A hierarchical Bayes model for rainfall with applications to index insurance

Kenneth Shirley, Daniel Osgood, and Andrew Robertson

The Earth Institute at Columbia University, New York, NY, USA

OVERVIEW

In order to efficiently price rainfall-based index insurance contracts, a statistical model for daily rainfall time series is needed. We fit a hierarchical Bayes model to a daily rainfall time series from a single weather station in Lilongwe, Malawi, using a Markov model for the occurrence of rain and a gamma distribution for the intensity of rain. We use posterior predictive checks to evaluate the fit of the model and find that it fits well with respect to most statistics; one exception is the variance of annual rainfall, which is underpredicted. A secondary problem we mention is how to design a payout function that minimizes the variance of annual income, subject to certain assumptions.

INTRODUCTION TO INDEX INSURANCE

Rainfall-based index insurance is a financial instrument that is being used to insure small farmers against drought. It is an exciting new tool that may help poor people in subsistence economies stabilize their annual income, begin to save money, and eventually escape a poverty trap.

Index insurance is a simple but powerful concept: instead of insuring farmers against crop failure, insure them against one or more risks of crop failure, such as drought. Because index insurance provides payouts in the event of drought rather than crop failure, two classic problems with traditional insurance - moral hazard and adverse selection - are avoided. First, farmers aren't incentivized to let their crops fail in order to get an insurance payout (moral hazard), thus they will try to maximize their yield in all circumstances. Second, all farmers within a given region are equally exposed to the risk of drought, thus insurance isn't being bought disproportionately by high-risk individuals (adverse selection).

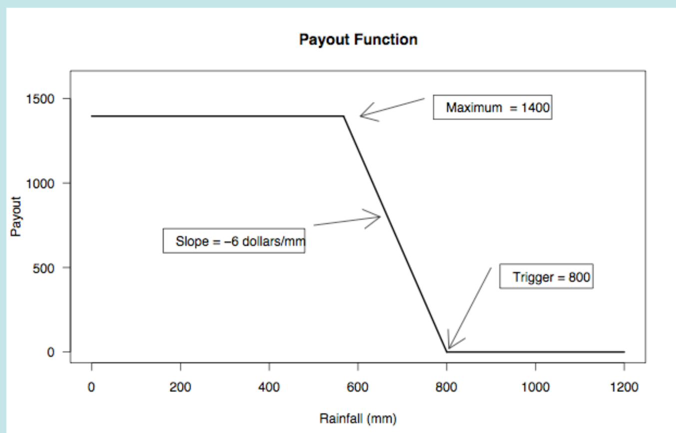
CONTRACT DESIGN AND PRICING

The price of an insurance contract is:

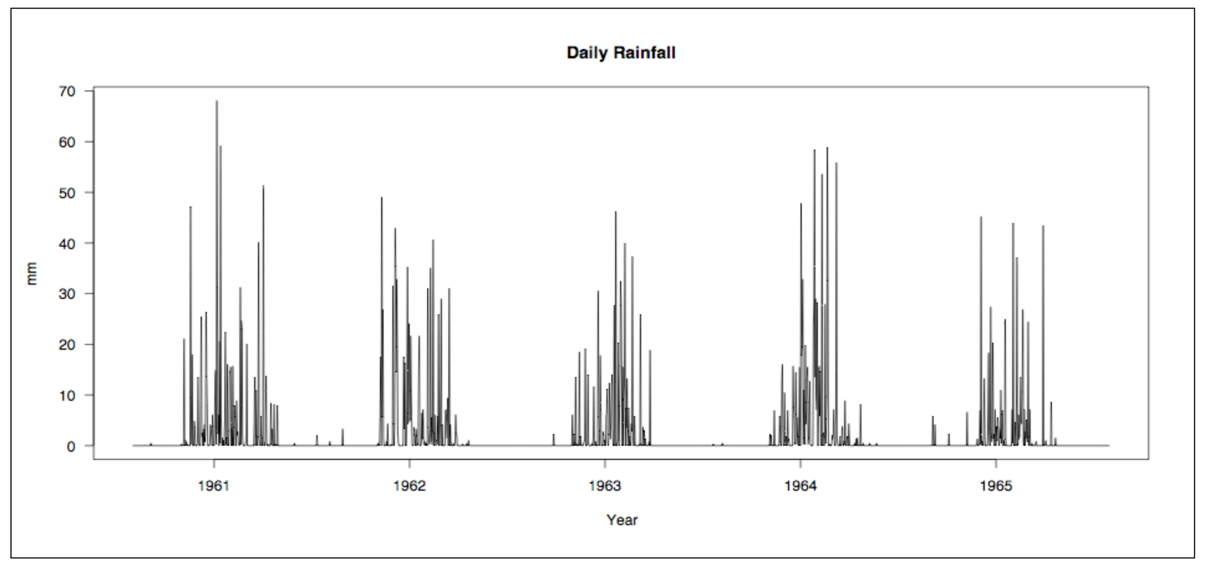
Price = Average Payout + Loading*(Value at Risk - Average Payout)

where the second term on the right side of the equation is the premium. The Value at Risk is the amount of money that the insurance company needs to hold in order to cover a 99th percentile event, i.e. a huge drought, and the loading is set roughly to the interest rate that the insurance company could charge if they were able to lend their money, instead of holding onto it (6% for example).

The payout is a linear, decreasing, deterministic function of the rainfall, and in a simple contract, there is one payout function per growing season. The key statistical problem is to estimate the mean and 99th percentile of the payout distribution (for a fixed contract design). Current practitioners use the empirical distribution for these estimates, which is especially bad when estimating the 99th percentile with as few as 10 - 20 years of data. Technically, one only needs to model rainfall on the same time scale as the contract, but it is customary to model daily rainfall, as daily time series of rainfall are often used in crop yield simulations.



The secondary problem of optimizing the payout function to minimize the variance of annual income is discussed in the lower right panel. This involves tuning the 3 parameters (trigger, maximum, and slope) in the figure to the left.



THE DATA

The daily rainfall time series we use for this model fit is from Lilongwe, Malawi, from the years 1961 - 2005 [See above figures]. The weather station there receives an average of 850 mm of rain per year, with a standard deviation of 180 mm. There is one wet season, extending roughly from November through March. There is no evidence of a long-term trend in rainfall frequency or intensity.

Let $Y_{t[j]}$ denote the amount of rainfall (mm) and let $Z_{t[j]} = 1\{Y_{t[j]} > 0\}$, the indicator of a wet day, on day t, in year j, for $t = 1, .., 16709$, and $j = 1, ..., 45$. (The nesting structure is inconvenient because of leap years).

The model for the occurrence of rainfall is:

$$\text{logit}\Big(p(Z_t = 1 \mid Z_{t-1} = i)\Big) = \beta_{i0} + \beta_{i1}\cos(2\pi t/365.25) + \beta_{i2}\sin(2\pi t/365.25) + \alpha_{it[j]},$$

$$\alpha_{ij} \sim N(0, \sigma_i^2),$$

$$\beta_{ik} \sim N(0, \infty),$$

$$p(\sigma_i) \propto 1,$$

for wet/dry previous-day states $i = 0, 1$, days $t = 1, ..., 16709$, and years $j = 1, .., 45$. The α_j's act as random effects for each year.

The model for the intensity of rainfall is as follows:

$$Y_t \sim \text{gamma}(\mu_t, s_t),$$

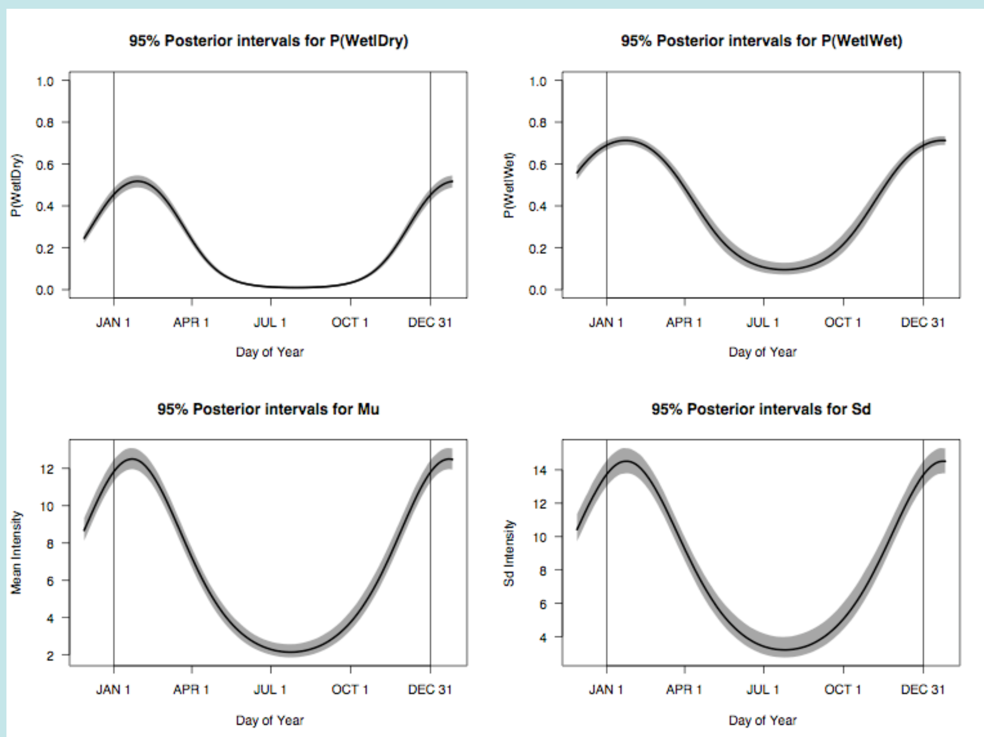
$$\log(\mu_t) = \eta_{10} + \eta_{11}\cos(2\pi t/365.25) + \eta_{12}\sin(2\pi t/365.25) + \gamma_{1t[j]},$$

$$\log(s_t) = \eta_{20} + \eta_{21}\cos(2\pi t/365.25) + \eta_{22}\sin(2\pi t/365.25) + \gamma_{2t[j]},$$

$$\gamma_{ij} \sim N(0, \tau_i^2),$$

$$\eta_{ik} \sim N(0, \infty),$$

$$p(\tau_i) \propto 1.$$



THE MODEL

We fit a Markov model for the occurrence of rain on a given day, and a gamma distribution for the intensity of rainfall on wet days. This is consistent with the basic models introduced in Richardson & Wright (1984). Alternatives to this model include higher-order Markov chains for the occurrence of rain (Stern & Coe, 1984), and mixtures of exponential distributions for the intensity of rain (Hansen & Mavromatis, 2001).

THE MODEL FIT

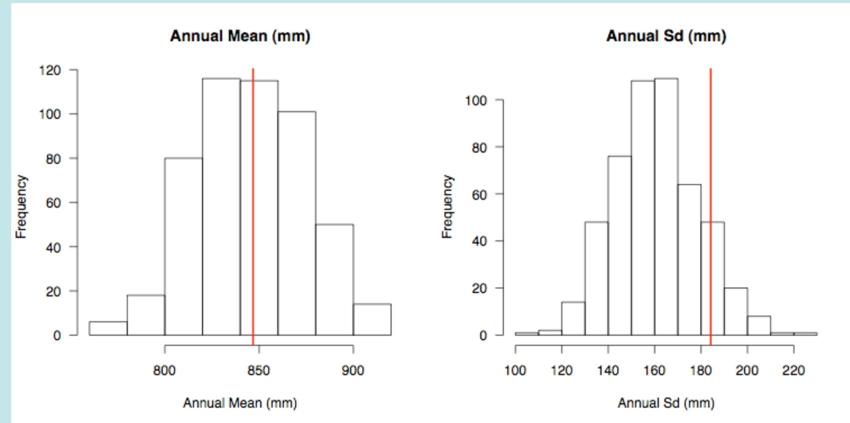
The model was fit using R. Three chains were run from random starting points for 5,000 iterations; convergence was very fast for the β parameters (coefficients associated with the Markov model), and slightly slower for the η parameters (coefficients associated with the gamma distributions), but all parameters converged within about 2,000 - 3,000 iterations.

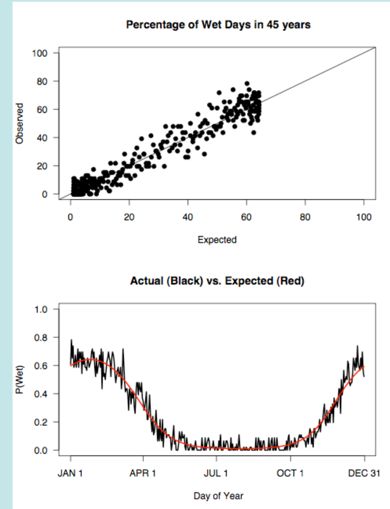
The group-level variance parameters τ_i shrunk to zero, so the year-level model was eliminated from the intensity model (i.e. the γ_{ij}'s were all set to zero).

The fit of the model is displayed in the figure to the left. The four panels contain posterior intervals throughout the year for the transition probabilities (top two panels) and the mean and sd of the gamma distribution (bottom two panels).

POSTERIOR PREDICTIVE CHECKS

Using the posterior samples, we simulated 500 copies of the data set (each one consists of 45 years of daily rainfall) and compared statistics from the simulations to the observed statistics. The model captured the annual mean well, but underpredicted the interannual variability (a well-known problem for many weather generators, Mavromatis & Hansen, 2001). The model did well capturing the occurrence of rain through the year.





DISCUSSION

The fit of the model allows us to calculate some interesting statistics:

1. The wettest day of the year (in expectation) in January 22nd. There is about a 65% chance of rain that day, with an expected intensity of 12.5 mm.
2. *The price of uncertainty*. If we subsample 5 or 10 years out of the 45 year time series, fit the model, and simulate data, the simulated data has fatter tails, due to the fact that there was less data to train the model. Larger standard errors for the parameters lead to more extreme values in the rainfall simulations (including droughts), which leads to a higher average payout (because of the truncation at a payout of zero) and a higher estimated 99th percentile of payouts. When using only 5 or 10 years of data, compared to 45 years, the price of insurance increased by 22% and 15%, respectively. This is especially relevant when designing contracts for sites with short rainfall records.

RELATED PROBLEMS & CONCLUSIONS

The second statistical problem with index insurance has to do with the design of contracts, with the goal of minimizing the variance of a farmer's annual income and maximizing its expected value. Under a few assumptions (a bivariate normal model for annual rainfall and yield), we can solve for the expected value and variance of annual income in quasi-closed form (involving some normal CDF's). We then see a classic tradeoff between maximizing expected income and minimizing its variance, and a function of these two can be optimized.

A variety of extensions to the rainfall model presented here are available and will be pursued. Among them are including cosine curves with frequencies other than 1/365 in the model, incorporating exogenous variables such as ENSO state into the model, and modeling higher-order dependence in the frequency model.

Innovations in this work include modeling the standard deviation of the gamma distribution using a cosine function, and fitting a fully Bayesian model for this type of data.

The presenter would like to thank his coauthors for their help, and Andrew Gelman and Paul Block for their comments.