)$ and $\overline{I(\theta)}$ is convex. Conversely, $\underline{I(\theta)} \leq I(\theta)$ and $\underline{I(\theta)}$ is concave.
- We replace $I(\theta)$ with either $\overline{I(\theta)}$ or $\underline{I(\theta)}$ where necessary to obtain conservative and convex approximations.
- We also need approximations or proxies for $E[I(\theta)]$ in constraint . We use $\pi_{SQ} = E[I_{SQ}(\theta)]$, where I_{SQ} is the contract designed using the status quo method, as a proxy for $E[I(\theta)]$ in constraint .