使用可解释提升算法对环境数据和农业数据建模

Fabia n Obster[1, 2], Christian Heumann[2], Heidi Bohle[3], 和 Paul Pechan[3,*]

[1]慕尼黑联邦国防军大学，工商管理系，纽比贝格，85577，德国[2]慕尼黑路德维希-马克西米利安大学，统计系，city，80539，德国[3]慕尼黑路德维希-马克西米利安大学，传播与媒体研究所，80539，德国[*]paul.pechan@ifkw.lmu.de

arXiv:2305.02699v1 [stat.ML] 4 May 2023

摘要

我们介绍如何使用基于岭正则化广义线性模型的可解释提升算法分析高维环境数据。我们通过运用环境、社会、人类和生物物理数据，预测智利和突尼斯农民面对气候灾害时的财务脆弱性，以此加以说明。我们展示了如何考虑组结构，以及如何利用一种新颖的两步提升方法，在高维数据集中发现交互作用。本文展示并讨论了所提方法的优势与有效性。结果表明，只有在纳入两步提升时，交互效应才会提高预测能力。预测各类脆弱性时，最重要的变量是自然资产。其他重要变量包括灌溉类型、经济资产，以及附近农场是否存在作物损害。

引言

在这项研究中，我们展示了如何利用可解释提升算法，根据环境因素以及人为、社会和生物物理因素及其交互作用，预测多种危害下的财务漏洞。为寻找交互作用，我们提出了一种基于两步提升的新方法，该方法仍具可解释性，并可与分量式提升结合使用。我们运用变量重要性、效应量和偏效应等可解释性工具，更好地理解可能导致这些气候变化相关漏洞的潜在因素。

基于模型的提升算法已用于环境科学领域的多种用途。例如，根据土壤样本量化多种土壤参数[1]，根据环境因素预测农民的财务状况[2]，以及根据气候变量预测动物园游客人数[3]。此外，基于分类树或回归树等不可解释提升算法（如Adaboost）[4] 也因其较强的预测能力而被用于基于环境数据进行环境预测。其应用包括滑坡易发性评估[5] 以及根据环境因素预测幼年海鳟的出现情况[6]。

通过提出的提升模型，我们希望实现以下目标：

- **预测能力**：模型不仅应对已分析的数据拟合良好，也应能良好地拟合同一领域中的未见数据，前提是变量的分布相似。
- **可解释性**：我们关注哪些变量与结果变量相关这一问题。但我们也想了解这些关联呈现出怎样的形式。在农学情形下，我们希望制定行动，以降低对不利环境变化的脆弱性。只有了解适应性措施的效应，才能做到这一点。只有在了解这些关联后，才能提出因果假设，并通过新的特定实验对其进行检验。我们还希望在保留模型能力的同时，尽可能简单地对这些效应进行建模。线性效应应优先于非线性效应和交互效应。在这种情况下，应避免使用黑箱。
- **稀疏性**：我们考虑高维数据集，其中变量的数量p相对于观测值n的数量较大；如果考虑交互作用的情况，变量数量甚至可能更多。在众多可能的变量中，我们想知道哪些变量确实与结果变量相关，哪些无关。因此，模型应执行变量选择，以实现稀疏性。目标是找到仍具有较高预测能力的最小变量子集。稀疏性还能提升可解释性，因为科学家和利益相关者只需关注真正相关的变量，便可忽略不重要的变量。在脆弱性情境下，这意味着农民可以关注灌溉系统类型等入选变量，而不是金融适应性措施等未入选变量。

• **复杂性**：模型应尽可能简单，同时满足必要的复杂性。复杂性是平衡前面所有要点的特征。在两个预测能力相同的解释中，模型应选择更简单的那个。这里所说的更简单，是指更加稀疏、更具可解释性且不包含交互作用。另一方面，我们也不希望忽视非线性和交互作用等重要的复杂性。识别某些变量是否会受到其他变量影响至关重要。此外，还可能存在非层级交互作用，即某个变量本身对结果变量没有影响，但在数据的一个子集中可能产生正向影响，而在另一个子集中则产生负向影响。例如，在某个地区，作物多样性较高可能会增加脆弱性，而在另一个地区则可能降低脆弱性。

• **组结构** 数据中的变量可以聚类成组。“气候变化经历”就是一个例子，它包含二元变量“气温升高”“干旱加剧”“极端天气增多”和“降雨减少”。问题在于，结果变量是受到每个单独变量的影响，还是只受到其中部分变量的影响，抑或是这些变量因彼此相似而作为一个组共同发挥作用。组结构也能提升可解释性，因为对人类而言，理解某个抽象概念的总体影响，往往比逐一审视其各个方面更容易。

有许多方法可以处理上述每一项要求。例如，可以通过套索回归实现稀疏性[7] 或提升套索[8]，可以通过种类繁多的模型实现预测能力，而组结构则可以通过稀疏组套索纳入[9]。

在本研究中，我们重点探讨如何利用提升算法实现这些目标，具体包括分量式提升（mb）、含交互作用的分量式提升（mb int）、稀疏组提升（sgb）以及用于交互作用的两步提升（2-boost）。我们比较这些算法的预测能力、效应大小，以及变量/组的相对重要性。下文将介绍用于分析的方法，并讨论如何通过对通用提升算法进行修改，利用这些方法实现上述目标。

方法

数据介绍

数据通过焦点小组会议以及对智利的 400 名樱桃种植户和突尼斯的 401 名桃种植户进行面对面问卷访谈收集。入选条件是：农民负责经营农场，且农业活动收入占其总收入的 70% 以上。为收集数据，我们向本机构（ICMR、LMU）咨询了涉及人类受试者参与的问卷实施和数据使用事宜。农场数据按照各国适用的数据收集程序收集。问卷参与者均已就数据收集提供知情同意。我们未收集任何可识别个人身份的数据，以确保完全匿名。数据的描述性概述[10] 以及基于类似数据开展的脆弱性混合方法分析[11] 均已完成。分析所用的 R 代码可在 https://github.com/FabianObster/boostingEcology 找到。

自变量

分析的变量可归为若干组，包括

- 气候体验组（气温升高、降雨减少、干旱加剧、极端天气增多）
- 自然资产组（地理区域）
- 社会资产组（对信息的依赖／使用、对信息来源的信任、社区、科学或宗教）
- 人力资产组（年龄、性别、教育程度）
- 生物物理资产组（农场规模、农场使用的水资源管理系统、所种作物的多样性、适应性措施）
- 经济资产类别（农场债务、农场绩效、对果园收入的依赖）
- 目标组（保持传统延续，独立工作）
- 危害组（气候威胁农场，乐观）
- 空间组（农场附近的作物损害，该国农场的作物损害）

所有变量及其所属组的概览可见于表格4 和 5。此外，还可以找到每个类别中的农民人数。

结果变量

结果变量用于衡量农民根据自我评估所反映的、针对五种气候灾害的财务脆弱性：冬季气温升高、夏季气温升高、降雨量减少、干旱加剧以及极端天气增多。对于每一种气候灾害，均定义一个二元变量，用于表示农民是否容易受到该灾害的影响，并将其作为结果变量。主要类别包括财务上不脆弱的农民，参照类别则包括财务上脆弱的农民。各类别中的农民人数见表 3。

交互变量

22 个变量被用作可能与其他变量对结果变量产生交互效应的变量。交互变量包括地区和社会人口学变量等，并在表格中以粗体标出 4和5。加上所有其他变量后，共有 1366 个交互项，以及 4000 多个待估计的模型参数。由于数据中有 801 位农民，寻找交互作用会导致“p>n”问题，即设计矩阵中的变量数量超过了观测值数量。

总体设置、模型构建与评估

所有分析均使用 R[12] 完成，提升模型则使用软件包“mboost”[13]进行拟合。由于所有结果变量都是二元变量，我们采用二项分布的岭惩罚负对数似然作为损失函数，并使用逻辑连接函数，由此得到

$$h(\beta, X_i) = P(y_i == 1) = \frac{1}{1+\exp\left(-X_i^T\beta\right)},$$

$$l(y,h) = -\left[\sum_{i=1}^{n} y_i \log\left(h(\beta,X_i)\right) + (1-y_i)\log\left(1-h(\beta,X_i)\right)\right] + \lambda \left\|\beta\right\|_2^2.$$

进行任何分析之前，先将数据划分为 70% 的训练数据和 30% 的测试数据，测试数据仅用于最终评估。完成预测分析后，使用全部数据计算变量重要性和偏效应。模型评估基于受试者工作特征曲线（ROC）下面积，并使用测试数据计算。ROC 曲线下面积（AUC）通过考察预测模型计算出的预测概率的所有可能阈值，同时考虑真阳性率和假阳性率。

在分析中，我们针对不同目的使用了多种提升模型。所有提升模型均使用 R 软件包“mboost”[14]进行拟合。对于提前停止，在每一步提升过程中均通过十折交叉验证确定停止参数。第一个也是最简单的模型是逐分量模型式提升（mb），该模型对所有变量采用岭正则化线性效应，使得所有自由度均等于 1。该模型使我们能够进行变量选择，并比较所有变量的相对重要性。对于第二个模型，我们使用了混合参数为 $\alpha = 0.5$ 的稀疏组提升，该参数用于平衡组选择与个体变量选择。通过这种方式，可以判断变量是单独对结果变量具有重要影响，还是更倾向于作为变量组发挥作用。

为了在数据中找到交互作用，我们采用两种方法。第一种是标准方法，即在每次迭代中同时定义线性效应和交互效应。随后，模型可以决定选择主效应还是交互效应。第二种方法采用两阶段提升模型。第一步使用已经拟合好的 mb 模型，该模型仅使用单个线性基学习器。第二步则只使用交互作用。由于线性基学习器先行拟合，这种方式会优先考虑线性基学习器，而非交互基学习器。

方法部分的其余内容更具技术性，以应用为导向的读者可以跳过。

通用提升算法

我们将从提升算法的一般表述开始介绍，该算法也可以描述为函数梯度下降算法。目标是找到一个函数 f^*，使其最小化某个损失函数 $l(y,\ f)$。这里，我们只考虑可微凸损失函数。损失函数有两个参数。第一个参数 $y \in \{1,\ldots,n\}$ 是结果变量，其中包含 n 个观测值。第二个参数 f 是一个实值函数 $f\colon\ \mathbb{R}^{n\times p} \mapsto \mathbb{R}$，它是数据 $X\ \in \mathbb{R}^{n\times p}$的函数。

另一种拟合稀疏回归模型的方法是采用提升法。其拟合策略是通过向给定模型中添加一个基学习器，逐步改进该模型。在本文中，我们将基学习器定义为设计矩阵中与某个实值函数相关联的列子集。为了保证稀疏性，每个基学习器在每一步仅考虑变量的一个子集[15]。在逐分量 $\mathscr{L}^2$ 提升法中，每个变量都将作为一个基学习器。对于一维 B样条，基学习器是表示 B样条基函数的设计矩阵。一般而言，提升法的目标是找到一个实值函数，使其能够最小化通常可微且凸的损失函数 $l(\cdot,\cdot)$。

这里我们将负对数似然视为损失函数，以估计 f^* as

$$f^*(\cdot) = \arg\min_{f(\cdot)} \mathbb{E}[l(y,f)].$$

通用函数梯度下降算法[16]

1. 定义结构为h的基学习器： $\mathbb{R}^{n\times p} \to \mathbb{R}$
2. 初始化$m = 0$ 和 $\widehat{f}^{(0)} \equiv 0$ 或 $\widehat{f}^{(0)} \equiv y$ –
3. 设置$m = m+1$，并计算负梯度$\frac{\partial}{\partial f} l(y,f)$，然后在 $\widehat{f}^{[m-1]}$处对其求值。由此得到伪残差$u1,\ldots,u_n$，其中

$$u_i^{[m]} = \frac{\partial}{\partial f} l(y_i,f)|_{f=\widehat{f}^{[m-1]}},$$

对于所有$i = 1,\ldots,n$

4. 使用响应值为$(u_1^{[m]},\ldots,u_n^{[m]})$的数据拟合基学习器$h$。由此得到 $\widehat{h}^{[m]}$，它是负梯度的近似值

5. 更新

$$\widehat{f}^{[m]} = \widehat{f}^{[m-1]} + \eta \cdot \widehat{h}^{[m]}$$

这里的 η 可以视为学习率，其中 $\eta \in]0,\ 1[$

6. 重复步骤 3、4 和 5，直到 $m = M$

增强岭回归

损失函数 $l(\cdot,\cdot)$ 可以设为任意函数。对于可解释提升法，负对数似然是一个合理的选择。还可以使用岭惩罚项修改对数似然。通过引入超参数 $\lambda > 0$，可以修改损失函数 l。设 h 是参数向量 $\beta \in \mathbb{R}^p$ 和设计矩阵 $X \in \mathbb{R}^{n\times p}$的函数，那么

$$l_{\text{Rige}}(u,h) = l(u,h) + \lambda \|\beta\|_2^2$$

是岭惩罚损失函数。通过增大 λ，可以将参数向量 β 向零收缩。与λ 密切相关的是自由度。令S 为命题3中所示的近似广义岭帽子矩阵，见[17]。我们指出，在普通最小二乘岭回归这一特殊情形下，有$S = X(X^TX + \lambda I)^{-1}X^T$。通常，自由度可定义为

$$\text{df}(\lambda)) = \text{tr}(2S - (S)^TS)).$$

建议为每个基学习器设置正则化参数，使所有基学习器的自由度相等。因此，正则化参数使得我们能够同时使用多项式效应等复杂基学习器，以及线性效应等简单效应。由于复杂基学习器受到的正则化程度高于简单基学习器，因此可以优先选择简单且更具可解释性的基学习器，而非复杂基学习器，从而引入一种倾向于可解释性的归纳偏置，正如我们在问题陈述中所要求的那样。

分量式提升和分组逐分量提升

在通用函数梯度下降算法的第 4 步中，应用函数h。除了使用单个函数外，也可以使用一组R 个函数 $\{(h_r)_{r\le R}\}$。然后，第 5 步仅使用对数据应用后损失函数最低的那个函数进行更新，即$r^* = \arg\min_{r\le R} \mathbb{E}[l(u,h_r)]$。在分量式提升中，对于数据集中的每个变量，都会使用一个仅以该变量为自变量、而不涉及其他变量的函数。这样，每一步只会选取一个变量。随后，通过提前停止，或将M 设置得相对于数据集中变量的数量较小，可以拟合出一个整体稀疏模型。这满足了问题陈述部分中的稀疏性要求。对于分组变量，也可以将基学习器定义为变量组，使其仅由属于某一组的变量构成。这些变量可以是社会学数据[18] 中属于某个特定构念的所有题项变量，也可以是农业和环境数据[2]中的所有气候变化相关变量。这样便可进行分组变量选择，即只选择部分组，从而开展以组/构念为中心的分析，而非以单个变量为基础的分析。这样就能将组结构纳入考虑。

稀疏组提升

It is an 同时使用个体基学习器和基于组的基学习器。这样，在每一步中，要么选择一个个体变量，要么选择一组变量。借鉴稀疏组套索中的类似思想[9], ，可以定义稀疏组提升[19]。我们再次通过修改自由度来实现这一点。每个变量都将获得其自身的基学习器，而每组变量将获得一个基学习器，其中包含该组的所有变量。令G 表示组的数量，p_g 表示组g中的变量数量。那么，对于个体基学习器$x^{(g)}{}_j \in \mathbb{R}^{n\times 1}$ 的自由度，我们将使用

$$\mathrm{df}(\lambda_j^{(g)}) = \frac{1}{p_g}\cdot\alpha.$$

对于组基学习器，我们使用

$$\mathrm{df}(\tilde{\lambda}^{(g)}) = \frac{1}{p_g}\cdot(1-\alpha).$$

混合参数 $\alpha \in [0,1]$ 允许在选择过程中调整组与个体变量之间的优先级。如果 $\mathrm{df}(\lambda) = 0$ 表示 $\lambda \to \infty$， $\alpha = 1$ 则得到分量式提升，而 $\alpha = 0$ 则得到组提升。

两步提升

在通用提升算法中，同一组函数依次应用于数据。尽管这组函数内部会进行变量选择，但在提升过程中，这组函数本身不会改变。我们介绍对通用算法的一种改进，使其更加灵活，即两步提升。遗传研究中曾采用一种类似的两步提升思路，称为分层提升[20] ，该方法也用于迁移学习[21], 以及深度学习应用[22]。在大多数情况下，当结果变量由层级类别结构构成时，会使用分层提升[23]。与文献中分析的数据不同，我们在此分析的数据不包含层级类别结构。因此，我们不像文献中大多数情况那样使用分层提升，而是将其用于检测层级和非层级交互作用。我们对两步提升进行形式化并加以推广。令 K 为步骤数，对于每个步骤 $k \leq K$ ，令 H_k 为基学习器的集合。

***K*步提升算法**

1. 将K 设为步数
2. 对于每一步 $k \leq K$ ，定义要使用的基学习器集合 H_k ，并将 M_k 设为提升迭代的次数。
3. 初始化 $m_0 = 0$ 和 $\hat{f}^{(0)} \equiv 0$ 或 $\hat{f}^{(0)}$4. 对于 $k \leq K$ 重复执行:
5. 对 $m_k \leq M_k$ 执行通用提升算法的步骤 2-6。6. 设置初始化 $m_k = 0$ 和 $u^{[0]} = u^{[M_{k-1}]}$

有人可能会问，如果在原始方法中可以直接并行使用更多基学习器，为什么还需要依次运行多个提升算法？先前的研究表明，某些步骤组合能够实现较高的预测能力。然而，正如我们在问题陈述中所描述的，预测能力只是我们要求的一部分，而且如果这种能力的提升是以牺牲可解释性和对数据的理解为代价的，那么这种提升未必值得追求。此外，该算法的顺序执行特性限制了并行化所带来的计算效率提升，因为并非所有基学习器都能在同一次提升迭代中并行拟合。K 步提升算法也可以被视为一般提升算法的一种特殊情况，其中基学习器本身就是提升算法。

变量重要性

对于前文所述的每种提升法，都可以计算变量重要性度量。在每一步中，都会计算对数似然，这意味着可以计算该步骤中所选基学习器带来的对数似然降幅。完成每个基学习器的拟合后，可以计算似然的总降幅。这样，无论基学习器的类型如何，都可以计算归因于各个基学习器的负对数似然降幅百分比。变量重要性让我们能够比较各个变量之间的相对重要性，这与显著性或p值的概念不同；后者是在一组假设的基础上检验参数不为零的假设。因此，某个变量在提升法中可能很重要，但按照经典回归方法却不显著，反之亦然。

部分效应和效应量

对于提升广义线性模型，可以计算部分效应[13]。与经典逻辑回归类似，可以通过首先对某个基学习器的所有线性预测量求和，来计算所有基学习器的优势比。随后，可以按照逻辑回归中效应量的类似方式解释这些优势比。基于线性预测量，如果将所有其他基学习器设为平均值，还可以计算变量各类别的预测概率。通过这种方式，可以绘制偏效应图，既包括单个变量的基学习器，也包括交互基学习器。因此，与只能推导出事后解释的其他机器学习模型相比，基于模型的提升模型本身就具有可解释性。此外，还可以跟踪每次提升迭代中被选中的变量，从而理解模型的内部工作机制。

应用

可预测性

参见表 1 和图 1 可知，除采用并行估计的交互作用模型外，各提升模型的曲线下面积值相近。对五种脆弱性的曲线下面积值取平均后，sgb 为 0.752，mb 和 2-boost 均为 0.745，mb int 为 0.613。对于降水和干旱脆弱性，并行估计交互作用时没有选出任何变量，因此曲线下面积为 0.5。在 2-boost 中，第二次估计也没有选出任何变量，因此得到的模型与 mb 相同；对于这些结果变量脆弱性，mb 的曲线下面积最高。对于夏季温度脆弱性，sgb 的曲线下面积最高；对于冬季气温和极端天气，2-boost 的曲线下面积最高。比较各项脆弱性的可预测性可知，利用给定变量对降雨减少所致脆弱性的预测效果最好，其次依次是极端天气加剧、干旱加剧、冬季温度升高和夏季温度所致的脆弱性。

个体变量和组的重要性

比较图2中稀疏组提升（sgb）和分量式提升（mb）的变量重要性，可以明显看出，尽管两者有一些重叠，但也存在一些不同的变量。对于所有结果变量而言，最重要的单个变量都是“自然资产”，它表示农场的四个地区。然而，对于五种脆弱性，sgb 中自然资产的相对重要性均高于 mb。与其他脆弱性相比，组似乎对解释气温升高脆弱性更为重要，因为经济资产类别是夏季温度脆弱性的第二重要变量，而目标组是冬季温度脆弱性的第二重要变量。空间组是降雨量减少脆弱性的第三重要变量，但与最重要变量相比，其相对重要性较低。

交互作用的重要性

在可预测性部分，我们已经看到，两步估计和并行估计在交互效应方面存在一些差异。在进行预测时，模型评估仅使用基于训练数据训练的模型，并在测试数据上进行。对于图3 中的变量重要性和表2 中的稀疏性分析，则使用了全部数据。并行估计仅选择了交互效应，而未选择主效应（个体变量）；两步估计则两者都选了。

参见表 2 可以明显看出，变量选择存在显著差异。总体而言，2-boost 的两步估计得出的交互作用少得多。对于夏季温度脆弱性，没有选择任何交互项；而并行估计则选择了 13 个交互效应。对于降雨量减少脆弱性，差异同样显著。两步估计仅选择了一个交互作用，即农艺措施与对电视的信任之间的交互作用，而 mb int 选择了 48 个。对于干旱脆弱性，差异最小：并行估计选择了 27 个交互作用，两步估计则选择了 16 个。2-boost 选出的交互作用中，五分之四的占比低于百分之一；而 mb int 的这一比例则有五分之四高于百分之一。

不仅稀疏性存在差异，所选的交互作用本身也有所不同。参见图 3，对于冬季温度脆弱性，基于两个模型，“自然资产”–“成为盈利性企业”和“国家”–“农场债务负担”这两个交互作用的相对重要性都很高。但除此之外，两个模型选出的交互作用几乎没有重叠。例如，对于降雨量减少脆弱性，唯一选出的“农艺措施”–“对电视的信任”交互作用在 2-boost 中的相对重要性为 1，但未被 mb int 选中；而 mb int 则选中了另外 48 个交互作用。
在图 4、5、6、7和8 中，我们根据仅使用一个交互项的经典逻辑回归，绘制了 mb int 和 2-boost 中针对每种脆弱性发现的四种最重要的交互效应。图中根据交互项的联合类别绘制了不存在脆弱性的概率。我们分别针对智利、突尼斯的数据以及全部数据进行了分析。举例来说，我们解读了两种模型都选出的冬季温度脆弱性中两个共同的交互效应：“自然资产”与“成为盈利性企业”，以及“国家”与“农场债务负担”。在智利北部地区，与没有成为盈利性企业这一目标相比，拥有这一目标与较高的免受冬季气温升高影响的概率相关。在智利南部地区，这种关联恰好相反，也就是说，拥有

与没有成为盈利性企业这一目标相比，拥有这一目标与较低的免受冬季气温升高影响的概率相关。在突尼斯的两个地区，拥有成为盈利性企业这一目标都与负相关联，但南部地区的负相关程度高于北部地区。根据“国家”与“农场债务负担”这一交互项，较高的农场债务负担与较高的免受冬季气温升高影响的概率正相关，而在突尼斯，这种关联为负。在智利，北部地区的正相关更强；在突尼斯，南部地区的负相关更强。

讨论

结果表明，可以利用分析中采用的变量及变量组，通过可解释提升算法及其变体来预测智利和突尼斯农民面临气候灾害的脆弱性。所有模型都进行了变量选择。无论采用哪种提升方法，对于降雨减少导致的脆弱性，曲线下面积（AUC）所衡量的预测能力最高；对于夏季气温升高导致的脆弱性，预测能力则最低。在预测夏季温度脆弱性时，稀疏组提升的表现优于所有其他模型，这表明，影响可能由潜在变量而非单个变量造成。组变量重要性主要指向经济和生物物理资产，包括适应性措施；这些因素可能是夏季温度脆弱性的潜在决定因素。变量重要性强烈指向自然资产，其由智利和突尼斯的四个不同地区构成，是各种脆弱性的主要决定因素。这表明，国家内部及国家之间存在显著差异。交互作用分析也证实了区域性的重要性，因为某些效应受到国家和南北比较的显著调节。债务负担受到地区调节的效应，可能表明地区间存在经济差异，或与大城市的距离不同；也可能是不同气候带所致。

尽管从单变量交互作用分析中可以看出，数据中存在显著的交互效应，但在高维情境下，要将这些交互效应转化为更强的预测能力并非易事。这一点显而易见：在所有情况下，除主效应外还纳入交互作用基学习器的模型，表现都不如不含交互作用的同一模型。原因之一可能是过拟合，因为需要估计的参数数量是变量数量的四倍多。因此，交互作用模型没有纳入任何主效应，只纳入了交互作用。我们认为，在高维数据中，过拟合问题会比纯粹随机的情况更具系统性，因为如果数据集中有 p 个变量，那么就有 $\mathcal{O}(p^2)$个可能的交互项。因此，随着 p 增大，无论实际效应量如何，选择交互项而非主效应的概率都会上升。所提出的两步提升方法可以有效解决这种隐含的交互作用选择偏差。两步提升带来了更强的预测能力和更高的稀疏性，最终模型中包含的交互作用也更少。因此，我们认为，相较于在提升法的主模型公式中纳入交互项进行“经典”的并行估计，这种方法更优。我们认为唯一的缺点是必须估计两个模型，而不是一个，这略微增加了编程工作量；此外，由于模型是按顺序而非并行拟合的，也降低了进一步并行化的可能性。不过，如果要进行交互作用分析，按照惯例并遵循稀疏性原则，通常总是先拟合一个仅包含单个变量的模型[24]。在这种情况下，两步提升在计算上也更高效，因为可以基于第一个模型继续建模，避免重新拟合主效应。

在环境研究中，要在高维数据集中持续发现关联，需要采用新方法来推动知识发展。这些新方法具有更大的灵活性，但往往要以牺牲经典统计推断为代价，包括无法像提升法那样获得p值和标准误估计[25]。

对于一种现象，往往存在多种合理的解释。本文提出的方法能够直接比较大量解释，并估计它们对结果变量的解释重要性。通过收集所有可能与结果变量相关的变量并对其进行观测抽样，这种方法可以加快对现象的理解，尤其是对气候变化等较新的现象。从相对较小的样本量着手，研究者可以估计各项假设的相对重要性，并根据结果确定未来研究的优先次序。

采用先验可解释方法（例如前文所述的方法），其一大优势是能够评估给定的一组解释对结果变量的可预测性。相比之下，将事后可解释性工具应用于黑箱，只能简化说明黑箱预测可能是如何得出的，却无法评估这些解释本身对结果变量的预测效果如何。

资金支持

本研究是在“气温升高的物候与社会影响——气候变化对突尼斯、智利和德国水果生产的影响”项目（PASIT；德国联邦教育与研究部资助，项目编号 031B0467B）框架内开展的。LMU 提供了开放获取出版资助。dtec.bw 还提供了额外资助

由 NextGenerationEU 资助。

参考文献

1. Li, B., Chakraborty, S., Weindorf, D. C. & Yu, Q. 基于模型的提升法进行数据整合。《SN Comput. Sci.》2, 400，DOI：10.1007/s42979-021-00797-0（2021）。2. Obster, F., Bohle, H. & Pechan, P. M. 在预测水果农场的财务经营状况时，气候变化以外的因素目前更为重要，DOI：10.48550/arXiv.2301.07685（2023）。ArXiv：2301.07685 [cs, stat]。3. Obster, F., Brand, J., Ciolacu, M. & Humpe, A. 利用随机森林改进提升广义加性模型：面向智慧旅游的动物园游客案例研究。《Procedia Comput. Sci.》217, 187–197，DOI：10.1016/j.procs.2022.12.214（2023）。4. Freund, Y. & Schapire, R. E. 在线学习的决策理论推广及其在提升法中的应用。《J. Comput. Syst. Sci.》55, 119–139，DOI：10.1006/jcss.1997.1504（1997）。5. Jennifer, J. J. 滑坡易发性制图中的特征消除与机器学习算法比较。《Environ. Earth Sci.》81, 489，DOI：10.1007/s12665-022-10620-5（2022）。6. Froeschke, J. T. & Froeschke, B. F. 基于环境因素的时空预测模型：利用提升回归树预测得克萨斯河口幼年斑点海鳟。《Fish. Res.》111, 131–138，DOI：10.1016/j.fishres.2011.07.008（2011）。7. Tibshirani, R. 基于套索的回归收缩与变量选择。《J. Royal Stat. Soc. Ser. B (Methodological)》58, 267–288（1996）。8. Zhao, P. & Yu, B. 提升套索。技术报告，加州大学伯克利分校统计系（2004）。章节：技术报告。9. Simon, N., Friedman, J., Hastie, T. & Tibshirani, R. 稀疏组套索。《J. Comput. Graph. Stat.》（2013）。10. Pechan, P. M., Obster, F., Marchioro, L. & Bohle, H. 气候变化对智利和突尼斯水果农场运营的影响。《agriRxiv》2023, 20230025166，DOI：10.31220/agriRxiv.2023.00171（2023）。出版方：CABI International agriRxiv。11. Pechan, P. M., Bohle, H. & Obster, F. 降低果园对气候变化的脆弱性。《agriRxiv》2023, 20230025167，DOI：10.31220/agriRxiv.2023.00172（2023）。出版方：CABI International agriRxiv。12. Team, R. RStudio：R 的集成开发环境（2020）。13. Hofner, B., Mayr, A., Robinzonov, N. & Schmid, M. R 中基于模型的提升法：使用 R 软件包 mboost 的实践教程。《Comput. Stat.》29, 3–35，DOI：10.1007/s00180-012-0382-5（2014）。14. Hothorn, T., Buehlmann, P., Kneib, T., Schmid, M. & Hofner, B. mboost：基于模型的提升法。CRAN（2022）。15. Bühlmann, P. & Hothorn, T. 提升算法：正则化、预测与模型拟合。《Stat. Sci.》22, 477–505，DOI：10.1214/07-STS242（2007）。16. Friedman, J. H. 贪心函数逼近：梯度提升机器。《The Annals Stat.》29, 1189–1232，DOI：10.1214/aos/1013203451（2001）。17. Tutz, G. & Binder, H. 提升岭回归。《Comput. Stat. & Data Analysis》51, 6044–6059，DOI：10.1016/j.csda. 2006.11.041（2007）。18. Agarwal, N. K. 验证调查条目的构念效度：信息行为研究中问卷设计的两阶段分类程序。《Proc. Am. Soc. for Inf. Sci. Technol.》48, 1–8，DOI：10.1002/meet.2011.14504801166（2011）。19. Obster, F. & Heumann, C. 稀疏组提升法——无偏组选择与变量选择，DOI：10.48550/arXiv.2206. 06344（2022）。ArXiv：2206.06344 [stat]。20. Pybus, M. 等。分层提升法：一种用于检测和分类人类群体中难以识别的选择性扫荡的机器学习框架。《Bioinformatics》31, 3946–3952，DOI：10.1093/bioinformatics/btv493（2015）。21. Wang, C., Wu, Y. & Liu, Z. 面向多源迁移学习的分层提升法。载于《国际人工智能与机器人会议及国际自动化、控制与机器人工程会议论文集》，ICAIR-CACRE ’16, 1–5，DOI：10.1145/2952744.2952756（计算机协会，纽约州纽约市，美国，2016）。

22. Yang, F.等。 用于路面裂缝检测的特征金字塔与层次增强网络，DOI： 10.48550/arXiv.1901.06340(2019年)。ArXiv:1901.06340 [cs].**23.** Valentini, G. 用于蛋白质功能预测的层次集成方法。 *ISRN*生物信息学。 **2014年**，901419，DOI：10.1155/2014/901419 (2014年)。**24.** Aguinis, H. 与 Gottfredson, R. K. 使用调节多元回归估计交互效应的最佳实践建议。 组织行为学杂志。**31**，776–786，DOI： 10.1002/job.686 (2010年)。**25.** Hofner, B.、Mayr, A. 与 Schmid, M. gamboostLSS：GAMLSS框架下用于模型构建和变量选择的R软件包。 *arXiv:1407.1774 [stat]*(2014年)。ArXiv: 1407.1774。

作者贡献声明

P.P. 和 H.B. 收集了数据，F.O. 进行了机器学习和统计建模，并分析了结果。所有作者均审阅了稿件。

补充信息

按以下顺序列出：**登录号** （如适用）；**竞争利益** （必须声明）。

曲线下面积 sgb	曲线下面积 mb	曲线下面积 2-boost	曲线下面积 mb int	结果变量脆弱性
0.656	0.619	0.608	0.587	夏季气温
0.707	0.708	0.713	0.705	冬季气温
0.852	0.852	0.852	0.500	降雨减少
0.768	0.768	0.768	0.500	干旱
0.776	0.778	0.783	0.773	极端天气

表1。 稀疏组提升（sgb）、分量式提升（mb）、带交互作用的并行提升（mb int）和带交互作用的两步提升（2-boost）在测试数据上针对所有脆弱性结果变量的 AUC值。

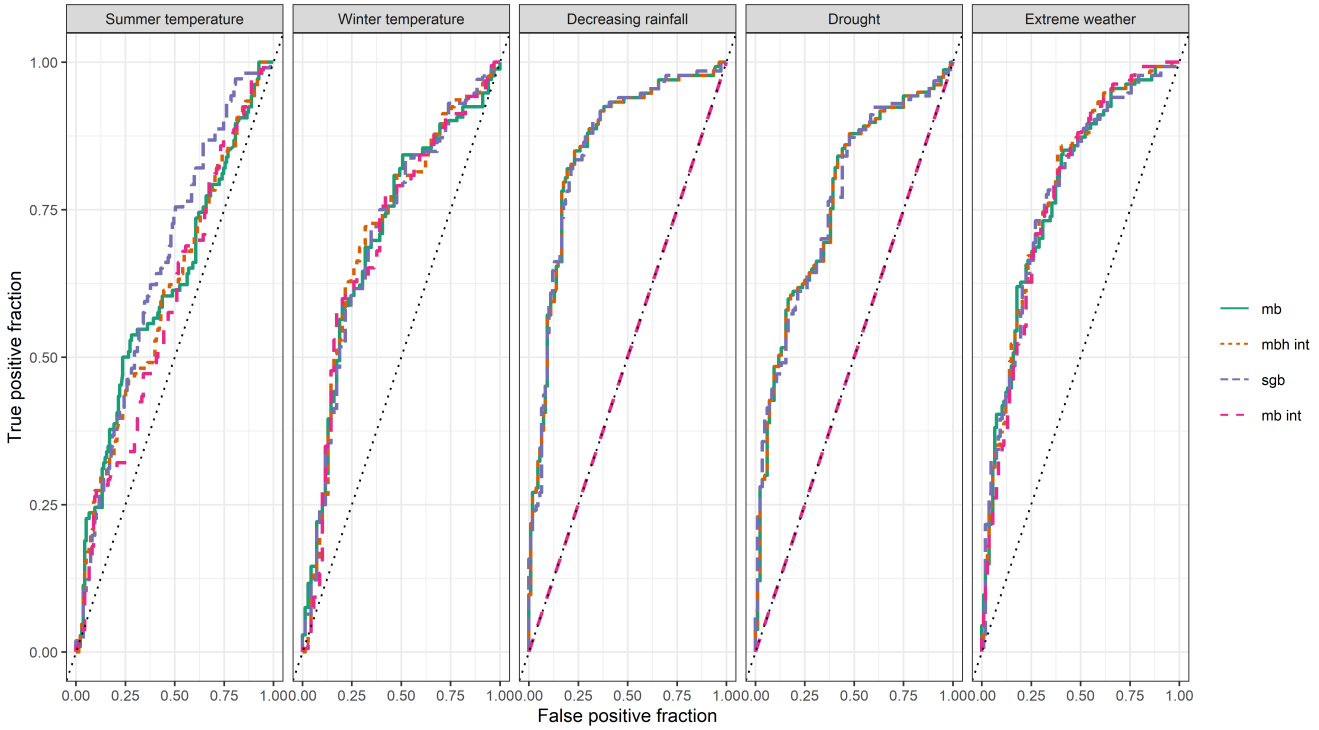

图1。 稀疏组提升（sgb）、分量式提升（mb）、带交互作用的并行提升（mb int）和带交互作用的两步提升（2-boost）在测试数据上针对所有脆弱性结果变量的受试者工作特征曲线。

模型	选定的交互项数量	百分比中的 1-稀疏性	结果变量脆弱性
mb int	13	0.95	夏季气温
2-boost	0	0	夏季气温
mb int	38	2.78	冬季气温
2-boost	12	0.88	冬季气温
mb int	48	3.51	降雨减少
2-boost	1	0.07	降雨减少
mb int	27	1.98	干旱
2-boost	16	1.17	干旱
mb int	32	2.34	极端天气
2-boost	10	0.73	极端天气

表2。 基于两步估计（2-boost）和平行估计（mb int）所选交互项数量的比较，以及在1366个交互项中所选交互作用的百分比（1-稀疏性）。

变量	类别	n
无夏季温度脆弱性	yes	358
无冬季温度脆弱性	yes	579
无降雨量减少脆弱性	yes	451
无干旱脆弱性	yes	492
无极端天气脆弱性	yes	453

表3. 结果变量概览。面对气候灾害的财务脆弱性。“n”列给出了对各类灾害均不存在财务脆弱性的农民人数。

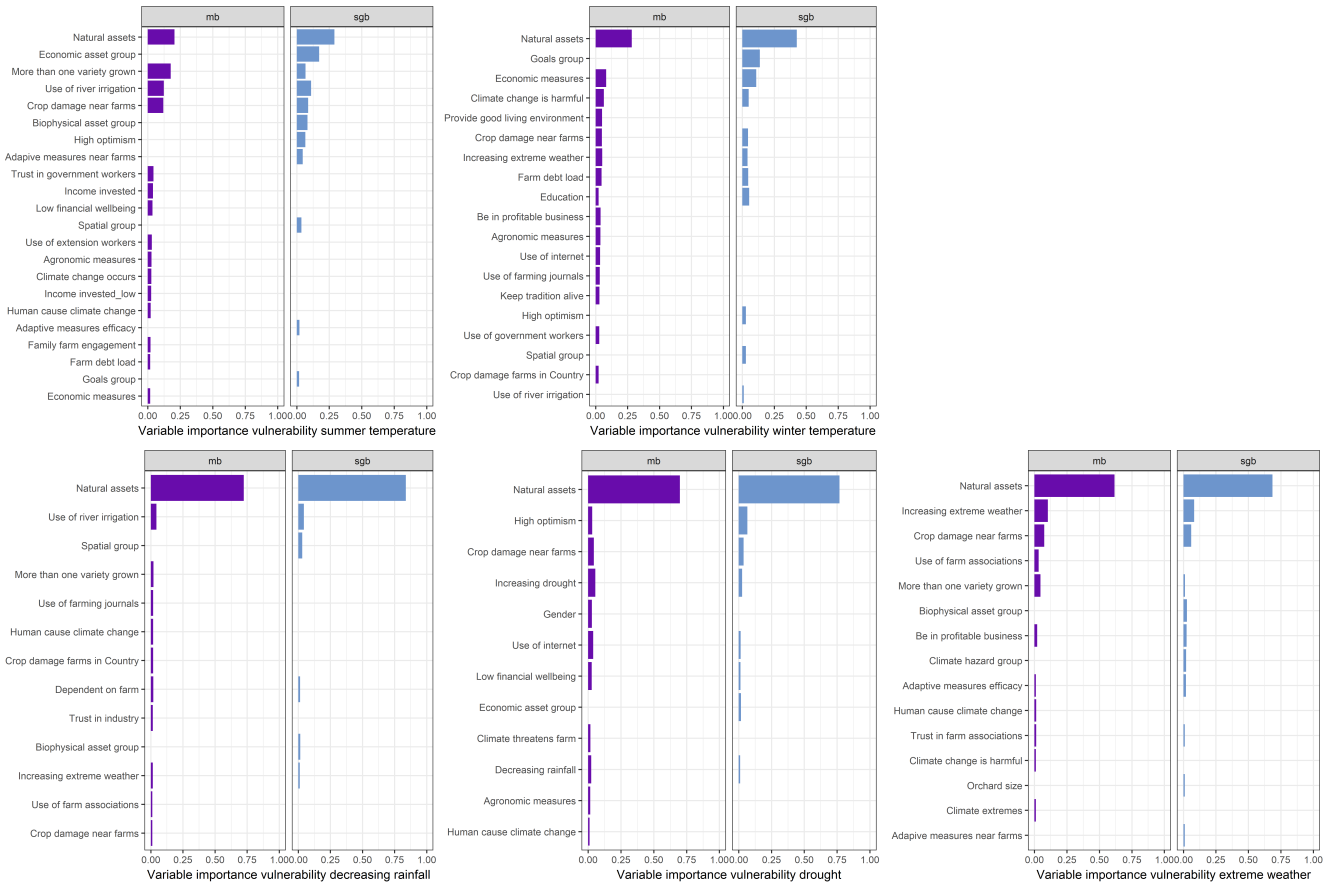

图 2. 比较基于分量式提升（mb）和稀疏组提升（sgb）的各项脆弱性变量重要性。变量的排序依据两个模型的相对重要性之和。图中仅显示相对贡献至少为 1% 的变量，且每个模型最多显示 15 个变量。

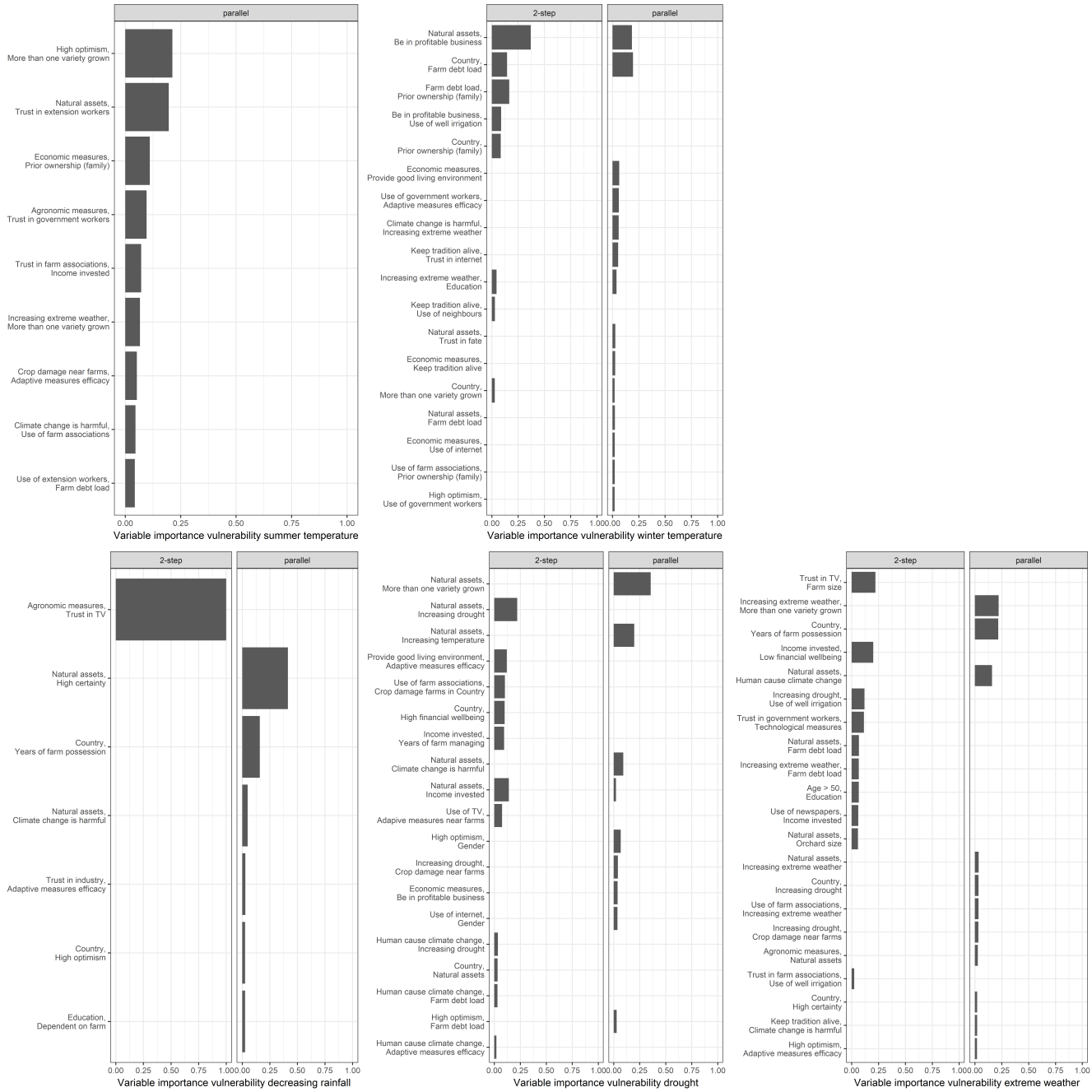

图3. 各漏洞在两步估计（2-boost）和并行估计（2-boost）中的交互项变量重要性。变量排序依据两个模型的相对重要性之和。仅展示相对贡献至少为 2% 的变量，且每个模型最多展示 15 个变量。

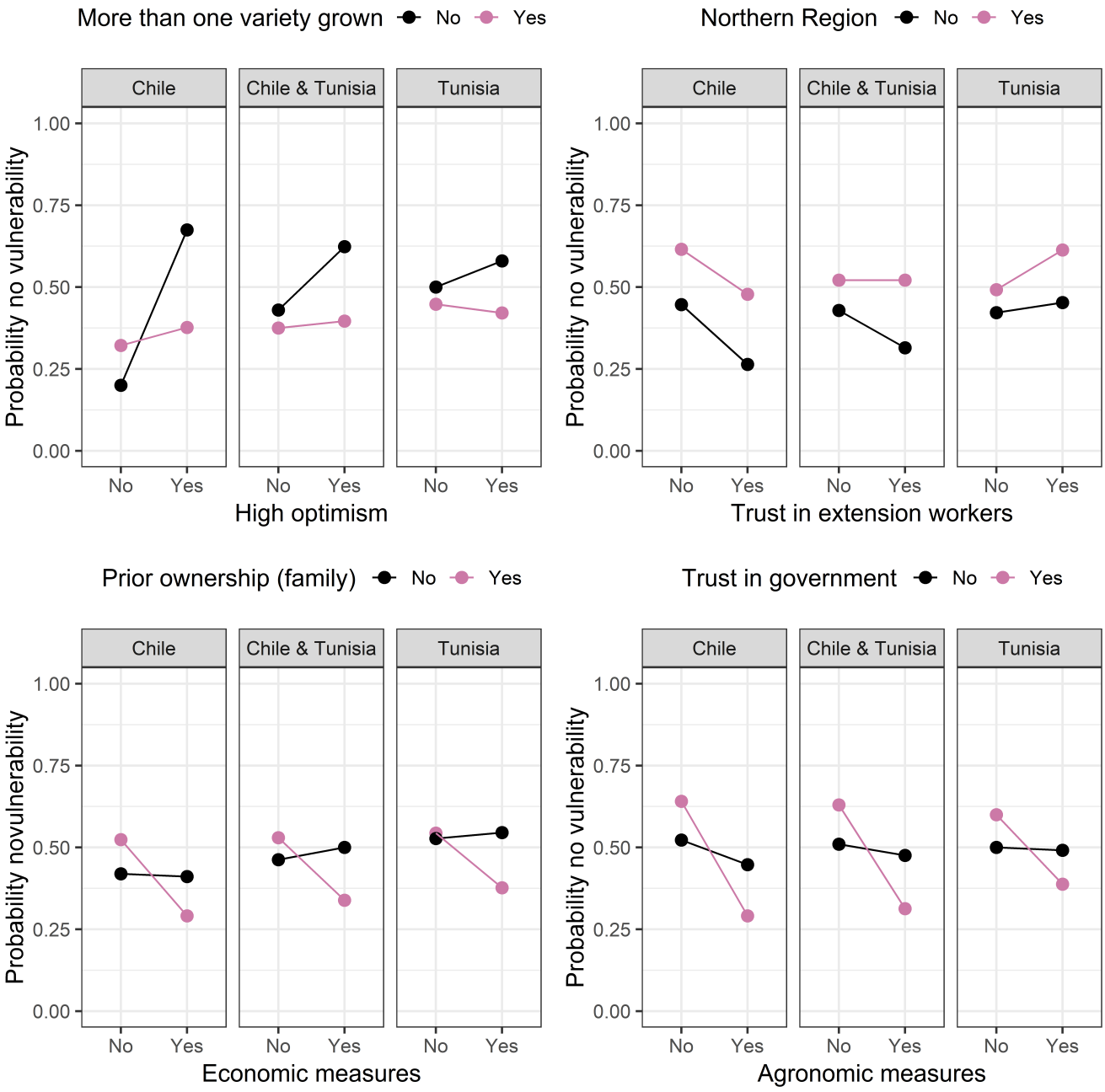

图4. 根据 mb int 和 2-boost 中发现的四个最重要交互效应类别，未处于脆弱状态的概率随夏季气温升高而变化。概率仅基于经典逻辑回归模型计算，每次只使用一个交互项。结果分别按国家（智利、突尼斯）分层，以及基于全部数据进行估算。

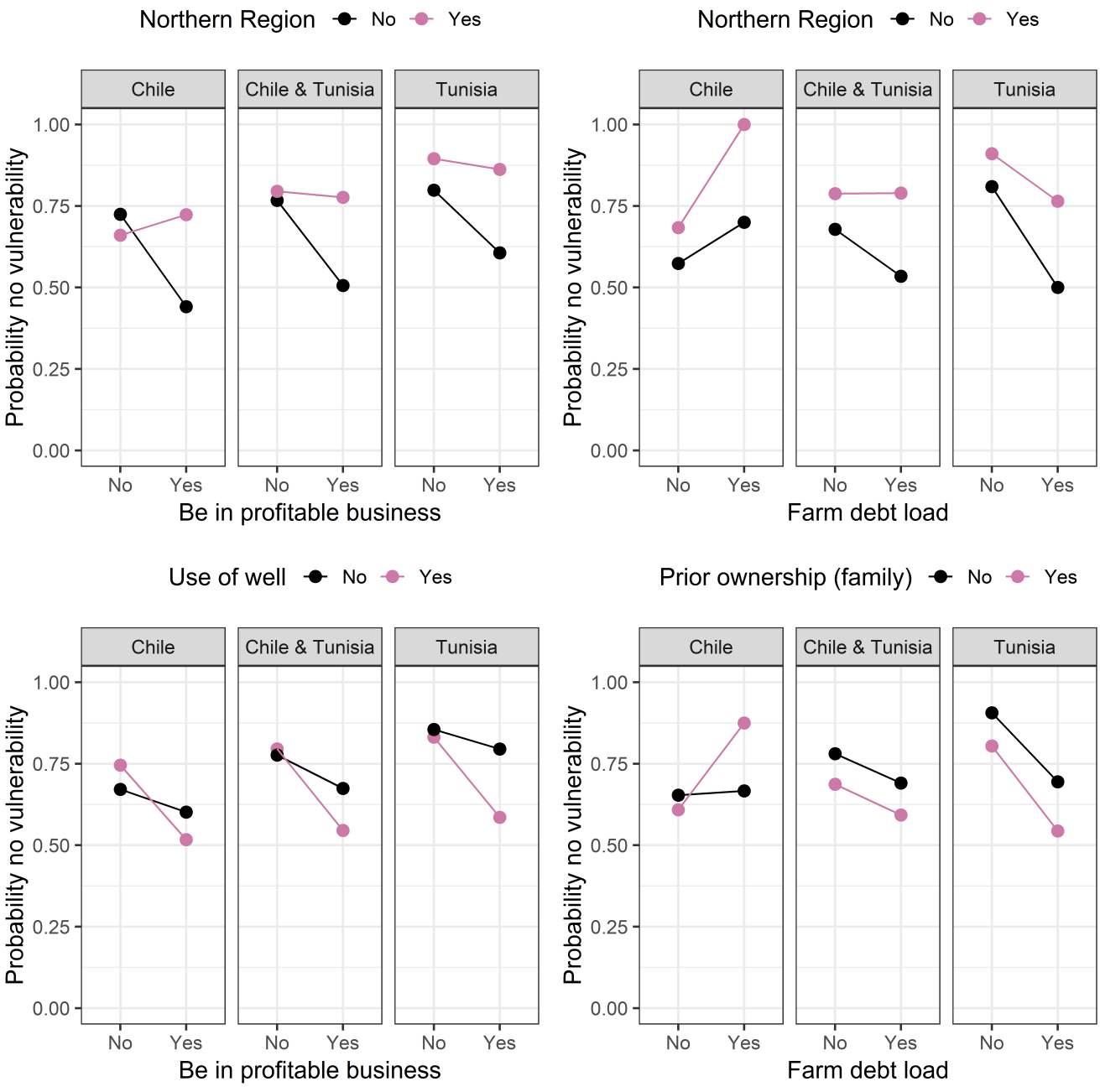

图5. 基于 mb int 和 2-boost 中发现的四个最重要交互效应的类别，冬季气温升高时不易受脆弱性影响的概率。概率仅通过经典逻辑回归计算，每次仅使用一个交互项。结果分别按国家（智利、突尼斯）进行分层估计，并在全部数据上进行一次整体估计。

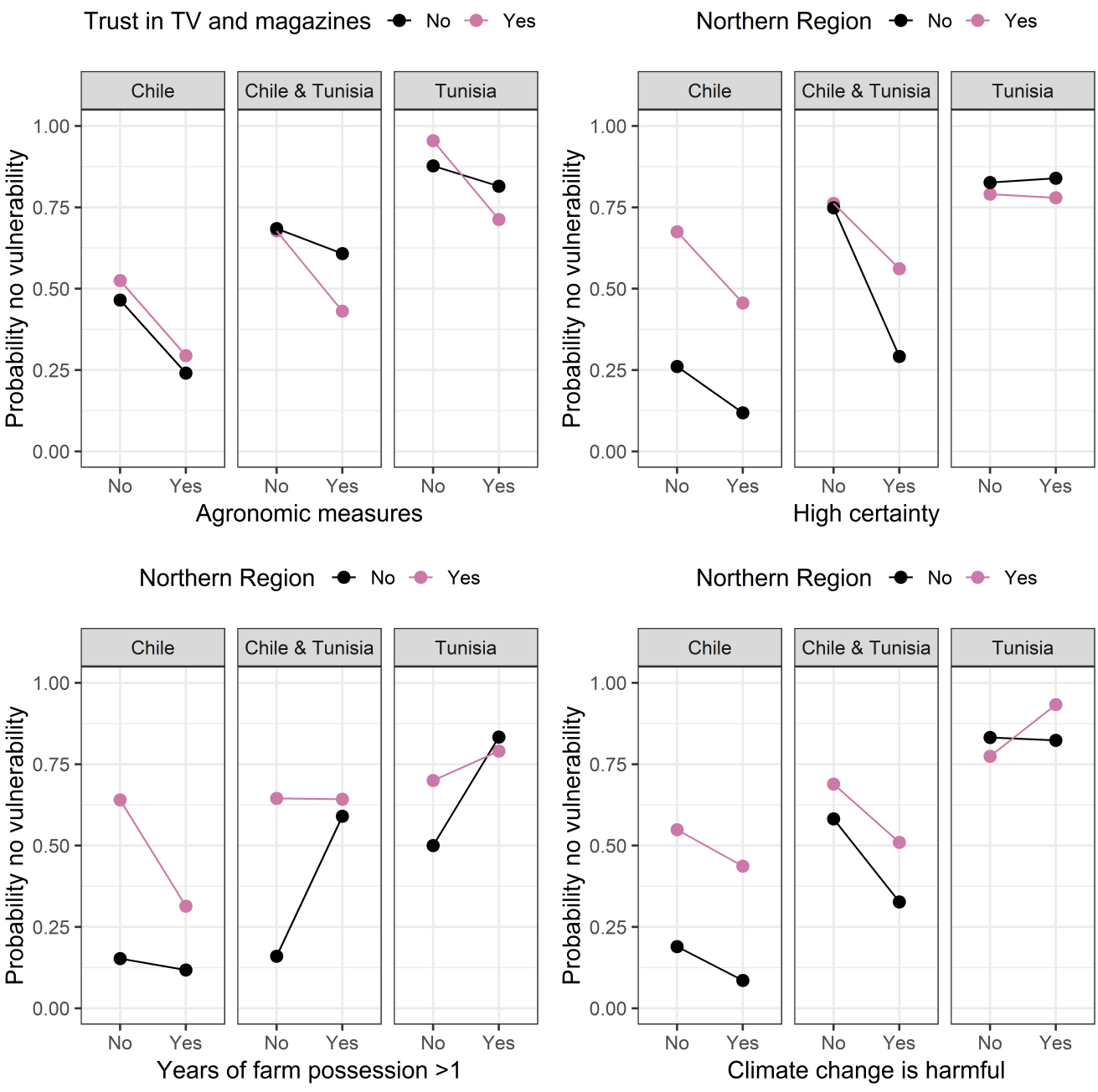

图6. 根据 mb int 和 2-boost 中发现的四个最重要交互效应类别，分析未受降雨减少影响而变得脆弱的概率。概率仅基于每次只纳入一个交互项的经典逻辑回归计算。结果分别按国家（智利、突尼斯）进行分层，以及在 wholedata 上进行估计。

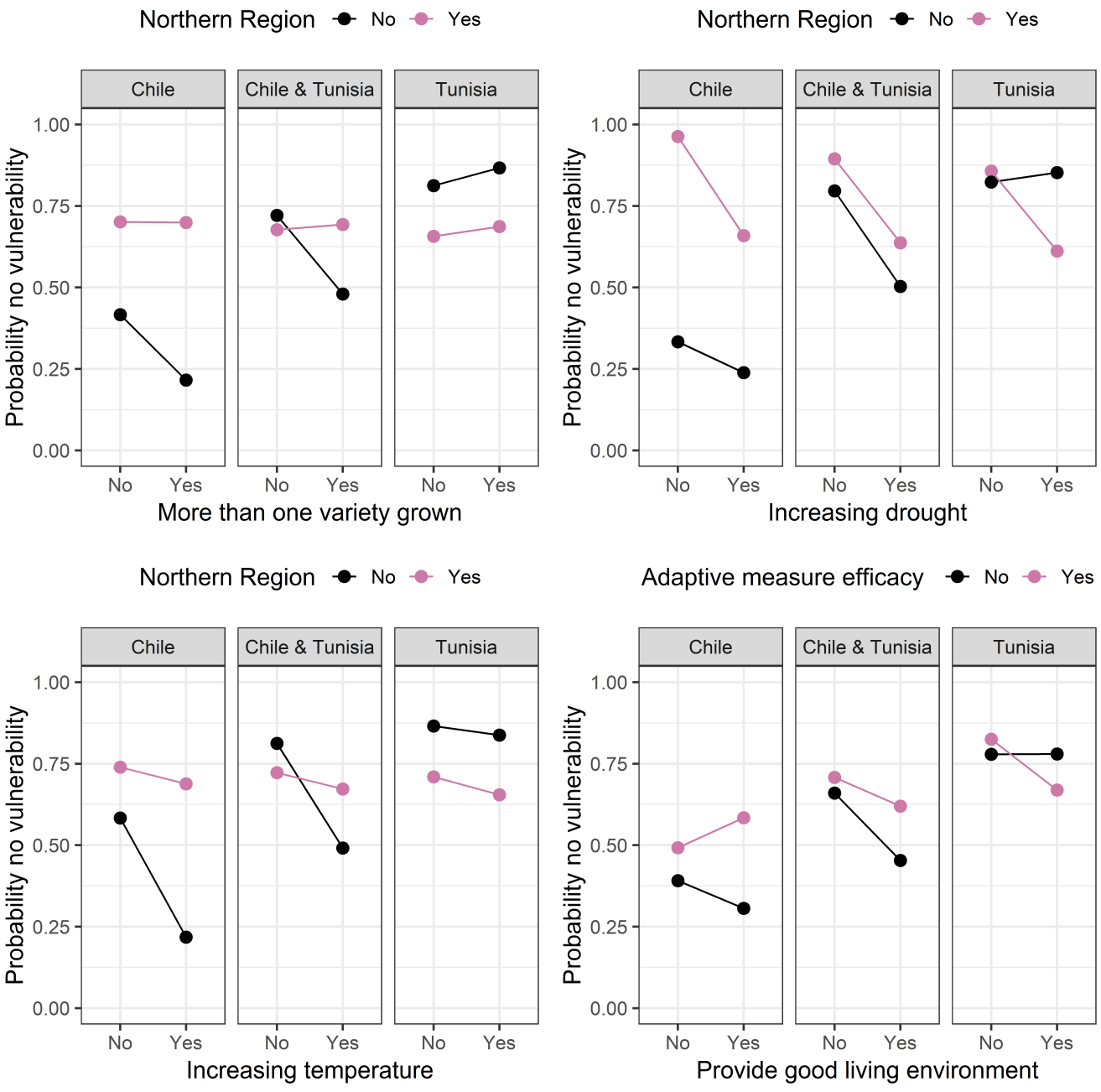

图 7. 根据在 mb int 和 2-boost 中发现的四个最重要交互效应的类别，预测不易受到干旱影响的概率。概率仅基于经典逻辑回归，每次只使用一个交互项。结果分别按国家（智利、突尼斯）分层，以及基于全部数据估算。

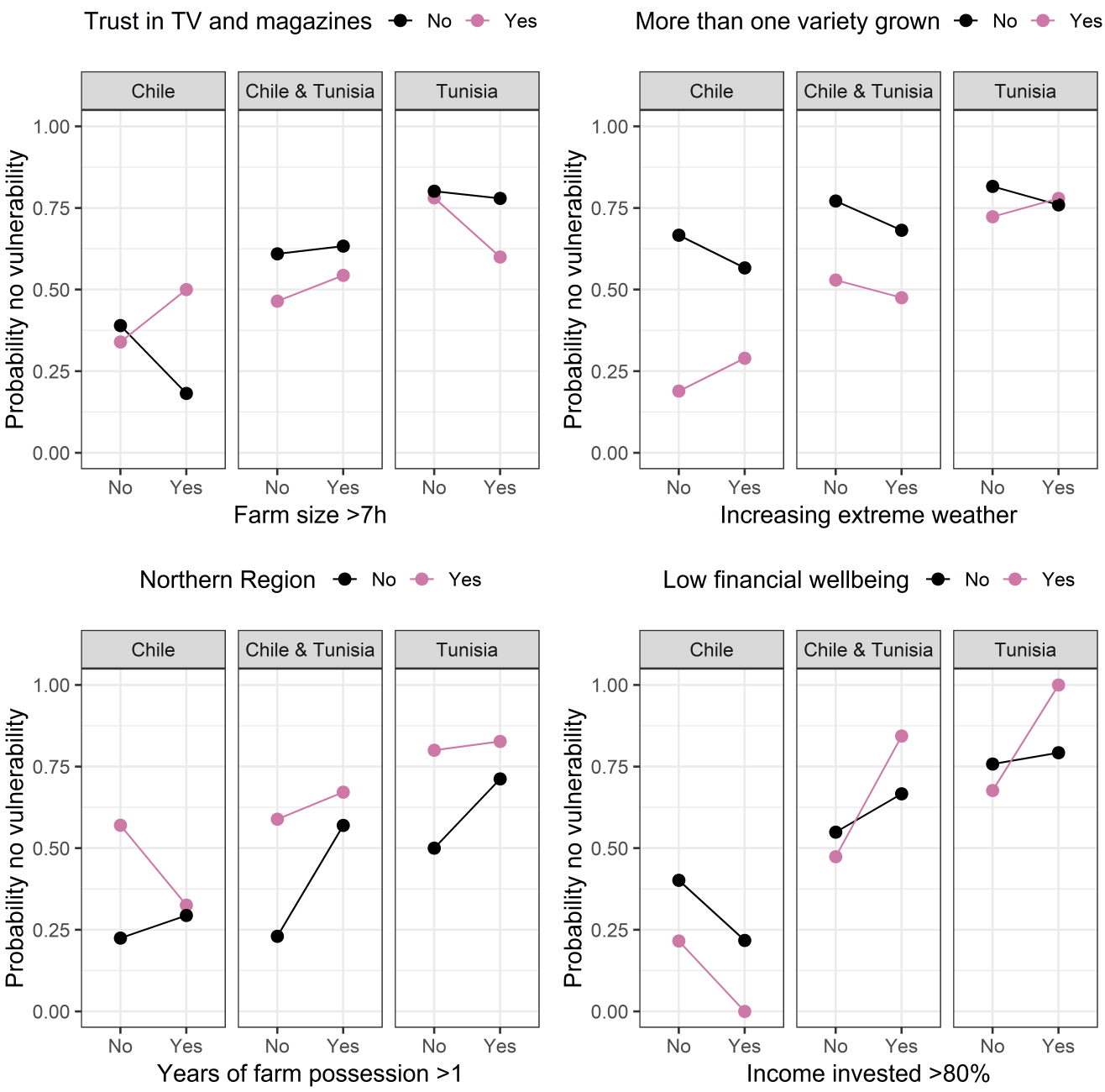

图8. 根据 mb int 和 2-boost 中发现的四个最重要交互效应的类别，预测不易受到极端天气影响的概率。概率仅基于经典逻辑回归计算，每次只使用一个交互项。结果分别按国家（智利、突尼斯）分层，以及基于 wholedata 进行估算。

变量名称	类别	n	组名
农艺措施	yes	647	生物物理资产组
经济措施	yes	464	生物物理资产组
河流灌溉的使用	yes	138	生物物理资产组
井灌的使用	yes	231	生物物理资产组
农场规模	yes	283	生物物理资产组
果园规模	yes	318	生物物理资产组
种植了多个品种	yes	508	生物物理资产组
其他产品	yes	571	生物物理资产组
技术措施	yes	721	生物物理资产组
气温升高	yes	629	气候体验组
降雨减少	yes	659	气候体验组
干旱加剧	yes	671	气候体验组
极端天气增多	yes	542	气候体验组
收入投资 >80百分比	yes	137	经济资产类别
收入投资占比 <40百分比	yes	358	经济资产类别
财务状况良好	yes	346	经济资产类别
财务福祉水平低	yes	148	经济资产类别
农场债务负担	high	96	经济资产类别
依赖农场	yes	528	经济资产类别
家庭农场参与度	yes	203	经济资产类别
适应措施效能	high	490	效能组
独立工作	yes	635	目标组
保持传统延续	yes	460	目标组
提供良好的生活环境	yes	466	目标组
从事盈利性业务	yes	320	目标组
气候变化有害	yes	258	危害组
高度乐观	yes	446	危害组
高度确定性	yes	470	危害组
气候威胁农场	yes	629	危害组
气候风险 >效益	yes	648	危害组
对气候变化的接受度	yes	676	人力资产组
人为导致气候变化	yes	685	人力资产组
气候极端事件	yes	755	人力资产组
年龄 > 50	yes	438	人力资产组
性别	F	121	人力资产组
性别	M	680	人力资产组
教育程度	yes	459	人力资产组
农场拥有年限	yes	577	人力资产组
先前所有权（家庭）	yes	399	人力资产组
农场管理年限	yes	437	人力资产组
自然资产	智利中部	200	自然资产组
自然资产	突尼斯中部	200	自然资产组
自然资产	突尼斯北部	201	自然资产组
自然资产	智利南部	200	自然资产组
农场附近的适应措施	1	424	规范组
农场附近的适应措施	2	151	规范组
农场附近的适应措施	3	226	规范组
高度乐观	yes	446	感知组

表4。 变量和组概览。用作交互变量（潜在调节变量）的22个变量以粗体显示。每个变量各类别中的观测值数量列于 n 列。对于二元变量，仅列出一个类别；若所示类别为“是”，则另一类别为“否”；若所示类别为“高”，则另一类别为“低”。

变量	类别	n	组
报纸使用情况	yes	95	社会资产组
农业期刊使用情况	yes	161	社会资产组
电视使用情况	yes	415	社会资产组
使用收音机	yes	219	社会资产组
使用互联网	yes	319	社会资产组
使用推广人员	yes	346	社会资产组
使用政府工作人员	yes	166	社会资产组
使用邻居	yes	313	社会资产组
行业使用情况	yes	192	社会资产组
农场协会使用情况	yes	97	社会资产组
对报纸的信任	yes	174	社会资产组
对农业期刊的信任	yes	291	社会资产组
对电视的信任	yes	329	社会资产组
对广播的信任	yes	241	社会资产组
对互联网的信任	yes	319	社会资产组
对推广人员的信任	yes	433	社会资产组
对政府工作人员的信任	yes	268	社会资产组
对邻居的信任	yes	319	社会资产组
对行业的信任	yes	215	社会资产组
对农场协会的信任	yes	184	社会资产组
对政府机构的信任	yes	312	社会资产组
对其他农民的信任	yes	351	社会资产组
对宗教的信任	yes	317	社会资产组
对命运的信任	yes	360	社会资产组
农场附近的作物损害	yes	643	空间组
国家的作物受损农场	yes	673	空间组
气候变化发生了	yes	592	空间组

表5. 变量和组概述（续）。用作交互变量（潜在调节变量）的22个变量以粗体显示。每个变量各类别中的观测值数量列于n列中。对于二元变量，仅显示其中一个类别；如果所显示的类别为“是”，则剩余类别为“否”；如果所显示的类别为“高”，则剩余类别为“低”。