On-line appendix to "Satellite Data and Machine Learning for Weather Risk Management and Food Security"

Enrico Biffis* Erik Chavez†

June 1, 2017

Abstract

This Appendix provides additional technical details for the analysis contained in Biffis and Chavez (2017).

1 Model setup: Details and extensions

In this section we provide a more rigorous formalization of the model discussed in Biffis and Chavez (2017, section 3). Different production sites in the region of interest[1] are mapped in terms of pixels with centroid coordinates given by the pair (i, j), where $i \in [1, \dots, L]$ denotes longitude and $j \in [1, l]$ denotes latitude. A fraction $F^c_{i,j} \in [0, 1]$ of land in pixel (i, j) is planted with crop of type $c \in \mathcal{C} = \{c_1, \dots, c_n\}$, where $\mathcal{C}$ denotes the set of agricultural products. We denote by $A^c_{i,j} = F^c_{i,j} H_{i,j}$ the area planted with agricultural commodity c, where $H_{i,j}$ is the size of pixel (i, j) expressed in hectares. Farmers plant seeds at time zero, and collect harvest at time one (end of the harvesting season).[2] The harvest can be sold by the producers in the spot market at price $P^c > 0$, or at a forward price $\overline{P}^c > 0$ agreed at

*Biffis (e.biffis@imperial.ac.uk) is at the Department of Finance, Imperial College Business School, Imperial College London, South Kensington Campus, London SW7 2AZ, United Kingdom.

†Chavez (erik.chavez07@imperial.ac.uk) is at the Centre for Environmental Policy, Imperial College London, South Kensington Campus, London SW7 1NA, United Kingdom.

[1] Country-wide production sites in the Mozambique case study of Biffis and Chavez (2017).

[2] This is for ease of exposition of the model. The numerical examples of the Mozambique case study take into account actual sowing and harvesting dates.

time zero. In general, we should allow the spot prices to depend on the realization of the collection of weather indices $W = \{W_{i,j}\}$, and write $P^x(W)$. For each location (i,j), the vector[3] $W_{i,j} = (W^1_{i,j}, \ldots, W^k_{i,j})^\mathsf{T}$ provides the realizations of k weather indices between time zero and time one.

The producer can employ different production technologies in each pixel (i,j). Possible examples include the following:

- **Irrigation-based vs. rainfed production**. A fraction $\alpha^c_{i,j} \in [0,1]$ of area $A^c_{i,j}$ relies on irrigation to grow agricultural commodity $c \in \mathcal{C}$ in pixel (i,j). The cost per hectare of deploying irrigation is denoted by $C^I_{i,j} \geq 0$.

- **Crop varieties**. Each crop c can be sown with different crop varieties, depending on how long they take to reach maturity. The idea is that some short cycle varieties may aovid potential exposure to drought or heat wave stresses, and that long cycle varieties may better smooth shocks over time. For example, short cycle varieties of corn can grow in under 80 days, whereas long cycle varieties can reach maturity in over 150 days. We denote by $v^{c,h}_{i,j}$ (with $h = 1, \ldots, m_c$) the fraction of land in pixel (i,j) planted with crop variety h for product c. We denote by $(S^c_{i,j}, N^c_{i,j}, L^c_{i,j})$ the fractions of land in pixel (i,j) planted with short, normal average, and long cycle varieties of product c. The cost per hectare of using different seed varieties for c is encoded in the positive real vector $(C^c_S,\, C^c_N,\, C^c_L)$.

- **Sowing dates**. Sowing dates can be chosen to limit the exposure of critical phases of crop growth to periods associated with high probability of erratic rainfalls or peak temperatures. Using historical average sowing dates as the baseline (zero delay), we denote by $d^c_{i,j}(l)$ the fraction of land in pixel (i,j) sown according to a (positive or negative) delay $l \in [\underline{d}^c, \overline{d}^c]$ relative to the historical average, with the constraint that $\int_{\underline{d}^c}^{\overline{d}^c} d^c_{i,j}(u)\mathrm{d}u = 1$. Examples of sowing strategies include the following: i) a baseline

[3]Here and in Biffis and Chavez (2017), all vectors are column vectors. For $v \in \mathbb{R}^n$, v^T denotes the transpose of vector v.

case where sowing is based on historical average sowing dates, ii) a scenario where sowing is anticipated by ten days relative to the baseline, and iii) a scenario where sowing is postponed by ten days relative to the baseline. We then denote by $D^{c,0}_{i,j}$, $D^{c,-}_{i,j} = D^{c,0}_{i,j} + \underline{d}^c$, and $D^{c,+}_{i,j} = D^{c,0}_{i,j} + \overline{d}^c$ the fraction of land in pixel (i,j) sown according to strategies i), ii) and iii), respectively, requiring in turn that $D^{c,0}_{i,j} + D^{c,-}_{i,j} + D^{c,+}_{i,j} = F^c_{i,j}$.

We denote by $u = (u^c_{i,j})_{c,i,j}$ the family of production configurations available to the producer, with $u^c_{i,j} \in \mathcal{U}^c_{i,j}$. The set of admissible configurations, $\mathcal{U}^c_{i,j}$, for the examples illustrated above takes the form $[0,1] \times [0,1] \times \{v^{c,1}_{i,j}, \ldots, v^{x,m^c}_{i,j}\} \times [\underline{d}^c, \overline{d}^c]$.

The total cost of production of commodity c in pixel (i,j) is denoted $K^c_{i,j}$. Assuming no post-harvest loss (i.e., any potential losses affect the harvesting season only), total gross revenue is given by

$$R_{i,j}(u,W) = \sum_{c \in \mathcal{C}} Y^c_{i,j}(u,W) P^c(W)$$

It is important to note that although spot prices are allowed to depend on W, they could be taken as exogenously given if production across locations is marginal relative to the global market. This simplification (which is adopted in Biffis and Chavez, 2017, section 5), however, does not prevent spot prices from being correlated across agricultural products. The total net revenue to the agent is finally given by

$$R(u,W) - K(u) = \sum_{i,j} \sum_{c \in \mathcal{C}} \left(R^c_{i,j}(u,W) - K^c_{i,j}(u) \right),$$

with obvious meaning of notation, where each production cost at location, $K^c_{i,j}(u)$, is given by

$$K^c_{i,j}(u) = A^x_{i,j} \left(C_I \alpha^c_{i,j} + S^c_{i,j} C^c_S + N^c_{i,j} C^c_N + L^c_{i,j} C^c_L \right).$$

2 Data mining and weather index design

The machine learning methodology discussed in Section 4 and used in Section 5 of Biffis and Chavez (2017) relies on a non-parametric regression approach known as recursive partitioning. Recursive partitioning is used to construct classification and regression trees where groups of response parameters are successively separated based on their similar response values. Contrary to linear regression models, tree partitioning allows one to capture non-linear and high order interactions (Breiman, 2001; Friedman et al., 2001).

The sequential splitting of a group of response variables is generally carried out by means of impurity reduction – where impurity refers to a measure of variable response similarity. Minimum impurity or entropy in a dataset is achieved when the relative frequency of one of the response classes is zero, while maximum impurity is attained when there is the same relative frequency for two response classes. Hence in each node tree construction, the first split carried out is the one generating the highest impurity reduction, where the variable selected for splitting (the weather index in our case) is the most strongly associated with the response variable (the crop yield in our case). For our purposes we use an ensemble-based recursive partitioning methodology known as Random Forest (RF) (see Breiman, 2001). In the following we first introduce regression trees, and then outline the use of RF for variable selection.

Two types of impurity measures are usualy employed for split selection: i) entropy measures such as the Gini Index or the Shannon entropy; ii) p-values of association tests with the response variable. Each successive node is split where the highest entropy level is detected or lowest p-value determined. The analysis of Biffis and Chavez (2017) relies on the second splitting criterion.

For all weather index realizations $z^k_{i,j}$, with $k = 1, ..., n$ (and $n = 25$ in the application discussed in the Mozambique case study), at each node we compute $\hat{c}_1$ and $\hat{c}_2$ as the mean of observations below and above the candidate split level t. The optimal level to split the tree is found by considering the residual sum of squares as impurity measure $I^k_{i,j}$ for each pixel

as follows:

$$I_{i,j}^{k} = \sum_{z_{i,j}^{k} \in \mathcal{Z}_t^-} \left(z_{i,j}^{k} - \hat{c}_1\right)^2 + \sum_{z_{i,j}^{k} \in \mathcal{Z}_t^+} \left(z_{i,j}^{k} - \hat{c}_2\right)^2 .$$

In the above, the $z_{i,j}^{k}$'s are realizations of weather variables capturing rainfall or maximum temperature variability during a specific period of the crop growth within pixel (i, j). For each threshold level t, the sets $\mathcal{Z}_t^-$ and $\mathcal{Z}_t^+$ collect realizations below and above the threshold, respectively. The pair of (t, k) that minimizes the impurity reduction at each node is used to determine the binary split. Figure 4 in Biffis and Chavez (2017) illustrates the different periods of crop growth captured by each pixel-level weather variable $z_{i,j}^{k}$. The computation of each $z_{i,j}^{k}$ is carried out either by computing cumulative daily rainfall or by counting the number of days for which the maximum daily temperature exceeds a critical threshold during the portion k of crop growth considered.

Parametric and non-parametric approaches to variable selection. Several classical parametric statistical techniques have been developed to assess the importance of the effect of predictor variables on quantities of interest. Most popular techniques include linear and logistic regressions. Given the quantitative and qualitative features of weather indices time series, however, parametric techniques for variable importance selection suffer from important limitations. First, the number of weather indices tested is large relative to the number of dependent variables (the commodity yield time series in our case), making the performance of standard parametric approaches inadequate in the presence of complex interactions and high order effects. Second, parametric methods limit the variety of association patterns that can be captured. In principle one could address some of these issues with dimension reduction techniques such as Principal Component Analysis before applying parametric regression models, but the use of projected variables (into a reduced variable coordinates space) makes the approach unappealing for our applications, for which preserving variable units is important.

Recursive partitioning and variable importance assessment Recursive partitioning RF techniques are applied in different fields confronted with the issue of selecting important variables within a high number of variables. RF-based predictor variable selection is, for instance, used in genetics where a high number of genes can be involved in the expression of a pathology, but a specific gene needs to be identified to be targeted in the drug molecule design phase (Segal et al., 2004). The selection of variables through RF is achieved by computing a variable importance measure that quantifies the predictiveness of the different variables from an initial sample. Two variable importance measures are used to assess the predictive power of a variable: i) "splitting improvement" and ii) "permutation accuracy importance". Splitting improvement is a measure that sums up the total average impurity reduction measured, for instance, by a Gini index gain, that a variable achieves in all of its positions in the forest. This variable importance measure is not used in the present study due to its potential bias (e.g., Strobl et al., 2007b), and therefore is not detailed here. We will focus instead on permutation accuracy, which is computed as follows.

For each tree, so called Out-of-Bag (OOB) data is run down each tree by making a random left-right split assignment when reaching a node splitting at a variable whose importance is being assessed. The OOB ensemble is then computed. Dropping dependence on location for ease of notation, we have that if Z^j is an important variable, the distance of the terminal node for the data point from the original terminal node assignment position increases. The OOB ensemble is computed and the prediction error is computed. A measure of importance of weather variable Z^j, $VI(Z^j)$, is the difference between the OOB prediction error, PE_j, for the noisy variable, and the original ensemble prediction error, PE: $VI(Z^j) = PE_j - PE$.

In case a given weather variable is not associated with the yield response variable, it is either not included in the tree as it is assigned a zero importance value, or is included in the tree by chance. Thus, if a given weather variable's prediction importance is close to zero, it was either not associated with the yield response or was included in the tree by chance only. If included by chance, the random permutation of this variable will only lead to a small

increase in the prediction error. Therefore, the permutation accuracy variable importance selection method can also generate small negative importance values. In case a preliminary screening of weather variables is considered, all weather indices with variable importance scores that are positive and higher than the largest absolute value of negative scores should be considered. In contrast, the weather index variable with the highest positive variable importance score is also the one showing the highest association with the yield response variable.

For a variable Z^j of an OOB sample $\mathcal{O}^{(t)}$ for a tree t, with $t \in [1, ntree]$, the variable importance is defined as

$$VI^t\left(Z^j\right) = \frac{\sum_{i \in \mathcal{O}^{(t)}} I\left(\gamma_i = \gamma_i^{(t)}\right)}{|\mathcal{O}^{(t)}|} - \frac{\sum_{i \in \mathcal{O}^{(t)}} I\left(\gamma_i = \gamma_{i,\pi_j}^{(t)}\right)}{|\mathcal{O}^{(t)}|},$$

with $I(\cdot)$ denoting the indicator function. In the above, $\hat{\gamma}_i^{(t)} = f^{(t)}(\mathsf{z}_i)$ and $\hat{\gamma}_{i,\pi_j}^{(t)} = f^{(t)}(\mathsf{z}_{i,\pi_j})$ are the predicted class for observation i before and after the permutation of the value of variable Z^j, respectively, where $\mathsf{z}_{i,\pi_j} = \left(z_i^1, \ldots, z_i^{j-1}, z_{\pi_j(i)}^j, z_i^{j+1}, \ldots, z_i^n\right)$. The variable importance of Z^j is therefore obtained by determining the decrease in prediction accuracy when the variable considered is randomly permuted. If the weather variable is highly associated with the response variable (i.e., the yield), the different prediction accuracy before and after permuting the variable considered increases and is therefore used as a measure of variable importance.

A first importance score for each variable is computed by averaging over all the RF trees:

$$VI(Z^j) = \frac{\sum_{t=1}^{ntree} VI^t(Z^j)}{ntree}.$$

The more widely known standardized variable importance measure, the z-score, is obtained by dividing $VI(Z^j)$ by its standard error, $\sigma/ntree$.

The variable selection procedure has been shown to suffer from bias in favor of variables with missing values (Strobl et al., 2007a). The bias propagates in an ensemble of trees and

is reflected in variable importance measures – in particular when predictor variables are of different types (e.g., Strobl et al., 2007b). Furthermore, the RF variable selection approach can favor correlated variables over weakly correlated ones. This latter bias also affects permutation importance measures. Unbiased variable selection can be achieved when node splitting subsamples are drawn without replacement, unlike in the bootstrapping approach used in Breiman's original contribution, together with unbiased splitting at each node (see Strobl et al., 2007b).

2.0.1 Probability distribution functions

The probability distribution functions of pixel-level weather indices are characterized by using mixture models including a Generalized Pareto distribution (GPD). The latter allows one to accurately estimate the risk of occurrence of events that are both rare and extreme, within a modified GPD framework across the whole gridded domain studied. We use the approach with unsupervised threshold selection proposed by Frigessi et al. (2002). For each pixel (i, j), the following mixture model is estimated by maximum likelihood:

$$f_{i,j}(w) = (1 - \alpha(\theta_{i,j}, w)) \times g_1(w) + \alpha(\theta_{i,j}, W) \times g_2(w),$$

where g_1 denotes the pdf of a Gamma distribution, g_2 the pdf of a GDP, and α the dynamic mixing function defined as

$$\alpha(w; \theta_{i,j}) = \frac{1}{2} + \frac{1}{\pi} \arctan\left(\frac{w - \mu_{i,j}}{\tau_{i,j}}\right), \quad \theta_{i,j} = (\mu_{i,j}, \tau_{i,j}),$$

where each $\mu_{i,j} > 0$ is a location parameter and $\tau_{i,j} > 0$ a transition speed parameter.

2.0.2 Copula-based framework

A copula-based approach is used to characterize the bivariate distribution of rainfall-heat indices. Looking at the more general case of k-variate indices, a k-dimensional copula function

is defined as[4] the joint cumulative distribution function

$$C(u_1, \ldots, u_k) = \mathbb{P}(U_1 \leq u_1, \ldots, U_k \leq u_k), \tag{1}$$

where the uniformly distributed random variables $U_1, \ldots, U_k$ represent the marginal cumulative distribution functions of a k-dimensional vector of weather indices. Biffis and Chavez (2017) use an Archimedean copula to model the dependency of deficit rainfall and excess temperature events. Archimedean copulas are constructed with Archimedean generator functions of different classes, which provide a flexible framework for capturing different dependency structures of the marginals based on the tuning of a small set of parameter values. On our k-dimensional space, a copula is Archimedean if it admits the following representation:

$$C(u) = \Psi(\Psi^{-1}(u_1) + \Psi^{-1}(u_2) + \ldots + \Psi^{-1}(u_k)), \tag{2}$$

where Ψ is the generator. We use the Gumbel-Hougaard generator, which has inverse $\Psi_\theta(t) = \exp(-t^{1/\theta})$, for $\theta \geq 1$, the limiting case $\theta = 1$ providing the independent copula function. The copula for the joint probability distribution of deficit rainfall and excess heat is therefore given by $C_\theta(u_R, u_T) = -\left((-\log(u_R))^\theta + (-\log(u_T))^\theta\right)^{-1/\theta}$, where u_R and u_T are the uniform transforms of the marginal probability distribution functions of deficit precipitation and excess heat indices.

3 Optimization problem: Technical details

The optimization problem (5) considered in Biffis and Chavez (2017) reads as follows:

$$\begin{cases} \max_{(\pi,X)\in\mathcal{A}} & E\left[U\left(H - K(u) - \max\left(\mu_R - R(u,W), 0\right) - \pi + X - \text{LGD} \times 1_D\right)\right] \\ \text{subject to} & E\left[\max(A + \pi - X, 0)\right] \geq \underline{v}, \end{cases} \tag{3}$$

[4]See McNeil et al. (2015) for details on the risk modelling tools discussed in this section.

where 1_D represents the default indicator of the risk transfer facility, the default event being defined as $D := \{A + \pi - X < 0\}$, where $A \geq 0$ represents the end-of-period random market value of the assets backing the facility. The quantity LGD represents the Loss Given Default of the facility, which is given by $X - \gamma(A + \pi)$, with $\gamma \in [0, 1]$ denoting the fractional recovery of the residual assets $(A + \pi)$. The upper bound $(\gamma = 1)$ captures the case of full recovery, the lower bound $(\gamma = 0)$ that of zero recovery.

The standard concavity assumptions ensuring the existence of optimizers do not apply to problem (3), but Biffis and Millossovich (2013) show that a Cesari-type growth condition is satisfied, and hence the problem admits a generalized solution in the sense of Zavlaski (2001): there exists a dense subset of parameters for which the problem has a unique solution, and this solution is stable under small perturbations. The optimal risk transfer entails a premium level π^* related to the optimal payout X^* as follows:

$$\pi^* = E[X^*] + (\underline{v} - E[A]) - E\left[\max\left(X^* - (A + \pi^*), 0\right)\right], \tag{4}$$

where the first term on the right hand side represents the actuarially fair premium, the second term represents the excess return required by the risk taker (the investor in the risk facility), and the third term captures a discount reflecting the credit risk carried by the risk facility. Under suitable conditions, the optimal payout X^* takes a form similar to an insurance layer, as discussed in Biffis and Chavez (2017). In particular, the case in which A is independent of H, $LGD < 1$, and H is independent of $\max\left(\mu_R - R(u, W), 0\right)$, results in a strictly positive attachment point (a deductible in the terminology of Biffis and Millossovich, 2013) followed by partial coverage, which gradually reduces, thus approximating a detachment point; see Biffis and Millossovich (2013, figure 3) or Biffis and Chavez (2017, figure 6). The case in which H is stochastically decreasing[5] in $\max\left(\mu_R - R(u, W), 0\right)$ gives rise to a similar

[5]A random vector X is stochastically increasing in another random vector Y if the map $y \mapsto E\left[g(X)|Y = y\right]$ is nondecreasing for every nondecreasing function g such that the expectation of $g(X)$ exists. A random vector X is stochastically decreasing in another random vector Y if X is stochastically increasing in $-Y$. See Müller and Stoyan (2002).

result. An example of this situation is represented by the following specification: $H = \max(R(u, W), \mu_R)$, meaning that the decision maker benefits directly from any realized production revenue above the historical average.

References

Biffis, E. and Chavez, E. (2017). Satellite data and machine learning for weather risk management and food security. Technical report, Imperial College Business School and Centre for Environmental Policy, Imperial College London. To appear in *Risk Analysis*.

Biffis, E. and Millossovich, P. (2013). Optimal insurance with counterparty default risk. Technical report, SSRN Working paper. `http://dx.doi.org/10.2139/ssrn.1634883`.

Breiman, L. (2001). Random forests. *Machine Learning*, 45:5–32.

Friedman, J., Hastie, T., and Tibshirani, R. (2001). *The elements of statistical learning*. Springer Series in Statistics. Springer, Berlin.

Frigessi, A., Haug, O., and Rue, H. (2002). A dynamic mixture model for unsupervised tail estimation without threshold selection. *Extremes*, 5:219–235.

McNeil, A. J., Frey, R., and Embrechts, P. (2015). *Quantitative risk management: Concepts, techniques and tools*. Princeton University Press.

Müller, A. and Stoyan, D. (2002). *Comparison methods for stochastic models and risks*. John Wiley & Sons, Chichester.

Segal, M., Barbour, J., and R.M., G. (2004). Relating HIV-1 sequence variation to replication capacity via trees and forests. *Statistical Applications in Genetics and Molecular Biology*, 3:Article2.

Strobl, C., Boulesteix, A., and Augustin, T. (2007a). Unbiased split selection for classification trees based on the Gini index. *Computational Statistics and Data Analysis*, 52:483–501.

Strobl, C., Boulesteix, A., Zeileis, A., and Hothorn, T. (2007b). Bias in random forest variable importance measures: Illustration, sources and solutions. *BMC Bioinformatics*, 8:8–25.

Zavlaski, A. (2001). Existence of solutions of optimal control problems for a generic integrand without convexity assumptions. *Nonlinear Analysis*, 43:339–361.