An Inferential Framework to Reduce Climate Risk*

Gustavo Canavire-Bacarreza† Alejandro Puerta-Cuartas‡
Carlos Rodriguez-Castelán§ Carolina Vélez-Ospina¶

October 18, 2025

Abstract

With the increase of climatic shocks, quantifying their impact on vulnerability to poverty has gained significant attention. This paper extends the simulation approach to climate vulnerability of Hill and Porter (2017) for the design of targeted, place-based public policies to prevent and mitigate climate risk. We propose using SHAP values to characterize the most vulnerable to climate shocks and estimate the heterogeneous impact of specific climate shocks on future consumption. We illustrate our approach empirically by considering Ecuador, a biodiverse country with high exposure to climate risk and vulnerability. Our analysis reveals that the vulnerable are mostly informal individuals, working in the primary sector, and living in rural areas, located in the Amazonian Region, which motivates implementing a targeted place-based formalization policy. Our analysis highlights the importance of implementing preventive measures in Imbabura and Pastaza. While the former ranks among the three provinces most affected by droughts and floods, the latter is one of the most affected by maximum temperatures and droughts.

Keywords: Poverty Vulnerability, Climate Shocks, Machine Learning.
JEL Codes: I32, I38, C14, C15.

*The opinions expressed in this paper are those of the authors and do not necessarily reflect the views of the World Bank, its Board of Directors, or the countries it represents.

†World Bank, Poverty and Equity Global Practice, and the Universidad Privada Boliviana gcanavire@worldbank.org

‡Department of Economics, Universidad Carlos III de Madrid, alpuerta@eco.uc3m.es

§World Bank, Poverty and Equity Global Practice, crodriguezc@worldbank.org

¶World Bank, Poverty and Equity Global Practice, cvelezospina@worldbank.org

1 Introduction

In recent decades, assessing poverty vulnerability has become increasingly important in the development literature, largely due to the various shocks that the world has faced. Natural disasters have become more frequent and severe (Cred, 2020), motivating the enhancement of the existing tools for studying the impact of climate shocks on poverty vulnerability.

This paper proposes an inferential framework for designing targeted, place-based public policies to prevent and mitigate climate risk. For this purpose, we formalize and generalize the simulation approach to climate vulnerability of Hill and Porter (2017). Specifically, we show that their proposal can be posed as an out-of-sample predictive problem and applied to estimate functions of vulnerability measures, such as the characteristics of the most vulnerable. Furthermore, our generalization allows estimating a wide class of functions, e.g., the heterogeneous impact of specific climate shocks on future consumption.

Hill and Porter's approach-henceforth HPA-, consists of estimating poverty vulnerability by simulating the probable distribution of future consumption using historical data on different shocks. Specifically, it (1) estimates with ordinary least squares (OLS) the conditional mean function ($m(\cdot)$) relating consumption to household characteristics and shocks, (2) simulates future consumption using historical data on shocks with $\hat{m}(\cdot)$, and (3) computes a measure of vulnerability with the simulated future consumption. While the HPA considers death, job loss, price, and climate shocks, this paper focuses on the latter.

Our formalization of the HPA motivates adopting a Machine Learning (ML) approach for simulating future consumption. The second step of the HPA consists of predicting the dependent variable Y for test data ($\boldsymbol{X}_{\text{test}}$) as $\hat{Y} = \hat{m}(\boldsymbol{X}_{\text{test}}) = \boldsymbol{X}_{\text{test}}^{\intercal}\hat{\boldsymbol{\beta}}$. Thus, it simulates (predicts) future consumption using household characteristics and historical climate (test) data, which corresponds to an out-of-sample prediction problem. Consequently, estimating $m(\cdot)$ should involve regularization to avoid overfitting. Furthermore, since the effect of climate shocks on consumption is likely to be heterogeneous across households and regions, the conditional mean function $m(\cdot)$ is potentially nonlinear. As a result, utilizing ML to simulate future consumption provides an appealing approach, because it estimates $m(\cdot)$ non-parametrically, providing high model complexity while involving regularization to attain better out-of-sample predictive accuracy than OLS.

The third step of the HPA provides a measure of vulnerability. For instance, each

realization of future consumption for household i can be compared to the poverty line to establish the proportion of times (p_i) that consumption falls below the poverty line. This way, the vulnerability measure proposed by Pritchett et al. (2000) can be estimated as the proportion of households whose p_i exceeds 0.5. Alternatively, an analogous procedure can be undertaken for estimating different vulnerability measures, such as those proposed in Chaudhuri et al. (2002) and Calvo and Dercon (2013).

The estimated vulnerability can be further exploited to characterize the vulnerable population. Hill and Porter (2017) compute different vulnerability measures by gender and rural/urban status, which can be extended to unveil the most salient features of the vulnerable, e.g., uneducated, working in the primary sector. Moreover, calculating the vulnerability rate by region pinpoints the regions with the highest poverty risk. These estimates contribute to the effective design of place-based policies targeted to the most vulnerable.

Analyzing the effect of climate shocks on future consumption offers valuable information for preventing climate risk. Characterizing the vulnerable provides crucial insights for reducing climate risk. However, it does not address the impact of specific climate shocks. To overcome this limitation, we analyze the relationship between simulated future consumption and climate shocks to pinpoint the most affected regions by specific climate shocks.

We utilize the SHapley Additive exPlanations (SHAP) values proposed in Lundberg (2017) to unveil the most salient features of the vulnerable and estimate the heterogeneous impact of climate shocks on future consumption. SHAP values provide a measure of feature importance, enabling the identification of the most salient characteristics of vulnerable individuals. Furthermore, they measure the impact of the independent variables on the dependent variables by observation, which can be utilized to assess the heterogeneous effect of climate shocks on future consumption.

The main contribution of this paper is presenting an easy-to-implement procedure to design targeted, place-based public policies to reduce climate risk. To estimate functions of vulnerability, our proposal consists of four stages: (1) estimating the conditional mean function relating consumption (or income) to individual characteristics and climate shocks with Machine Learning, (2) simulating future consumption using individual characteristics and historical data on climate shocks with the estimated conditional mean, (3) predicting the individual vulnerability status according to their simulated future consumption, and (4) estimating the function of vulnerability. When the function is known, we propose to estimate its sample counterpart.[1] Conversely,

[1]For instance, to characterize the vulnerable individuals, we can compute averages for binary vari-

for unknown functions, we propose computing the SHAP values from a classification model by regressing the estimated vulnerability status on individual characteristics and climate shocks. To estimate unknown functions of future consumption, we follow the same procedure but omit the third step.

Our method can be applied to short panels and cross-sectional data. Similar to the multi-level (e.g., Skoufias and Quisumbing (2003) and Günther and Harttgen (2009)) and simulation approaches (e.g., Hill and Porter (2017)) to poverty vulnerability, our proposal does not require the availability of lengthy panel data on income or consumption. In particular, it only requires survey data with household or individual characteristics and historical climate data at the community or regional level. In our empirical application, we utilize publicly available climate data, illustrating that practitioners need only to count on survey data, and merge it with this dataset to implement our procedure.

We illustrate our proposal with an empirical application to Ecuador, a country with high exposure to climate risk and high vulnerability. Ecuador is one of the most biodiverse countries in the world (Kleemann et al., 2022), with 96 percent of its population living in coastal and mountainous regions that are exposed to several natural hazards (World Bank, 2014). According to the Survey of Life Conditions (Encuesta de Condiciones de Vida) carried out by the Instituto Nacional de Estadística y Censos (INEC), around one out of four households is at risk of poverty or social exclusion (INEC, 2023).

We construct a panel for the period 2007–2021 consisting of climate and individual data. We consider four climate shocks: maximum and minimum temperatures, floods, and droughts. We draw the data from the Climatic Research Unit of the University of East Anglia (Harris et al., 2020),[2] which provides gridded Time Series information at a spatial resolution of approximately 1 km^2. We merge the climate data with the National Survey conducted by INEC, yielding a pool of nearly 400,000 individuals.

Our empirical results suggest that individuals aged 25-45 with less than primary education are most likely to be vulnerable. Furthermore, we find that the vulnerable are mostly informal individuals working in the primary sector and living in rural areas, located in the Amazonian Region, which motivates implementing a targeted place-based formalization policy. Our analysis underscores the need for preventive measures in Imbabura and Pastaza. While the former ranks among the three provinces most

ables and the empirical cumulative distribution function for continuous variables. Furthermore, provided information on whether individuals live in rural areas and their age, we can estimate the proportion of vulnerable individuals working in the primary sector and the distribution of age of the vulnerable.

[2]The data can be accessed at `https://www.worldclim.org/data/`.

affected by droughts and floods, the latter is among the most affected by maximum temperatures and droughts.

The remainder of the paper proceeds as follows: section 2 formalizes the simulation approach proposed in Hill and Porter (2017) as an out-of-sample prediction problem, and motivates the adoption of a machine learning approach for simulating future consumption. Section 3 generalizes the simulation approach to climate vulnerability by extending it to prevent and mitigate climate risk, and presents the proposed estimation approach. Section 4 presents our application to Ecuador, and section 5 concludes.

2 A Formalization of the Simulation Approach to Climate Vulnerability

Our point of departure is the estimation of a measure of vulnerability $\mathcal{V}$, which depends on future consumption C_{t+1}, the poverty line z, and an exogenous parameter α. Accordingly, our object of interest can be defined as

$$\mathcal{V} \coloneqq g\left(C_{t+1}, z, \alpha\right), \tag{1}$$

where $g\left(\cdot\right)$ is known up to C_{t+1}, z, and α. Thus, if the distribution of future consumption were known, we could compute the vulnerability measure provided values for z and α.

Equation (1) encompasses a wide class of vulnerability measures. To begin with, consider the expectation of poverty as proposed by Chaudhuri et al. (2002)

$$\mathcal{V} = g\left(C_{t+1}, z, \alpha\right) = \mathbb{E}\left[\left(\max\left\{0, \frac{z - C_{t+1}}{z}\right\}\right)^{\alpha}\right],$$

where the parameter α determines whether the object of interest is the headcount ($\alpha = 0$), the poverty gap ($\alpha = 1$), or the squared gap ($\alpha = 2$). Another example of equation (1) is the measure proposed by Pritchett et al. (2000), where the higher the probability that the future consumption falls below the poverty line, the higher the probability that an individual is vulnerable, so that equation (1) boils down to

$$\mathcal{V} = g\left(C_{t+1}, z, \alpha\right) = P\left(C_{t+1} \leq z\right) > \alpha,$$

where α is commonly set to 0.5. Finally, assuming a constant relative risk-sensitive measure as in Calvo and Dercon (2013), would result in a vulnerability measure of the

type

$$\mathcal{V} = g\left(C_{t+1}, z, \alpha\right) = \frac{1}{\alpha}\left(1 - \mathbb{E}\left[\left(\frac{C_{t+1}}{z} \times 1\left(C_{t+1} < z\right) + z \times 1\left(C_{t+1} \geq z\right)\right)^{\alpha}\right]\right).$$

Because future consumption is unknown, we need to estimate it to measure vulnerability. The proposal of Hill and Porter (2017) consists of simulating the probable distribution of future consumption conditional using household characteristics and historical data on different shocks. In this paper, we focus on climate shocks. To formalize their approach, consider the decomposition of consumption into observed and unobserved components as

$$C = m\left(\boldsymbol{X}, \boldsymbol{S}; \boldsymbol{\eta}\right) + \epsilon, \quad \mathbb{E}\left[\epsilon | \boldsymbol{X}, \boldsymbol{S}\right] = 0, \tag{2}$$

where $\boldsymbol{X}$ and $\boldsymbol{S}$ are individual characteristics and climate shocks, respectively, $m\left(\cdot\right) \coloneqq \mathbb{E}\left[C|\boldsymbol{X}, \boldsymbol{S}\right]$ is the conditional mean function, known up to the nuisance parameter $\boldsymbol{\eta}$,[3] and ϵ is an idiosyncratic shock, corresponding to the unexplained component.

The Hill and Porter's approach is motivated by the fact that if we were to know $m\left(\cdot\right)$, $\boldsymbol{\eta}$, and the distributions of the climate and idiosyncratic shocks, future consumption can be simulated as

$$\begin{aligned} C_{t+1}^{m} &= m\left(\boldsymbol{X}, \boldsymbol{S}^{m}; \boldsymbol{\eta}\right) + \epsilon^{m}, \\ S^{m} &\sim F_{S}, \\ \epsilon^{m} &\sim F_{\epsilon}, \end{aligned} \tag{3}$$

where F_S and F_ϵ are the distribution of the climate and idiosyncratic shocks, respectively, and S^m and ϵ^m are draws from their respective distributions. Thus, by drawing C_{t+1}^m enough times (say M), we could approximate the distribution function of future consumption, and consequently compute the vulnerability measure in equation (1).

We now formalize the HPA in the following three-step procedure

1. Estimate the equation relating consumption (C) to household characteristics ($\boldsymbol{X}$) and climate shocks ($\boldsymbol{S}$)

$$C = m\left(\boldsymbol{X}, \boldsymbol{S}; \boldsymbol{\eta}\right) + \epsilon, \quad \mathbb{E}\left[\epsilon | \boldsymbol{X}, \boldsymbol{S}\right] = 0,$$

[3]For instance, if we assume that $m\left(\cdot\right)$ is linear, the nuisance parameter $\boldsymbol{\eta}$ corresponds to the regression coefficients. We provide a characterization of the conditional mean function and the nuisance parameter for the case of tree-based algorithms and kernel machines in the Appendix

by Ordinary Least Squares, assuming

$$m\left(\boldsymbol{X},\boldsymbol{S};\boldsymbol{\eta}\right)=\boldsymbol{X}'\boldsymbol{\beta}+\boldsymbol{S}'\boldsymbol{\delta}+\boldsymbol{S}'\tilde{\boldsymbol{X}}'\boldsymbol{\gamma}, \tag{4}$$

where $\tilde{\boldsymbol{X}}$ is a subset of $\boldsymbol{X}$ so the effect of climate shocks on consumption is allowed to depend on individual characteristics.

Thus, the estimated conditional mean function is given by

$$\hat{m}\coloneqq m\left(\boldsymbol{X},\boldsymbol{S};\hat{\boldsymbol{\eta}}\right)=\tilde{\boldsymbol{Z}}'\left(\tilde{\boldsymbol{Z}}\tilde{\boldsymbol{Z}}'\right)^{-1}\tilde{\boldsymbol{Z}}'C\coloneqq\tilde{\boldsymbol{Z}}'\hat{\boldsymbol{\eta}},\quad\tilde{\boldsymbol{Z}}'\coloneqq\left(\boldsymbol{X}',\boldsymbol{S}',\boldsymbol{S}'\tilde{\boldsymbol{X}}'\right)',$$

and the empirical cumulative distribution function (E.C.D.F) of the residuals $\hat{\epsilon}\coloneqq C-\tilde{\boldsymbol{Z}}'\hat{\boldsymbol{\eta}}$ is given by

$$\hat{F}_{\epsilon}(e)=\frac{1}{n}\sum_{i=1}^{n}1\left(C_i-\tilde{\boldsymbol{Z}}_i'\hat{\boldsymbol{\eta}}<e\right)\coloneqq\mathbb{E}_n\left[1\left(C_i-m\left(\boldsymbol{X}_i,\boldsymbol{S}_i;\hat{\boldsymbol{\eta}}\right)<e\right)\right],$$

where $\mathbb{E}_n\left[\cdot\right]$ is the empirical expectation operator.

2. Simulate M times the future consumption by drawing climate shocks and idiosyncratic shocks from their E.C.D.F

$$\begin{aligned}\hat{C}_{t+1}^{m}&=m\left(\boldsymbol{X},\boldsymbol{S}^{m};\hat{\boldsymbol{\eta}}\right)+\epsilon^{m}=\tilde{\boldsymbol{Z}}^{m'}\hat{\boldsymbol{\eta}}+\epsilon^{m},\quad m=1,...,M,\\ S^{m}&\sim\hat{F}_{S},\quad\hat{F}_{S}(s)=\mathbb{E}_n\left[1\left(\boldsymbol{S}_i<s\right)\right],\\ \epsilon^{m}&\sim\hat{F}_{\epsilon},\quad\hat{F}_{\epsilon}(e)=\mathbb{E}_n\left[1\left(C_i-m\left(\boldsymbol{X}_i,\boldsymbol{S}_i;\hat{\boldsymbol{\eta}}\right)\right)\right],\end{aligned} \tag{5}$$

where $\tilde{\boldsymbol{Z}}^{m'}\coloneqq\left(\boldsymbol{X}',\boldsymbol{S}^{m'}\right)'$.

Compute

$$\hat{C}_{t+1}\sim\hat{F}_{C},\quad\hat{F}_{C}(c)=\frac{1}{M}\sum_{m=1}^{M}1\left(\hat{C}_{t+1}^{m}<c\right),$$

so that

$$\hat{C}_{t+1}=f\left(C,\boldsymbol{X},\phi,m\right),\quad\phi\coloneqq\left(F_S,F_{\epsilon}\right). \tag{6}$$

Thus, simulated future consumption depends on the observed consumption, individual characteristics, the distribution function of climate and idiosyncratic shocks, and the conditional mean function.

3. Estimate the vulnerability measure in equation (1) as

$$\hat{\mathcal{V}} = g\left(\hat{C}_{t+1}, \alpha, z\right). \tag{7}$$

The aforementioned procedure involves estimating the unknown distribution functions and a conditional mean. To illustrate this point, consider plugging equation (6) into equation (7), so that we can reparametrize the object of interest as

$$\hat{\mathcal{V}} = g\left(\hat{C}_{t+1}, z, \alpha\right) \coloneqq \tilde{f}\left(C, \boldsymbol{X}, \phi, m, z, \alpha\right).$$

Thus, the estimated measure of vulnerability depends on consumption and individual characteristics $(C, \boldsymbol{X})$, unknown distribution functions (ϕ), the conditional mean function (m), and exogenous values (z, α). Because the data and the poverty line are given, and the exogenous value α is defined ad-hoc for the analysis, we only need to estimate the conditional mean and the unknown distribution functions.

The HPA estimates the distribution of future consumption, climate, and idiosyncratic shocks with the E.C.D.F. This is justified by the Glivenko-Cantelli Theorem, which establishes that

$$||F_n - F||_\infty \coloneqq \sup_{x\in\mathbb{R}}|F_n(x) - F(x)| \to 0 \ \ a.s., \tag{8}$$

where F is the common cumulative distribution of the random variables $X_1, ..., X_M$ which are independent and identically distributed in $\mathbb{R}$, and

$$F_n(x) \coloneqq \frac{1}{M}\sum_{m=1}^{M} 1\left(X_m < x\right)$$

is the empirical cumulative distribution function of $\{X_i\}_{i=1}^M$ evaluated at x. Equation (8) shows that the E.C.D.F. converges uniformly to the true distribution function almost surely. Thus, the E.C.D.F. is a consistent estimator for the C.D.F., which justifies the HPA to estimate the unknown distribution functions.

The key aspect of estimating the conditional mean function is that its purpose is to perform out-of-sample prediction. As illustrated by equation (5), to simulate future consumption, we evaluate the estimated conditional mean function on out-of-sample (test) data, which is, by definition, out-of-sample prediction. Thus, estimating $m\left(\cdot\right)$ should involve regularization to avoid overfitting. Additionally, since the effect of climate shocks on consumption is likely to be heterogeneous across households and regions, the conditional mean function $m\left(\cdot\right)$ is potentially non-linear.

Estimating the conditional mean function by OLS entails a substantial loss in the predictive accuracy of future consumption. Because the effect of climate shocks on consumption varies across households and regions, the conditional mean function $m(\cdot)$ is potentially non-linear. Furthermore, as illustrated by equation (5), the simulation of future consumption corresponds to an out-of-sample predictive problem, so the estimation of $m(\cdot)$ should be flexible enough to account for non-linearities and should also involve regularization to avoid overfitting.

The linearity assumption of OLS can be partially addressed by projecting the covariates into some high-dimensional space using basic feature-space functions. For instance, by projecting covariates into the space of powers $t(x) = (1, x, x^2, x^3, ...)'$, or incorporating interactions between them, in the spirit of equation (4). Whenever the functions are independent of the parameters, the model is still linear on the parameters (Williams and Rasmussen, 2006).

The HPA to overcome the linearity assumption is interacting a subset of individual characteristics with climate shocks based on economic rationale. Even though this practice helps to avoid overfitting by omitting variables, it is not data-driven and could entail a substantial loss in accuracy. Furthermore, this approach allows for shock heterogeneity across individuals but not across regions. While certain mechanisms are inherently tied to specific household characteristics (indirect effect), the forces propelling individuals into poverty are partly external (direct effect). Thus, the extent to which shocks affect individuals depends on their characteristics and the socioeconomic and geographic characteristics of the region in which they live. Despite this can be internalized by interacting climate shocks with regional dummies, it would exacerbate the need for regularization.

Adopting a ML approach incorporates the predictive nature of simulating future consumption. To begin with, it provides high model complexity by modeling non-parametrically the conditional mean function, relaxing the linearity assumption of OLS. Furthermore, it involves regularization, which avoids overfitting. Thus, our formalization of the HPA motivates adopting a ML approach for simulating future consumption.

Despite its benefits, the machine learning approach induces bias in predicting the vulnerability status. This issue arises from the regularization and model selection bias introduced by the models used to estimate the conditional mean (Chernozhukov et al., 2022). Thus, errors in the out-of-sample consumption prediction would affect the estimated vulnerability status. Specifically, individuals could be misclassified as vulnerable when their future consumption is not well approximated. However, this limitation also arises in OLS estimation for different reasons. First, the linearity assumption causes the

future consumption to be poorly approximated. Second, the inclusion of interactions without regularization can lead to over-fitting (Hawkins, 2004).

In this section, we have formalized the HPA to climate vulnerability, highlighting the advantages of adopting a ML approach for simulating future consumption. While ML shares some limitations with OLS, the former approach provides better predictive accuracy through higher model complexity and regularization, enabling more accurate estimation of vulnerability. We now turn to illustrate how the HPA can be extended to estimate functions of future consumption to reduce climate risk.

3 An Inferential Framework to Reduce Climate Risk

3.1 Estimating Functions of Future Consumption

Equation (1) shows that we can estimate a measure of vulnerability by simulating future consumption. We now generalize this object of interest to allow for a more general setting in which we are interested in functions of future consumption. In particular, we focus on the estimation of a parameter θ, which is a function $h(\cdot)$ depending on individual characteristics, climate shocks, future consumption, the poverty line z, and an exogenous parameter α

$$\theta = h\left(\boldsymbol{Z}, C_{t+1}, z, \alpha\right), \tag{9}$$

where $\left(\boldsymbol{Z} \coloneqq (\boldsymbol{X}', \boldsymbol{S}')'\right)$. We first focus on a particular case of equation (9), where the object of interest depends on individual characteristics, climate shocks, and the individual vulnerability status V_i, which is determined according to the vulnerability measure of Pritchett et al. (2000)

$$V_i = 1\left(P\left(C_{i,t+1} \leq z\right) > \alpha\right), \quad i = 1, ..., n. \tag{10}$$

In particular, our object of interest has the form

$$\theta = h\left(\boldsymbol{Z}, C_{t+1}, z, \alpha\right) = \tilde{h}\left(\boldsymbol{Z}, V\right), \tag{11}$$

where in the second equality we have used equations (9) and (10), and V is a vector with information on the vulnerability status for every individual.

Equation (11) generalizes the objects of interest in Hill and Porter (2017). In their empirical application, the authors estimate five different vulnerability measures by gender

and rural/urban status, which can be formalized as

$$\theta = \tilde{h}\left(\boldsymbol{Z}, V\right) = \mathcal{V}|\boldsymbol{Z},$$

where Z is the conditioning variable (gender and rural/urban status).[4]

The main implication of equation (11) is that the HPA can be used to estimate functions of vulnerability. More importantly, as we will now show, it can be applied to the design of targeted, place-based policies to prevent and mitigate climate risk. We now illustrate with a motivating example how equation (11) can be used to characterize the vulnerable

Consider the object of interest in equation (11) is given by

$$\theta = \tilde{h}\left(\boldsymbol{Z}, V\right) = F_{\boldsymbol{X}|V=1}\left(\boldsymbol{x}\right) = P\left(X_1 \leq x_1, ..., X_r \leq x_{ls} | V = 1\right), \tag{12}$$

where $F_{\boldsymbol{X}|V=1}\left(\boldsymbol{x}\right)$ is the joint cumulative distribution function of the vulnerable's characteristics, and l_s is the cardinality of $\boldsymbol{X}$. For instance, if individual characteristics only include the age of the individual (age) and a dummy variable ($prim$) which equals 1 if the individual works in the primary sector and zero otherwise, we have that equation (12) boils down to

$$F_{\boldsymbol{X}|V=1}\left(\boldsymbol{x}\right) = P\left(age_1 \leq x_1, prim \leq x_2 | V = 1\right),$$

with corresponding marginal distributions given by

$$F_{age}(a) = \mathbb{E}\left[1\left(age \leq a\right)|V = 1\right],$$

and

$$F_{prim}(a) = \mathbb{E}\left[1\left(prim \leq a\right)|V = 1\right].$$

If we were to find that the majority of the vulnerable are old individuals working in the primary sector, the government could target efforts towards this population aiming to prevent climate risk.

To motivate our proposal to identify the most salient characteristics of vulnerable, consider estimating equation (10) using the first two steps of the HPA. Thus, according

[4]Notice that the measure of vulnerability $\mathcal{V}$ summarizes in a scalar the vector V with information on the individual vulnerability status.

to equation (5), we have

$$\hat{V}_i = 1\left(P\left(\hat{C}_{i,t+1} \leq z\right) > \alpha\right) := h\left(C_i, \boldsymbol{X}_i, \hat{\phi}, \hat{m}, z, \alpha\right), \quad \hat{\phi} := \left(\hat{F}_S, \hat{F}_\epsilon\right), \tag{13}$$

so the individual vulnerability measure, can be expressed as a function of individual characteristics (and others). This arises from the fact that to simulate future consumption using the HPA, we keep fixed individual characteristics (see equation (5)). Accordingly, to identify the most salient characteristics of the vulnerable we propose estimating

$$\theta = \tilde{h}\left(\boldsymbol{Z}, V\right) = \tilde{h}\left(\boldsymbol{X}, V\right) = \nabla_{\boldsymbol{X}} V_i = \begin{bmatrix} \frac{\partial P(V_i=1)}{\partial X_1} \\ \vdots \\ \frac{\partial P(V_i=1)}{\partial X_{l_x}} \end{bmatrix}, \quad i = 1, ...n, \tag{14}$$

where $\nabla_{\boldsymbol{X}} V_i$ denotes the gradient of the individual probability of vulnerability with respect to individual characteristics.

Equation (14) captures the relationship between the vulnerability status and the individual characteristics. The usefulness of this object is two-fold. First, it provides a measure of variable importance, in the sense that, the variables with the highest impact on the probability of being vulnerable correspond to the most salient attributes of the vulnerable. Second, it allows us to characterize the most vulnerable households. For instance, if we were to find that $\frac{\partial P(V_i=1)}{\partial Age}$ is U-shaped, the youngest and oldest are the most vulnerable. Conversely, if the derivative is constant and negative, the most vulnerable are the youngest.

One limitation of the simulation approach is that the vulnerability status does not depend on the realized shocks individuals have been exposed to. Even though $\hat{V}_i$ depends on the estimated distribution of the climate shocks $\left(\hat{F}_s\right)$, it does not depend on the realized shocks $\boldsymbol{S}_i$. For this reason, we cannot undertake a similar approach to that of equation (14) to study the impact of climate shocks on vulnerability.

Although we cannot assess the effect of climate shocks on vulnerability, we can analyze their impact on future consumption. According to equation (5), simulated future consumption depends on individual characteristics and the draws of climate shocks. Thus, in the same spirit as equation (14), we can study the heterogeneous impact of

climate shocks on future consumption by estimating

$$\theta = h\left(\boldsymbol{Z}, C_{t+1}, z, \alpha\right) = h\left(\boldsymbol{S}, C^m_{t+1}\right) = \nabla_{\boldsymbol{S}} C^m_{i,t+1} = \begin{bmatrix} \frac{\partial^m_{i,t+1}}{\partial S_1} \\ \vdots \\ \frac{\partial^m_{i,t+1}}{\partial S_{l_s}} \end{bmatrix}, \quad i = 1, ...n, \tag{15}$$

where $\nabla_{\boldsymbol{S}} C^m_{i,t+1}$ is the gradient of a realization of the individual future consumption with respect to climate shocks, and l_s is the cardinality of $\boldsymbol{S}$. Similar to equation (14), equation (15) informs us about the most affected when a specific climate shock occurs. Since the derivates are computed for each individual, and we observe individual characteristics, we can identify the most affected individuals and regions by specific climate shocks. This way, the government can focus efforts on preventive actions to reduce climate risk. For instance, local governments in the most affected regions by floods can target efforts to reduce debris and waste in the drainage system to reduce blockages. Conversely, the most affected regions by maximum temperatures can invest in cool roofs and increase urban greenery.[5]

In this section, we have generalized the HPA by showing that their proposal can be extended to estimate functions of future consumption. Additionally, we have illustrated how estimating equations (14) and (15) allow the design of targeted, place-based public policies to prevent and mitigate climate risk. We now describe our proposal for estimating these two equations.

3.2 Estimation Procedure

In Section 2 we have highlighted the benefits of modeling consumption as a function of individual characteristics and climate shocks with ML. Before turning to the estimation procedure, we first illustrate how the ML approach can be complemented with novel tools from the explainable artificial intelligence literature to reduce poverty risk. In particular, we show that equations (14) and (15) can be estimated by computing the SHAP values proposed in Lundberg (2017).

The SHAP values build upon the Shapley values (Shapley, 1953), which dictate how the surplus in a cooperative game should be allocated among players. In particular, in a coalition C of N agents, the j-th Shapley value corresponds to the fair share of the

[5]For evidence on cool roofs and urban greenery reducing temperature, see Macintyre and Heaviside (2019) and Đekić et al. (2018), respectively

value of the coalition (V), that each agent j should receive, which is given by

$$\phi_j(V) = \frac{1}{N}\sum_{S}\frac{[V(S\cup\{j\}) - V(S)]}{\binom{N-1}{k_s}},$$

where the summation is over all the subsets S, of the team, $T = \{1, ..., N\}$, that one can construct after excluding j, k_s is the number of agents in the coalition S, $V(S)$ is the value achieved by subteam S, and $V(S\cup\{m\})$ is the realized value after j joins S. Thus, $\phi_j(V)$ measures the average contribution of j, which should be his/her fair share.

The Shapley values for regression evaluate the effect of every variable in a model in predicting an outcome for each observation. In a predictive framework, the agents correspond to each explanatory variable, and coalitions correspond to a given model using a subset of the explanatory variables. Intuitively, the Shapley values for regression are computed by estimating each possible model with and without each regressor, to assess its impact in predicting the dependent variable. However, since there are many potential combinations, the SHAP values rely on an approximation of the Shapley values (for a detailed formulation see Lundberg (2017)).

We propose computing the SHAP values to estimate equations (14) and (15). In our framework, the SHAP values measure the impact of every variable on vulnerability and welfare outcomes. Thus, they provide estimates for our two equations of interest. Consequently, by computing the SHAP values we can characterize the most vulnerable to climate shocks and estimate the heterogeneous effect of specific climate shocks on future consumption.

Our proposal to estimate functions of vulnerability can be formalized in the following four-step procedure

1. Estimate the equation relating consumption (C) to individual characteristics ($\boldsymbol{X}$) and climate shocks ($\boldsymbol{S}$)

$$C = m(\boldsymbol{X}, \boldsymbol{S}; \boldsymbol{\eta}) + \epsilon, \quad \mathbb{E}[\epsilon|\boldsymbol{X}, \boldsymbol{S}] = 0,$$

with machine learning, yielding the non-parametric estimate of the conditional mean $m(\boldsymbol{X}, \boldsymbol{S}; \hat{\boldsymbol{\eta}})$, and the E.C.D.F of the residuals $\hat{\epsilon} \coloneqq C - m(\boldsymbol{X}, \boldsymbol{S}; \hat{\boldsymbol{\eta}})$ is given by

$$\hat{F}_\epsilon(e) = \frac{1}{n}\sum_{i=1}^{n} 1\left(C_i - \tilde{\boldsymbol{Z}}_i'\hat{\boldsymbol{\eta}} < e\right) \coloneqq \mathbb{E}_n\left[1\left(C_i - m(\boldsymbol{X}_i, \boldsymbol{S}_i; \hat{\boldsymbol{\eta}}) < e\right)\right].$$

2. Simulate M times the future consumption by drawing climate and idiosyncratic shocks from their E.C.D.F

$$\begin{aligned}\hat{C}_{t+1}^{m} &= m\left(\boldsymbol{X}, \boldsymbol{S}^{m}; \hat{\boldsymbol{\eta}}\right) + \epsilon^{m}, \quad m = 1, ..., M, \\ S^{m} &\sim \hat{F}_{S}, \quad \hat{F}_{S}(s) = \mathbb{E}_{n}\left[1\left(\boldsymbol{S}_{i} < s\right)\right], \\ \epsilon^{m} &\sim \hat{F}_{\epsilon}, \quad \hat{F}_{\epsilon}(e) = \mathbb{E}_{n}\left[1\left(C_{i} - m\left(\boldsymbol{X}_{i}, \boldsymbol{S}_{i}; \hat{\boldsymbol{\eta}}\right)\right)\right].\end{aligned}$$

Compute

$$\hat{C}_{t+1} \sim \hat{F}_{C}, \quad \hat{F}_{C}(c) = \frac{1}{M}\sum_{m=1}^{M} 1\left(\hat{C}_{t+1}^{m} < c\right),$$

3. Compute the individual vulnerability status using

$$\hat{V}_{i} = 1\left(P\left(\hat{C}_{i,t+1} \leq z\right) > \alpha\right), \quad i = 1, ..., n.$$

4. Estimate equation (11) as

$$\hat{\theta} = \hat{\tilde{h}}\left(\boldsymbol{Z}, \hat{V}\right), \quad \hat{V} = \left(\hat{V}_{1}, ..., \hat{V}_{n}\right).$$

When $\tilde{h}$ is known (e.g., a conditional mean, or a vulnerability measure for a subgroup), then $\hat{\tilde{h}}$ is the sample counterpart of h. Conversely, when h is unknown, we propose estimating it using the SHAP values from a classification model by regressing the estimated vulnerability status on individual characteristics and climate shocks.

The main differences between our proposal and the HPA lie in steps 1 and 4. To begin with, instead of estimating the conditional mean with OLS, we perform non-parametric estimation with ML. In contrast, steps 2 and 3 are the same as in the HPA, with the subtle difference that in step 3, we do not compute a vulnerability measure, but keep the individual vulnerability status. Furthermore, step 4 encompasses a wide class of objects of interest, including vulnerability measures for the whole population and subpopulations (such as male/female and rural/urban status as in Hill and Porter (2017)), and equation (14).

A critical note on the first step is that, since machine learning methods are proposed, practitioners must conduct hyperparameter tuning to avoid overfitting. Care must also be taken to ensure this process does not assign zero weights to the climate shock variables, as doing so would restrict step 2 to only idiosyncratic shocks and individual

or household characteristics.

To estimate functions of future consumption, we follow the following procedure

1. Estimate the equation relating consumption (C) to individual characteristics ($\boldsymbol{X}$) and climate shocks ($\boldsymbol{S}$)

$$C = m\left(\boldsymbol{X}, \boldsymbol{S}; \boldsymbol{\eta}\right) + \epsilon, \quad \mathbb{E}\left[\epsilon | \boldsymbol{X}, \boldsymbol{S}\right] = 0,$$

with machine learning, yielding the non-parametric estimate of the conditional mean $m\left(\boldsymbol{X}, \boldsymbol{S}; \hat{\boldsymbol{\eta}}\right)$, and the E.C.D.F of the residuals $\hat{\epsilon} \coloneqq C - m\left(\boldsymbol{X}, \boldsymbol{S}; \hat{\boldsymbol{\eta}}\right)$ is given by

$$\hat{F}_{\epsilon}(e) = \frac{1}{n}\sum_{i=1}^{n} 1\left(C_i - \tilde{\boldsymbol{Z}}_i'\hat{\boldsymbol{\eta}} < e\right) \coloneqq \mathbb{E}_n\left[1\left(C_i - m\left(\boldsymbol{X}_i, \boldsymbol{S}_i; \hat{\boldsymbol{\eta}}\right) < e\right)\right].$$

2. Simulate future consumption one time by drawing climate and idiosyncratic shocks from their E.C.D.F

$$\begin{aligned}
\hat{C}_{t+1}^{m} &= m\left(\boldsymbol{X}, \boldsymbol{S}^{m}; \hat{\boldsymbol{\eta}}\right) + \epsilon^{m}, \\
S^{m} &\sim \hat{F}_{S}, \quad \hat{F}_{S}(s) = \mathbb{E}_n\left[1\left(\boldsymbol{S}_i < s\right)\right], \\
\epsilon^{m} &\sim \hat{F}_{\epsilon}, \quad \hat{F}_{\epsilon}(e) = \mathbb{E}_n\left[1\left(C_i - m\left(\boldsymbol{X}_i, \boldsymbol{S}_i; \hat{\boldsymbol{\eta}}\right)\right)\right].
\end{aligned}$$

3. Estimate equation (9) as

$$\hat{\theta} = \hat{h}\left(\boldsymbol{Z}, \hat{C}_{t+1}^{m}, z, \alpha\right).$$

When h is known, then $\hat{h}$ is the sample counterpart of h. Conversely, when h is unknown, we propose estimating it using the SHAP values from a regression model by regressing the simulated future consumption on individual characteristics and climate shocks.

We now illustrate how our proposal can be used to design targeted, place-based public policies to reduce climate risk, by applying it to the case of Ecuador.

4 Application

We construct a panel for the period 2007–2021 consisting of climate and individual-level data. We draw the temperature and precipitation data from the CRU-TS 4.06 (gridded Time Series of the Climatic Research Unit of the University of East Anglia,

for details see (Harris et al., 2020)) downscaled with WorldClim 2.1 (Fick and Hijmans, 2017). Our dataset comprises historical monthly weather data at a spatial resolution of approximately 1 km^2 for temperature and precipitation. This enables us to obtain the information at the parish level, which is the most detailed identification of the geographical location provided by the available household survey.

We gather information on four climate shocks: maximum and minimum temperature, floods, and droughts. The maximum and minimum temperatures correspond to the most extreme temperatures in a given parish during the year. The flood and drought variables are constructed based on the 3-month Standardised Precipitation Index (SPI-3), which measures the deficit and surplus of precipitation accumulated over 3 months. Following the literature (McKee et al., 1993), a drought event starts when the SPI-3 values fall below -1 and ends when the index returns positive. The magnitude of the drought is given by the sum of the SPI-3 during the drought event. Similarly, a flood even starts when the SPI-3 values fall above 1 and ends when the index returns negative, and the magnitude of the drought is given by the sum of the SPI-3 during the flood event.

We merge the climate data with data from the National Survey of Employment, Unemployment, and Underemployment (Encuesta Nacional de Empleo, Desempleo y Subempleo, ENEMDU), conducted by INEC. Our dependent variable is the per-capita household income variable used by INEC to compare with the poverty line to determine whether or not a household is poor. We also have information on whether the individuals live in a rural area, their relationship to the head of the household, informality and self-employment status, sex, whether the individual interviewed works in the primary sector (agriculture, hunting, forestry, fishing, and mining and quarrying), education level, and age. The pooling of data from the surveys of individuals merged with the climate data yields an individual-level data set with 395,988 observations. Our data set comprises information for every province in Ecuador, except for Galápagos, Santo Domingo, and Santa Elena. Thus, we have information for 206 parishes, belonging to 21 provinces.

As indicated in the four-step procedure, to estimate functions of a vulnerability measure, we first perform XGBoost Regression to estimate the equation relating consumption to individual characteristics (including year and province dummies) and climate shocks. In doing so, we obtain estimates for the conditional mean function and the E.C.D.F. of the residuals. The hyperparameters for the XGBoost Regression are selected via k-fold cross-validation. In particular, we split the dataset into training and test sets, and define a parameter grid covering various values for learning rate, maximum tree depth, number of estimators, and column sampling by tree. We then evaluate

combinations of these parameters using 2 folds. The grid search process identifies the parameter set that achieves the highest accuracy on the training data, enabling us to select the most effective model configuration while minimizing the risk of overfitting.

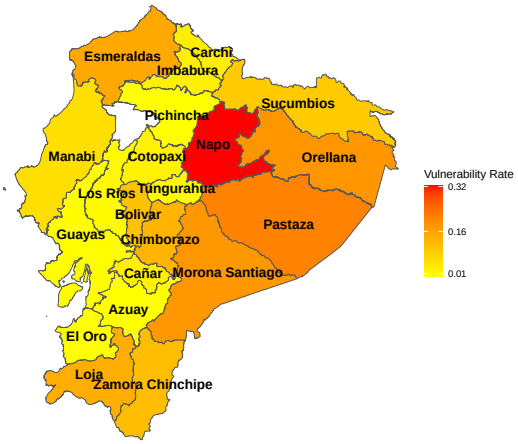
Second, we simulate $M = 5,000$ times future consumption by drawing climate and idiosyncratic shocks from their E.C.D.F. We draw the former at the province level, while the latter at the national level. That is, to simulate the climate shocks we utilize the E.C.D.F. within the province to which each individual belongs,[6] and for the idiosyncratic shocks, we use the E.C.D.F. for the whole sample. Third, we estimate the individual vulnerability based on the simulated future income. Finally, we estimate four variants of equation (11), namely, the regional vulnerability rate, summary statistics of individual characteristics, feature importance, and the effect of each characteristic on predicting poverty vulnerability. For the latter two objects, we compute the SHAP values from a classification model where the dependent variable is the estimated vulnerability status, and the independent variables are individual characteristics and climate shocks.

Our findings suggest that poverty vulnerability is concentrated in the Amazonian region of Ecuador. Figure 1 reports the vulnerability rate by province for the year 2021, depicting that Morona Santiago, Napo, and Pastaza are the most vulnerable provinces. We find that poverty vulnerability is concentrated in the Amazonian region of Ecuador, suggesting that place-based policies should target the eastern region of the country.

We summarize the regional dynamics of poverty vulnerability from 2007 to 2021 in Figures 2 and Table 1. Figure 2 depicts the evolution of poverty vulnerability by province, illustrating that poverty vulnerability in Bolivia has generally declined. However, most provinces experienced an increase in vulnerability between 2019 and 2020, followed by a subsequent decrease in 2021. Table 1 presents the ranking of each province based on its level of poverty vulnerability, where a higher rank indicates greater vulnerability. The table also displays the most frequent rank, as well as the minimum and maximum ranks observed. Table 1 shows that Morona Santiago, Napo, and Pastaza—the provinces with the highest vulnerability in 2021—have consistently ranked among the most vulnerable since 2007. Additionally, Manabi and Carchi, which ranked 4th and 5th in 2021, have remained in the top five most vulnerable provinces over the years. Surprisingly, Pastaza ranked 16th in 2007 and 14th in 2008. However, since 2009, it has remained among

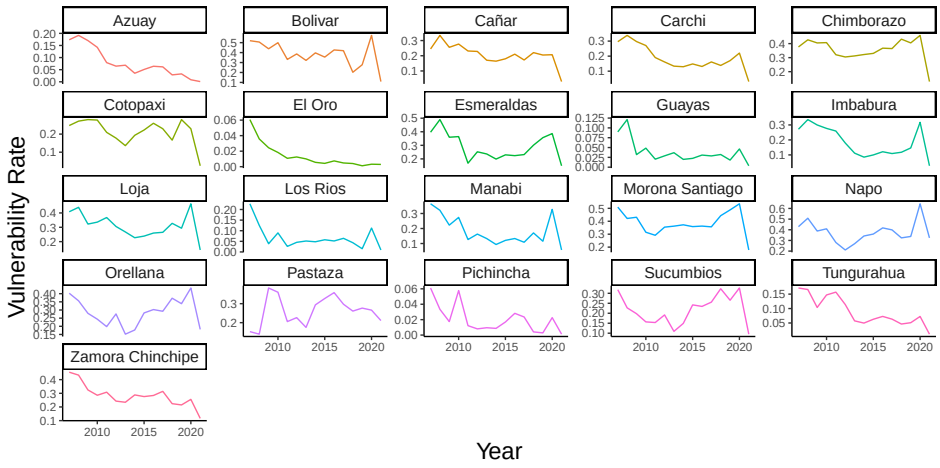
[6]The stratified sampling for the climate shocks aims to simulate "reasonable" climate shocks for every individual. Thus, we only consider climate shocks that have occurred in the individual's province. Even though we have information at the parish level, we stratify at the province level to obtain more variability.

Figure 1: Estimated Vulnerability Rate by Province, 2021



the most vulnerable provinces. Lastly, except for 2011, Morona Santiago has been the most vulnerable province in the country, underscoring the need for targeted efforts to reduce vulnerability.

Figure 2: Dynamics of Poverty Vulnerability by Province, 2007–2021



The vulnerable are mostly informal individuals, working in the primary sector, and living in rural areas. In Table 2 we report the average values for individual characteristics for the vulnerable and the entire population. By comparing both averages for the same variable, we isolate potential compositional effects.[7] We also find no systematic differences as regards gender, head, and civil status. Concerning self-employment status, we find that the proportion of vulnerable self-employed workers is higher than the

[7]If the vast majority of the sample is informal, finding that the vulnerable are mostly informal might result from the composition of the labor market.

Table 1: Rank of Provinces According to Poverty Vulnerability (descending), 2007–2021

	Morona	Napo	Manabi	Pastaza	Carchi	Sucumbios	Bolivar	Esmeraldas	Loja	Chimborazo	Zamora	Orellana	Pichincha	Imbabura	Azuay	Cotopaxi	Tungurahua	Los Rios	Cañar	Guayas	El Oro
2007	1	4	3	16	2	10	7	9	5	11	6	8	18	15	12	17	13	14	19	20	21
2008	1	2	7	14	5	13	6	3	4	9	8	16	19	12	15	17	18	11	10	20	21
2009	1	4	5	2	3	12	8	6	10	15	9	13	11	7	14	19	17	18	16	20	21
2010	1	2	3	4	5	14	6	7	9	11	10	19	8	17	15	12	18	16	13	20	21
2011	2	5	3	1	7	4	10	9	6	11	8	20	14	15	17	18	13	16	12	19	21
2012	1	8	4	5	3	7	2	9	6	12	11	10	14	15	16	17	13	18	19	20	21
2013	1	3	6	5	4	8	9	7	2	11	10	14	13	16	12	19	17	15	18	20	21
2014	1	3	9	2	4	8	5	10	7	11	6	12	17	19	16	13	18	15	14	20	21
2015	1	3	7	2	4	5	8	10	9	11	6	12	13	19	15	14	17	16	18	20	21
2016	1	2	10	3	6	7	5	11	9	8	4	12	14	18	16	13	15	19	17	20	21
2017	1	2	12	3	4	6	7	9	10	8	5	11	14	16	17	13	19	18	15	20	21
2018	1	5	6	3	2	4	12	7	9	10	11	8	17	13	14	15	16	18	19	20	21
2019	1	5	11	4	2	6	13	3	8	7	10	9	19	12	16	14	17	18	15	20	21
2020	1	2	3	8	6	4	9	7	5	13	11	10	14	12	20	17	16	15	19	18	21
2021	1	2	4	3	5	8	10	6	7	12	11	9	13	15	16	19	14	18	17	20	21
Mode	1	2	3	3	4	4	7	9	9	11	11	12	14	15	16	17	17	18	19	20	21
Min	1	2	3	1	2	4	2	3	2	7	4	8	8	7	12	12	13	11	10	18	21
Max	2	8	12	16	7	14	13	11	10	15	11	20	19	19	20	19	19	19	19	20	21

Table 2: Average Values for Individual Characteristics, Vulnerable vs. Entire Sample

Characteristic	Vulnerable	Entire Sample
Informal	0.95	0.61
Rural	0.95	0.40
Primary	0.86	0.31
Male	0.60	0.58
Head	0.47	0.47
Married	0.44	0.41
Self-employed	0.44	0.32
Age	36.96	39.29
Education	6.19	9.57

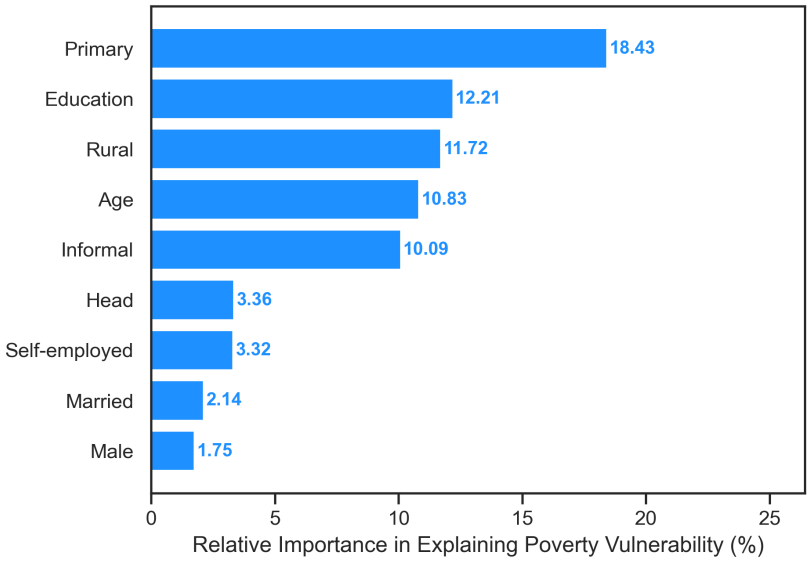
self-employed workers in the whole sample. Finally, we find that the vulnerable tend to be younger, and less educated.

The most relevant characteristics to predict poverty vulnerability are living in the rural/urban area, education level, working in the primary sector, informality status, and age. Figure 3 reports the variable importance of individual characteristics in explaining vulnerability. To compute variable importance, we follow the literature (Rodríguez-Pérez and Bajorath, 2019) by averaging the SHAP values across all individuals (in absolute value), and normalizing the outcomes so that the sum of variable importance adds up to 100.[8] Consistent with Table 2, Figure 3 illustrates that these five variables are the most salient characteristics of the vulnerable. Furthermore, the importance of the remaining variables is also in line with the mean difference in Table 2.

Figure 4 summarises the impact of the most relevant variables (as measured by the SHAP values) on poverty vulnerability. The dots in each figure depict the SHAP value of the corresponding variable for an individual in the sample, highlighting that the effect of each characteristic on the probability of poverty vulnerability is heterogeneous across

[8]The fact that education and age play a similarly important role than primary, rural and informal might arise from multicollinearity. For this reason, we complement the analysis with summary statistics and do not interpret individual variable importance.

Figure 3: Variable Importance of Individual Characteristics in Explaining Vulnerability



The importance of explaining poverty vulnerability in Figure 3 is computed according to the average SHAP values (in absolute value), and normalized so that the sum of variable importance adds up to 100.

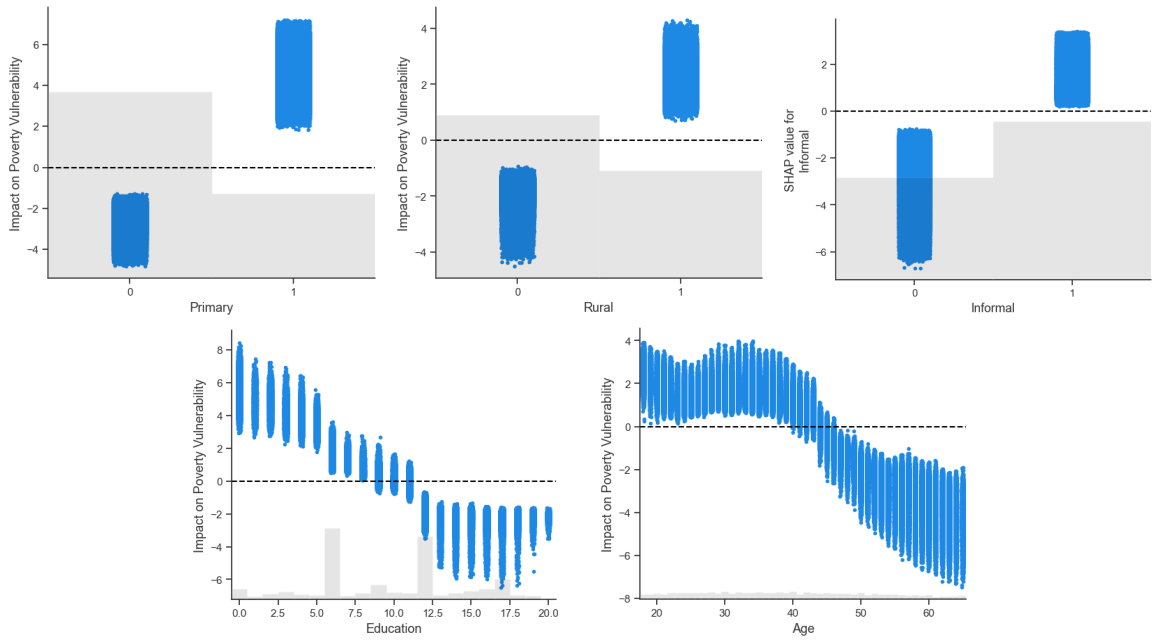
individuals. For instance, if working informally were to affect homogeneously individual vulnerability, all points would lie in the same value along the y-axis. However, we find that for every variable, the effect varies across individuals. In the case of the continuous variables (age and education), we find that the effect is not only heterogeneous for a given value but is also nonlinear.

The results in Figure 4 are in line with those of Table 2. In particular, we find that working in the primary sector, living in rural areas, and being an informal worker increase the probability of being vulnerable. Furthermore, older and more educated individuals are less likely to be vulnerable.

We find a negative relationship between education and vulnerability. Individuals who have completed at most primary schooling are more likely to be vulnerable, while those with secondary schooling or more are less likely to be vulnerable. Furthermore, we find that the relationship is non-linear, with a decreasing slope after two years of schooling, which turns flat for those having at least secondary schooling.

Individuals aged 25 to 45 are the most vulnerable, as indicated by the upward spikes on the graph. This age group may face challenges such as family responsibilities, unstable income, or job market pressures, which contribute to their vulnerability. Poverty vulnerability decreases steadily after the age of 45, with the trend moving closer to or below the zero impact line. Such a decline could be linked to increased financial stability, savings accumulation, or reduced household demands.

Figure 4: Effect of Individual Characteristics on Poverty Vulnerability by Informality Status



Note: the impact on poverty vulnerability is measured by the SHAP value for each of the five variables in the figure. The dots in each figure depict the SHAP value of the corresponding variable for an individual in the sample, highlighting that the effect of each characteristic on the probability of poverty vulnerability is heterogeneous across individuals.

Implementing a formalization policy can prevent poverty vulnerability. Moreover, our results suggest that this policy should prioritize low-educated individuals aged 25-45, working in the primary sector, living in rural areas in the Amazonian region of Ecuador. Our findings also suggest that long-term policies aiming to elevate the education level of those with less than high school can potentially reduce poverty vulnerability.

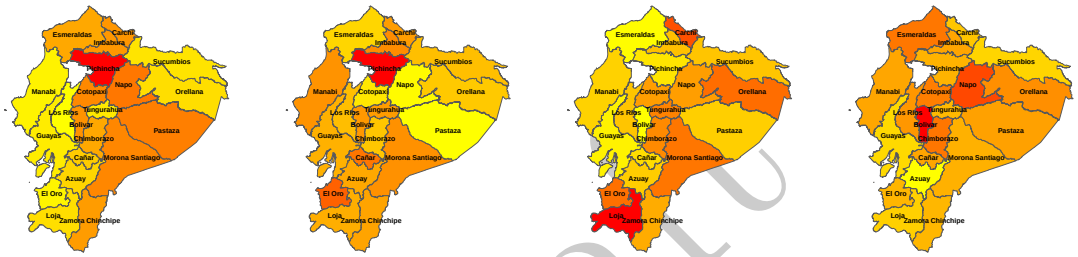
We now turn to estimating the heterogeneous impact of specific climate shocks on future income. We first identify the most affected provinces by each climate shock by focusing attention on the most extreme values of each climate variable. This is due to the fact that our definition of the climate variables captures low, moderate, and high levels of each. For instance, the lowest or average maximum temperature recorded in our data does not correspond to a climate shock. Similarly, a very mild flood or drought can be hardly considered a shock. Accordingly, to identify the most affected provinces by each climate shock we focus attention on the events that can be considered a shock.

For the case of extreme temperatures, we compute the average SHAP values for the four most extreme temperatures recorded in each province. Thus, for maximum temperatures, we only consider the highest four maximum temperatures by region, while for minimum temperatures, the lowest four. Similarly, we compute the average SHAP values for droughts and floods considering the tails of the distributions (the top quartile

for droughts, since they take negative values, and the bottom quartile for floods).

Figure 5 reports the impact of each climate shock on income by province. We find that the most affected regions by maximum temperatures are located in the center-east of the country. In particular, Pastaza, Pichincha, Azuay, and Morona Santiago emerge as the most vulnerable areas. In contrast to maximum temperatures, we find that minimum temperatures disproportionately affect the North and South regions. The most affected regions are located in the north and south of the Pacifical Coastal Region (Esmeraldas, and El Oro, respectively). However, we also find that minimum temperatures impact the northern part of the Andes and Amazonian region, since Napo Pichincha, and Orellana are the most affected apart from Esmeraldas and El Oro.

Figure 5: Impact of Specific Climate Shocks on Future Income by Province



(a) Maximum Temperature (b) Minimum Temperature (c) Floods (d) Droughts

The impact of each climate shock corresponds to the average SHAP value for the events that could be classified as shocks. For maximum temperatures, we only consider the highest four maximum temperatures by region, while for minimum temperatures, the lowest four. Similarly, we compute the average SHAP values for droughts and floods considering the tails of the distributions (the top quartile for droughts and the bottom quartile for floods)

Our findings suggest that Napo, Tungurahua, and Imbabura are the most affected provinces by extreme floods. Except for Napo and Zamora Chinchipe, we find that the Amazon Region is the least affected by extreme floods. Conversely, the Andes and Pacific Coastal regions experience the highest decrease in future income arising from floods. Regarding droughts, our analysis reveals that both the Andes and the Amazonian regions are heavily affected. The incidence of lack of precipitations is particularly pronounced in Imbabura, Pastaza, and Orellana. These findings highlight the heterogeneous nature of climate risks across regions, emphasizing the need for targeted and context-specific policy responses to mitigate both flood and drought impacts.

Figure 5 highlights the regions most affected by specific climate shocks. Imbabura ranks among the three provinces most affected by droughts and floods, underscoring the urgent need for preventive measures to mitigate vulnerability to excessive and deficient precipitation. Similarly, Pastaza stands out as highly vulnerable, being one of the provinces most affected by maximum temperatures and droughts.

5 Concluding Remarks

In this paper, we formalize and generalize the simulation approach of Hill and Porter (2017). We first show that their proposal involves out-of-sample prediction, which motivates the adoption of machine learning techniques to simulate future income. We then generalize their procedure to estimate functions of vulnerability and welfare outcomes.

To estimate unknown functions of vulnerability and welfare outcomes we draw from the explainable artificial intelligence literature, by computing SHapley Additive exPlanations (SHAP) values. This approach enables the characterization of the vulnerable population to poverty due to climate shocks. Furthermore, our proposal allows the identification of the regions with the greatest vulnerability rates. The former capability reveals the potential drivers of vulnerability, indicating who is the most vulnerable at the regional level, and the latter provides useful information for the design of place-based public policy.

This paper presents an easy-to-implement procedure to design targeted, place-based public policies to reduce climate risk. An important appeal of our proposal is that it can be applied to short panels and cross-sectional data. In our empirical application to Ecuador, we illustrate how our method provides useful information for the design of targeted place-based policies to prevent and mitigate climate risk.

References

Calvo, C. and Dercon, S. (2013). Vulnerability to individual and aggregate poverty. *Social Choice and Welfare*, 41(4):721–740.

Chaudhuri, S., Jalan, J., and Suryahadi, A. (2002). Assessing household vulnerability to poverty from cross-sectional data: A methodology and estimates from indonesia.

Chernozhukov, V., Escanciano, J. C., Ichimura, H., Newey, W. K., and Robins, J. M. (2022). Locally robust semiparametric estimation. *Econometrica*, 90(4):1501–1535.

Cred, U. (2020). Human cost of disasters. an overview of the last 20 years: 2000–2019. *CRED, UNDRR, Geneva*, 609.

Đekić, J. P., Mitković, P. B., Dinić-Branković, M. M., Igić, M. Z., Đekić, P. S., and Mitković, M. P. (2018). The study of effects of greenery on temperature reduction in urban areas. *Thermal Science*, 22(Suppl. 4):988–1000.

Fick, S. E. and Hijmans, R. J. (2017). Worldclim 2: new 1-km spatial resolution climate surfaces for global land areas. *International journal of climatology*, 37(12):4302–4315.

Günther, I. and Harttgen, K. (2009). Estimating households vulnerability to idiosyncratic and covariate shocks: A novel method applied in madagascar. *World Development*, 37(7):1222–1234.

Harris, I., Osborn, T. J., Jones, P., and Lister, D. (2020). Version 4 of the cru ts monthly high-resolution gridded multivariate climate dataset. *Scientific data*, 7(1):109.

Hawkins, D. M. (2004). The problem of overfitting. *Journal of chemical information and computer sciences*, 44(1):1–12.

Hill, R. V. and Porter, C. (2017). Vulnerability to drought and food price shocks: evidence from ethiopia. *World Development*, 96:65–77.

INEC (2023). Encuesta de condiciones de vida. Technical report.

Kleemann, J., Koo, H., Hensen, I., Mendieta-Leiva, G., Kahnt, B., Kurze, C., Inclan, D., Cuenca, P., Noh, J., Hoffmann, M., et al. (2022). Priorities of action and research for the protection of biodiversity and ecosystem services in continental ecuador. *Biological Conservation*, 265:109404.

Lundberg, S. (2017). A unified approach to interpreting model predictions. *arXiv preprint arXiv:1705.07874*.

Macintyre, H. and Heaviside, C. (2019). Potential benefits of cool roofs in reducing heat-related mortality during heatwaves in a european city. *Environment international*, 127:430–441.

McKee, T. B., Doesken, N. J., Kleist, J., et al. (1993). The relationship of drought frequency and duration to time scales. In *Proceedings of the 8th Conference on Applied Climatology*, volume 17, pages 179–183. Boston.

Pritchett, L., Suryahadi, A., and Sumarto, S. (2000). Quantifying vulnerability to poverty: A proposed measure, with application to indonesia.

Rodríguez-Pérez, R. and Bajorath, J. (2019). Interpretation of compound activity predictions from complex machine learning models using local approximations and shapley values. *Journal of medicinal chemistry*, 63(16):8761–8777.

Shapley, L. S. (1953). A value for n-person games. *Contribution to the Theory of Games*, 2.

Skoufias, E. and Quisumbing, A. R. (2003). Consumption insurance and vulnerability to poverty. Technical report, International Food Policy Research Institute (IFPRI).

Williams, C. K. and Rasmussen, C. E. (2006). *Gaussian processes for machine learning*, volume 2. MIT press Cambridge, MA.

World Bank (2014). Global facility for disaster reduction and recovery.

Appendix

Consider the decomposition of consumption into observed and unobserved components as

$$C = m\left(\boldsymbol{X}, \boldsymbol{S}; \boldsymbol{\eta}\right) + \epsilon, \quad \mathbb{E}\left[\epsilon | \boldsymbol{X}, \boldsymbol{S}\right] = 0,$$

where $\boldsymbol{X}$ and $\boldsymbol{S}$ are individual characteristics and climate shocks, respectively, $m\left(\cdot\right) \coloneqq \mathbb{E}\left[C|\boldsymbol{X}, \boldsymbol{S}\right]$ is the conditional mean function, known up to the nuisance parameter $\boldsymbol{\eta}$, and ϵ is an idiosyncratic shock, corresponding to the unexplained component. Assuming linearity of $m\left(\cdot\right)$, the nuisance parameter correspond to the regression coefficients. Since our proposal involves estimating $m\left(\cdot\right)$ non-parametrically, we now characterize it for the case of tree-based algorithms and kernel machines.

Trees form a partition[9] of the covariate space into J regions $\boldsymbol{R} \coloneqq (R_1, ..., R_J)$, by finding the J regions and coefficients $\boldsymbol{C} \coloneqq (C_1, ..., C_J)$ and solving the optimization OLS problem (for a fixed value of R)

$$\arg\min_{c} \frac{1}{n} \sum_{i=1}^{n} \left(Y_i - m\left(\boldsymbol{X}_i, \boldsymbol{S}_i, \eta\right)\right)^2, \quad m\left(\boldsymbol{X}_i, \boldsymbol{S}_i, \boldsymbol{\eta}\right) = \sum_{j=1}^{J} c_j 1\left(\boldsymbol{X}_i, \boldsymbol{S}_i \in R_j\right),$$

for the training set,[10] where the solution to the above optimization problem is given by $\hat{\boldsymbol{C}} = (\hat{c}_1, ..., \hat{c}_J)$ ($\hat{c}_j = \overline{Y}_j$ is the average outcome in the R_j region). Thus, for tree-based algorithms, we have

$$m\left(\boldsymbol{X}_i, \boldsymbol{S}_i, \boldsymbol{\eta}\right) = \sum_{j=1}^{J} c_j 1\left(\boldsymbol{X}_i, \boldsymbol{S}_i \in R_j\right),$$

and $\eta = (\boldsymbol{C}, \boldsymbol{R})$. As regards the estimation of $m\left(\cdot\right)$, any ensemble of trees can be used, e.g., bagging, boosting, or random forest.

Kernel machines, on the other hand, solve the optimization problem

$$\arg\min_{m \in \mathcal{H}, ||m||_{\mathcal{H}} \leq k} \frac{1}{n} \sum_{i=1}^{n} \left(Y_i - m\left(\boldsymbol{X}_i, \boldsymbol{S}_i, \boldsymbol{\eta}\right)\right)^2,$$

for the training set, where $\mathcal{H}$ is a reproducing kernel Hilbert space (RKHS). Correspond-

[9] A partition of a set is a grouping of its elements into subsets that are not empty, such that every element is included in exactly one subset. For example, $\mathcal{P} = \{[0, 0.5), [0.5, 1]\}$ is a partition of the set $[0, 1]$.

[10] Machine learning algorithms split the data into training and test sets. The training set is the subset of the data used to train a model and the test set is the subset of the data used to test the trained model.

ing to each $\mathcal{H}$ there exists a unique positive semidefinite symmetric kernel function $K\left(x_1, s1;, x_2, s_2\right)$ with the representation

$$K\left(x_1, s_1;, x_2, s_2\right) = \sum_{j=0}^{\infty} \alpha_j, \phi_j(x_1, s1)\phi_j(x_2, s2)$$

for a positive sequence of numbers α_j and linearly independent functions ϕ_j such that each element $m \in \mathcal{H}$ has the form

$$m\left(\boldsymbol{X}_i, \boldsymbol{S}_i, \boldsymbol{\eta}\right) = \sum_{j=0}^{\infty} m_j \phi_j(\boldsymbol{X}_i, \boldsymbol{S}_i), \quad \sum_{j=0}^{\infty} \frac{m_j}{\alpha_j} < \infty.$$

As regards estimation, the kernel machine optimization problem has the following closed-form solution:

$$\hat{\mu}\left(x, s\right) = \sum_{i=1}^{n} \left(\boldsymbol{K} + \lambda_n I_k\right)^{-1} K(x, s; \boldsymbol{X}_i, \boldsymbol{S}_i), \tag{16}$$

where $\boldsymbol{K}$ is the $n \times n$ matrix with ij-th element $K(\boldsymbol{X}_i, \boldsymbol{S}_i; \boldsymbol{X}_j, \boldsymbol{S}_j)$.[11]

A common choice of the kernel is the squared exponential Kernel (Williams and Rasmussen, 2006), which has the form

$$K(x_1, x_2) = \sigma_f^2 \times \exp\left\{-\frac{|x_1 - x_2|^2}{2l^2}\right\}.$$

Thus, for the squared exponential kernel we have that

$$m\left(\boldsymbol{X}_i, \boldsymbol{S}_i, \boldsymbol{\eta}\right) = \sum_{j=0}^{\infty} m_j \phi_j(\boldsymbol{X}_i, \boldsymbol{S}_i),$$

and $\boldsymbol{\eta} = \left(\lambda_n, \sigma_f^2, l^2\right)$.

[11] Notice that equation (16) illustrates why kernel machines are a generalization of ridge.