DIFFERENTIAL MEASUREMENT OF PROXY DISCRIMINATION

ZAHRA ABOOTALEBI[1,2], ANDREAS TSANAKAS[1], AND RUI ZHU[1]

Abstract. Excluding protected attributes from insurance pricing models does not guarantee the absence of discrimination, as remaining covariates may still act as proxies. This paper develops a differential framework for measuring proxy discrimination in fitted pricing models. The approach decomposes the sensitivity of prices to a covariate into components reflecting the direct impact on predictions and the proxying of a protected attribute. In applications with predominantly categorical features, we use Multiple Correspondence Analysis (MCA) to obtain a continuous latent representation that supports differential sensitivity analysis. In that setting, we decompose discrete price changes due to policyholder profile perturbations into direct and proxy effects, using Aumann-Shapley attributions. Two empirical illustrations on insurance claims datasets reveal weak but detectable proxy effects related to gender. The proposed approach provides a diagnostic framework for assessing the presence and magnitude of proxy-discriminatory effects at both the individual and portfolio level, which is applicable to differentiable pricing models, without requiring model refitting.

Keywords: Proxy discrimination, insurance pricing, fairness, sensitivity analysis, Multiple Correspondence Analysis.

1. Introduction

Concerns about fairness and discrimination have become increasingly prominent in data-driven decision-making (Barocas et al., 2023; Barocas and Selbst, 2016). In many settings, especially those involving personal or demographic data, predictions can encode dependence on sensitive characteristics, even when those characteristics are not explicitly included as inputs (Barocas and Selbst, 2016; Tschantz, 2022). In insurance, the concept of discrimination has a particular role. Insurers are expected to *discriminate* in the technical sense: to distinguish between policyholders by risk and charge premiums that reflect expected losses (Frees and Huang, 2023; Charpentier, 2022). This principle, often called actuarial fairness, is central to well-functioning insurance markets. However, problems arise when the pricing process disproportionately affects protected groups. This includes cases where sensitive characteristics, such as gender, ethnicity, or socioeconomic status, influence prices directly or through correlated variables. In such situations, pricing practices can be regarded by regulators as potentially unfair or unlawful (Frees and Huang, 2023; Xin and Huang, 2024).

Date: April 17, 2026.

[1] Bayes Business School, City St George's, University of London.

[2] Corresponding author: zahra.abootalebi-naeini@citystgeorges.ac.uk.

We are grateful to Mathias Lindholm, Rosalba Radice, and participants at the 17th Actuarial and Financial Mathematics Conference in Brussels, for helpful comments and discussions.

In the insurance context, the literature and regulatory guidance, e.g., the EU Council Directive 2004/113/EC on equal treatment in access to goods and services, emphasize a key distinction between *direct discrimination*, where protected attributes are explicitly used in pricing, and *indirect discrimination*, defined as arising when "an apparently neutral provision, criterion, or practice [...] disproportionately disadvantages persons of one sex" (European Council, 2004). Closely associated to this idea is the mechanism of *proxy discrimination* (Tschantz, 2022; Lindholm et al., 2022, 2024*b*), arising when permitted covariates operate as effective proxies of policyholders' sensitive attributes. Such proxies, then, partially replicate the pricing disparities that would arise from directly using sensitive attributes as inputs to the pricing model.

In light of these concerns, several strands of work propose ways to characterize and mitigate unfairness in pricing. Lindholm et al. (2022) give definitions of direct and indirect discrimination and derive a discrimination-free pricing adjustment that can be implemented by post-processing model predictions. Within the same framework, Lindholm et al. (2024*b*) formalize proxy discrimination as an individual fairness property distinct from group fairness notions (e.g., demographic parity), proving incompatibilities and showing that input pre-processing or output post-processing aimed at group fairness can leave the status of direct or proxy discrimination ambiguous. From an actuarial perspective, Frees and Huang (2023); Xin and Huang (2024) survey how big data expands the use of proxy variables, discuss regulatory prohibitions across jurisdictions, and motivate projection-type adjustments that break dependence on protected traits while retaining risk relevance. Related but distinct from these associational approaches, Araiza Iturria et al. (2024); Côté et al. (2025*a*) develop causal frameworks for insurance fairness; Côté et al. (2025*a*) in particular provide formal definitions of direct and indirect discrimination within a directed acyclic graph.

Despite the substantial efforts to define and mitigate proxy discrimination, diagnostic tools for measuring its practical materiality remain limited (Côté et al., 2025*b*; Lindholm et al., 2026). Furthermore, even though the use of differential methods is key to understanding model sensitivities (e.g., Pesenti et al. (2021, 2025)), few studies define fairness-related quantities explicitly in terms of model derivatives (Huang and Pesenti, 2025; Miao and Pesenti, 2026). To our knowledge, there is no derivative-based framework in the literature that targets proxy discrimination through the joint structure of the input space and the pricing function. This gap motivates our contribution.

This paper develops a framework for measuring *proxy discrimination* in insurance pricing using a derivatives-based diagnostic. Rather than modifying the model to enforce fairness, we assess the fitted pricing function by examining how its predictions respond to input perturbations. Through the statistical dependence between policyholders' sensitive and permitted characteristics, we study the "direct" and "proxy paths" leading from covariates to prices. Concretely, by modifying policyholder characteristics, we construct directions in the pricing model's input space which represent a change in the likelihood of a protected attribute. We then compute local, gradient-based sensitivities on these perturbations to the input profile, while holding the fitted model fixed. In this way,

changes in policyholder attributes indicate how predicted premiums respond along the direct and proxy pathways.

To distinguish between the direct and proxy effects of permitted covariates on prices, we use a regression representation of a protected attribute as a function of other covariates, following formulations used in the sensitivity analysis literature (Mara and Tarantola, 2012; Pesenti et al., 2021). This allows us to isolate indirect pathways through which pricing can be influenced via protected characteristics, even when those characteristics are not explicitly used. The central construct of the paper is a two-argument functional that decouples the direct impact of permitted covariates from their role as proxies operating via the conditional distribution of the protected attribute. The derivative-based quantities introduced subsequently are one natural consequence of this construction under smoothness assumptions. The same construction also opens the door to applying standard model-explainability tools separately to the direct and proxy channels, including for models that are not differentiable, thus widening the scope of our contribution.

We illustrate the framework in two empirical settings. Throughout, we work with deliberately simple and transparent model classes so that the proposed decomposition can be interpreted clearly – maximizing predictive accuracy is not the primary objective of the present paper. In the first case study, we consider a continuous covariate setting based on the `pg15training` motor portfolio (Dutang and Charpentier, 2024), where the proposed approach yields derivative-based direct and proxy effects for a focal rating factor, together with portfolio-level and policyholder-level diagnostics. In the second case study, the pricing inputs are predominantly categorical. Since gradient-based analysis is then no longer directly available in the original covariate space, we embed the categorical features into a continuous latent space via Multiple Correspondence Analysis (MCA; Abdi and Valentin, 2007), preserving input information while enabling sensitivity analysis in the latent space. In that setting, we combine the framework with counterfactual level changes and Aumann–Shapley attributions (Billera and Heath, 1978) to decompose finite premium changes into direct and proxy components. In both studies, the observed proxy effects related to gender are found to be weak but systematic. On the one hand, the proposed method is sensitive enough to detect even small proxy effects. On the other, establishing the low materiality of those effects is in itself useful for regulatory compliance of these insurance portfolios.

The rest of the paper is structured as follows. Section 2 provides the necessary background and formal setup: we discuss proxy discrimination, define unawareness and discrimination-free prices, and introduce differential direct and proxy effects. Section 2.4 presents the first case study for a continuous covariate setting. Section 3 develops the framework for categorical covariates, including the use of Multiple Correspondence Analysis and counterfactual profile changes; Section 3.3 applies these ideas in the second case study. Finally, we summarize our findings in Section 4.

2. Proxy Discrimination

2.1. **Background.** Following Lindholm et al. (2022), we adopt a standard insurance pricing setting and work on a probability space $(\Omega, \mathcal{F}, \mathbb{P})$, supporting a random vector $(Y, \boldsymbol{X}, D)$, where:

- Y is a real-valued random variable representing the loss quantity to be predicted and priced (e.g. claim cost or claim count).
- $\boldsymbol{X}$ is a k-dimensional vector of covariates representing policyholder characteristics whose use is permitted for pricing.
- D is a random variable corresponding to a sensitive or protected attribute. Usually D is categorical (e.g., gender or ethnicity) and then we assume that it takes values in a finite set $\mathfrak{D}$; in the case of a binary outcome, we write $\mathfrak{D} = \{0, 1\}$.

The joint distribution of $(\boldsymbol{X}, D)$ describes the portfolio composition. The split between permitted covariates $\boldsymbol{X}$ and the protected attribute D is taken as exogenous, e.g., determined by law or internal policy.

2.2. **Unawareness prices and proxy discrimination.** The basis of technical pricing is the conditional expectation of Y given the information available to the insurer. Using full information $(\boldsymbol{X}, D)$, define the best-estimate price

$$\mu(\boldsymbol{x}, d) \;:=\; \mathbb{E}[Y \mid \boldsymbol{X} = \boldsymbol{x},\, D = d], \tag{1}$$

which is the optimal predictor of Y among all measurable functions of $(\boldsymbol{X}, D)$. When the sensitive attribute is disallowed as an explicit rating factor, a natural alternative is the unawareness price

$$\mu(\boldsymbol{x}) \;:=\; \mathbb{E}[Y \mid \boldsymbol{X} = \boldsymbol{x}]. \tag{2}$$

Although this excludes D from the model inputs, it does not, in general, eliminate the influence of D on prices. Indeed,

$$\mu(\boldsymbol{x}) = \mathbb{E}\big[\mu(\boldsymbol{x}, D) \mid \boldsymbol{X} = \boldsymbol{x}\big] = \int \mu(\boldsymbol{x}, d)\, \mathrm{d}\mathbb{P}(d \mid \boldsymbol{X} = \boldsymbol{x}).$$

Thus, the unawareness price is obtained by averaging the full-information price surface $d \mapsto \mu(\boldsymbol{x}, d)$ with respect to the conditional law of D given $\boldsymbol{X} = \boldsymbol{x}$.

By this mechanism, proxy effects can arise: if D is dependent on $\boldsymbol{X}$, then excluding D from the regression does not prevent the fitted predictor from using sensitive information in pricing indirectly, via the dependence between $\boldsymbol{X}$ and D (Lindholm et al., 2022, 2024*b*). Hence, *proxy discrimination* occurs when two ingredients are present simultaneously:

(a) $\boldsymbol{X}$ and D are statistically dependent.

(b) The best-estimate surface depends on d; that is, holding $\boldsymbol{x}$ fixed, the map $d \mapsto \mu(\boldsymbol{x}, d)$ is not constant.

Condition (a) captures the presence of proxy information: some permitted characteristics carry signal about D through $\mathbb{P}(D \mid \boldsymbol{X})$. Condition (b) captures that D matters for predicted costs under full information; if (b) fails, then inference of D becomes irrelevant for pricing.

These effects also motivate the idea of discrimination-free pricing, which aims to remove the indirect impact of the sensitive attribute on prices. This is achieved by replacing the conditional law $\mathbb{P}(D \mid \boldsymbol{X} = \boldsymbol{x})$ in (2) with a fixed reference distribution $\mathbb{P}^*(D)$ of the protected attribute that does not depend on $\boldsymbol{x}$. Then, one can define a discrimination-free insurance price (Lindholm et al., 2022)

$$\mu^*(\boldsymbol{x}) \;:=\; \int \mu(\boldsymbol{x}, d)\, \mathrm{d}\mathbb{P}^*(d). \tag{3}$$

A natural choice is $\mathbb{P}^*(D) = \mathbb{P}(D)$, the unconditional portfolio distribution of D; for a justification of this choice by causal argument, see Lindholm et al. (2022); Côté et al. (2025*a*).

2.3. **Sensitivity of prices to direct and proxy channels.** We now make explicit the mechanisms by which permitted covariates drive the unawareness price (2), by separating their effect on fully modelled claims costs (the direct channel) from their effect on the conditional distribution of the sensitive attribute (the proxy channel).

Throughout this section, we assume that the covariates $\boldsymbol{X}$ are continuous; we will revisit this assumption in Section 3. It is always possible to construct a random variable U, independent of $\boldsymbol{X}$, and a function g such that

$$D = g(\boldsymbol{X}, U). \tag{4}$$

A standard choice is

$$g(\boldsymbol{x}, u) = F^{-1}_{D|\boldsymbol{X}}(u \mid \boldsymbol{x}), \quad U \sim \mathrm{Unif}[0,1], \tag{5}$$

where $F^{-1}_{D|\boldsymbol{X}}(\cdot \mid \boldsymbol{x})$ denotes the conditional quantile function. When D is continuous, one may take $U = F_{D|\boldsymbol{X}}(D \mid \boldsymbol{X})$. When D is not continuous, such a representation can be obtained via a suitable generalised probability integral transform; see Theorem 3 in Rüschendorf and de Valk (1993). Representations of the form (4), (5) are commonly used in sensitivity analysis when dependence between inputs is central to the problem (Mara and Tarantola, 2012; Pesenti et al., 2021). Using (4), the unawareness price for a policyholder with profile $\boldsymbol{X} = \boldsymbol{x}$ can be written as

$$\mu(\boldsymbol{x}) = \mathbb{E}[\mu(\boldsymbol{x}, D) \mid \boldsymbol{X} = \boldsymbol{x}] = \mathbb{E}\big[\mu\big(\boldsymbol{x}, g(\boldsymbol{x}, U)\big)\big]\,, \tag{6}$$

where the expectation is taken with respect to U only.

To distinguish the ways that permitted covariates act through the best-estimate surface versus through the dependence structure $\mathbb{P}(D \mid \boldsymbol{X})$, we define $\nu : \mathbb{R}^{2k} \to \mathbb{R}$ by

$$\nu(\boldsymbol{x}, \boldsymbol{x}') := \mathbb{E}\big[\mu\big(\boldsymbol{x}, g(\boldsymbol{x}', U)\big)\big]\,. \tag{7}$$

The function $\nu(\boldsymbol{x}, \boldsymbol{x}')$ is the central object of study in this paper. Its two arguments play distinct roles. The first argument, $\boldsymbol{x}$, is the profile at which the best-estimate surface $\mu(\boldsymbol{x}, d)$ is evaluated, and therefore represents the direct impact of $\boldsymbol{x}$ on predicted costs. The second argument, $\boldsymbol{x}'$, is used only to generate the protected attribute through $g(\boldsymbol{x}', U)$, and therefore relates to the conditional distribution of D used in the averaging step. Thus, $\nu(\boldsymbol{x}, \boldsymbol{x}')$ evaluates the price surface at profile $\boldsymbol{x}$, while using the distribution of the protected attribute associated with the possibly different profile

$\boldsymbol{x}'$. When $\boldsymbol{x}' = \boldsymbol{x}$, we recover the unawareness price, that is, $\mu(\boldsymbol{x}) = \nu(\boldsymbol{x}, \boldsymbol{x})$. When $\boldsymbol{x}' \neq \boldsymbol{x}$, $\nu(\boldsymbol{x}, \boldsymbol{x}')$ is a counterfactual quantity introduced to isolate proxy effects on pricing.

Studying the sensitivity of $\nu(\boldsymbol{x}, \boldsymbol{x}')$ in (7) to $\boldsymbol{x}'$, using standard model explainability or sensitivity analysis tools, allows the quantification of proxy discrimination. We now define the direct and proxy effects of a permitted covariate as partial derivatives of the unawareness price across the direct and proxy channels. These are the constructs of central interest in this paper.

Definition 1. For a given policyholder profile $\boldsymbol{x}$, assume that the function $\nu(\boldsymbol{x}, \boldsymbol{x}')$ defined in (7) is differentiable at $(\boldsymbol{x}, \boldsymbol{x})$ with respect to its i-th coordinate in the first argument and its i-th coordinate in the second argument. We define the *direct and proxy effects* of the covariate X_i, respectively, by

$$\begin{aligned} \mathrm{DE}_i(\boldsymbol{x}) &= \frac{\partial}{\partial x_i} \nu(\boldsymbol{x}, \boldsymbol{x}') \Big|_{\boldsymbol{x}'=\boldsymbol{x}}, \\ \mathrm{PE}_i(\boldsymbol{x}) &= \frac{\partial}{\partial x_i'} \nu(\boldsymbol{x}, \boldsymbol{x}') \Big|_{\boldsymbol{x}'=\boldsymbol{x}}. \end{aligned} \tag{8}$$

We note that, by construction, we have

$$\frac{\partial}{\partial x_i} \mu(\boldsymbol{x}) = \mathrm{DE}_i(\boldsymbol{x}) + \mathrm{PE}_i(\boldsymbol{x}),$$

such that the sensitivity of the unawareness price is additively decomposed into the direct and proxy effects.

In order to derive specific expressions for the direct and proxy effects, we need to specify what kind of variable D is.

Discrete D. First, we assume that D is discrete, taking values in a finite set $\mathfrak{D}$. Then, the conditional quantile (5) becomes a step function and one can show that

$$\nu(\boldsymbol{x}, \boldsymbol{x}') = \sum_{d \in \mathfrak{D}} \mu(\boldsymbol{x}, d) h(\boldsymbol{x}', d),$$

where $h(\boldsymbol{x}', d) = \mathbb{P}(D = d \mid \boldsymbol{X} = \boldsymbol{x}')$. Assuming that the following derivatives are well defined,

$$\partial_{x_i} \mu(\boldsymbol{x}, d) := \frac{\partial}{\partial x_i} \mu(\boldsymbol{x}, d), \quad \partial_{x_i} h(\boldsymbol{x}, d) := \frac{\partial}{\partial x_i} h(\boldsymbol{x}, d), \tag{9}$$

it follows that:

$$\begin{aligned} \mathrm{DE}_i(\boldsymbol{x}) &= \textstyle\sum_{d \in \mathfrak{D}} \partial_{x_i} \mu(\boldsymbol{x}, d) h(\boldsymbol{x}, d) \\ \mathrm{PE}_i(\boldsymbol{x}) &= \textstyle\sum_{d \in \mathfrak{D}} \mu(\boldsymbol{x}, d) \partial_{x_i} h(\boldsymbol{x}, d) \end{aligned} \tag{10}$$

Matters simplify further when $D \in \{0, 1\}$ is binary. In that case, denoting $h(\boldsymbol{x}) = h(\boldsymbol{x}, 1)$, we have

$$\begin{aligned} \mathrm{DE}_i(\boldsymbol{x}) &= \partial_{x_i} \mu(\boldsymbol{x}, 0) + h(\boldsymbol{x}) \big(\partial_{x_i} \mu(\boldsymbol{x}, 1) - \partial_{x_i} \mu(\boldsymbol{x}, 0) \big) \\ \mathrm{PE}_i(\boldsymbol{x}) &= \partial_{x_i} h(\boldsymbol{x}) \big(\mu(\boldsymbol{x}, 1) - \mu(\boldsymbol{x}, 0) \big) \end{aligned} \tag{11}$$

Expression (11) makes the channel interpretation explicit. The direct effect $\mathrm{DE}_i(\boldsymbol{x})$ works through the sensitivities of the best-estimate surface to x_i for each group d, mixed according to the prediction weight $h(\boldsymbol{x})$. In contrast, the proxy effect $\mathrm{PE}_i(\boldsymbol{x})$ works through the sensitivity of the conditional probability $h(\boldsymbol{x}) = \mathbb{P}(D = 1 \mid \boldsymbol{X} = \boldsymbol{x})$, scaled by the subgroup pricing gap $\mu(\boldsymbol{x}, 1) - \mu(\boldsymbol{x}, 0)$. In particular, $\mathrm{PE}_i(\boldsymbol{x})$ can only be material when both (i) X_i is informative about D (so $\partial_{x_i} h(\boldsymbol{x}) \neq 0$)

and (ii) the best-estimate price differs across d (so $\mu(\boldsymbol{x}, 1) \neq \mu(\boldsymbol{x}, 0)$), matching the two conditions for proxy discrimination discussed in Section 2.2.

Continuous D. Let D be continuous and assume the derivatives

$$\partial_d \mu(\boldsymbol{x}, d) := \frac{\partial}{\partial d}\mu(\boldsymbol{x}, d), \quad \partial_{x_i'} g(\boldsymbol{x}', u) := \frac{\partial}{\partial x_i'} g(\boldsymbol{x}', u) \tag{12}$$

are well defined and that all relevant derivatives in (9), (12) are continuous. Then, differentiation of $\nu(\boldsymbol{x}, \boldsymbol{x}')$ in (7) leads to:

$$\begin{aligned} \mathrm{DE}_i(\boldsymbol{x}) &= \mathbb{E}\big[\partial_{x_i}\mu\big(\boldsymbol{x}, g(\boldsymbol{x}, U)\big)\big] \\ \mathrm{PE}_i(\boldsymbol{x}) &= \mathbb{E}\big[\partial_d \mu\big(\boldsymbol{x}, g(\boldsymbol{x}, U)\big)\, \partial_{x_i} g(\boldsymbol{x}, U)\big]\,. \end{aligned} \tag{13}$$

Once again, the interpretation is consistent with the previous discussion. Specifically, a non-zero proxy effect $\mathrm{PE}_i(\boldsymbol{x})$ appears when $\mu(\boldsymbol{x}, d)$ varies in d and the conditional law $D \mid \boldsymbol{X} = \boldsymbol{x}$ is sensitive to x_i.

Differentiability and limitations. Naturally, differentiability is a strong assumption. When elements of $\boldsymbol{X}$ are discrete or categorical, derivatives with respect to $\boldsymbol{x}$ are generally not meaningful. We return to the specific issue of categorical variables in Section 3.1. More generally, both smooth models (such as GAMs and neural networks) and non-smooth models (such as tree-based methods) are widely used in pricing. Our definitions of DE_i and PE_i assume differentiability of the various prediction functions used, which restricts the model space considered; note however that relevant derivatives can also be defined for discontinuous models, using the methods of Pesenti et al. (2025).

The key conceptual contribution of this paper is the consideration of the function $\nu(\boldsymbol{x}, \boldsymbol{x}')$ as a tool for quantifying proxy discrimination, via its sensitivity in the second argument. By explicitly separating the direct and proxy channels by which permitted covariates act on predictions, standard model explainability or sensitivity analysis tools can be applied to ν in order to study direct and proxy impacts, including in settings where derivatives are not available.

We do not pursue such methods in full generality here, and instead give as an example Accumulated Local Effects (ALE) (Apley and Zhu, 2020), which connect directly to our framework. First, note that the derivatives in Definition 1 depend on the full profile $\boldsymbol{x}$, which makes direct visualisation difficult. One way to obtain global (i.e., portfolio-wide) interpretable one-dimensional summaries is to average over non-focal covariates, yielding the conditional mean derivative curves

$$x_i \mapsto \mathbb{E}[\mathrm{DE}_i(\boldsymbol{X}) \mid X_i = x_i]\,, \qquad x_i \mapsto \mathbb{E}[\mathrm{PE}_i(\boldsymbol{X}) \mid X_i = x_i]\,. \tag{14}$$

Integrating the curves in (14) yields the (uncentred) Accumulated Local Effects (ALE) in the i-th and $(k+i)$-th arguments of ν under differentiability; see Apley and Zhu (2020, eq. 7). More broadly, because ALEs can also be defined without differentiability, the same two-argument construction $\nu(\boldsymbol{x}, \boldsymbol{x}')$ provides a practical route to visualising direct and proxy impacts even for non-smooth prediction models.

2.4. **Case Study I: `pg15training` data.**

2.4.1. *Overview.* We apply the proposed framework to the `pg15training` dataset from the 2015 Pricing Game portfolio, publicly distributed with the `CASdatasets` R package (Dutang and Charpentier, 2024). The data contain third-party liability motor policies observed for up to one year, with policy exposure, claim outcomes, and a range of standard actuarial rating factors. The dataset has been used in several actuarial pricing and fairness studies, including Xin and Huang (2024), and related applications (Dutang and Charpentier, 2024; Ponnet et al., 2021; Chevalier and Côté, 2025).

We focus on material-damage claim amounts, $Y = \texttt{Indtppd}$ and consider as the protected attribute $D = \texttt{Gender}$, with $D \in \{0, 1\}$ $D = 1$ (female), $D = 0$ (male). In the vector of permitted rating factors $\boldsymbol{X}$, we designate $X_1 = \texttt{Age}$ as the focal permitted covariate. The workflow is: (i) estimate the classifier $h(\boldsymbol{x}) = \mathbb{P}(D = 1 \mid \boldsymbol{X} = \boldsymbol{x})$; (ii) estimate subgroup best-estimate prices $\mu(\boldsymbol{x}, 0)$ and $\mu(\boldsymbol{x}, 1)$; (iii) construct unawareness and discrimination-free prices; and (iv) calculate the direct and proxy effects of `Age` on unawareness prices.

2.4.2. *Data analysis.*

Classification model. We model $h(x_1) = \mathbb{P}(D = 1 \mid X_1 = x_1)$ by logistic regression,

$$\log\left(\frac{h(x_1)}{1 - h(x_1)}\right) = \alpha_0 + \alpha_1 \, x_1. \tag{15}$$

The fitted model gives $\hat{\alpha}_1 = -0.0104$ ($p < 2 \times 10^{-16}$), indicating that the inferred probability of $D = 1$ decreases with `Age`. We use a univariate specification with only `Age` as a covariate for simplicity of exposition. We focus on `Age` because it emerges as the covariate with by far the strongest effect size and also the only clearly statistically significant predictor of D.

The remaining rating factors are controlled for in the best-estimate model for $\mu(\boldsymbol{x}, d)$; full coefficient tables for all fitted models are reported in Appendix A.1.

Best-estimate model. Following Xin and Huang (2024), we remove inconsistent records with non-zero claim count and zero claim amount, i.e. observations with $\texttt{Numtppd} > 0$ but $\texttt{Indtppd} = 0$. We take the annualised material-damage claim amount as the pricing target Y.

We estimate the best-estimate surface $\mu(\boldsymbol{x}, d)$ using a Tweedie generalised additive model with log-link via distributional regression in `GJRM::gamlss` (Marra and Radice, 2025, 2024). For policyholder i, we assume

$$Y_i \mid (\boldsymbol{X}_i = \boldsymbol{x}_i, D_i = d_i) \sim \mathrm{Tw}(\mu_i, \phi_i, p),$$

where $\mu_i = \mathbb{E}[Y_i \mid \boldsymbol{X}_i = \boldsymbol{x}_i, D_i = d_i]$ is the conditional mean, $\phi_i > 0$ is the dispersion parameter, and $p \in (1, 2)$ is the Tweedie power parameter. For policyholder i, we use the offset

$$o_i = \log(\texttt{Exppdays}_i / 365).$$

The Tweedie family is well suited to claims amounts as it accommodates a point mass at zero together with a continuous, right-skewed distribution for positive outcomes. In addition to modelling the

conditional mean, we allow the dispersion parameter ϕ_i to vary with covariates to accommodate heteroskedasticity.

`Age` enters the regressions via a smooth term $s(\texttt{Age})$. In compact form, the mean submodel is

$$\log \mu(\boldsymbol{x}, d) = \beta_0 + \beta_D d + s(\texttt{Age}) + (\text{other covariate effects}) + o_i. \tag{16}$$

A separate log-link regression is used for the dispersion parameter ϕ_i, with the corresponding coefficient estimates reported in Appendix A.1 where we also report the fitted mean equation in tables 5–6. The coefficient on D is -0.396 ($p < 2 \times 10^{-16}$), implying that, holding permitted covariates fixed, the fitted mean for women is approximately 33% lower expected value compared to men. The smooth term $s(\texttt{Age})$ is the component of the fitted mean equation through which `Age` enters directly. It is highly significant, indicating a pronounced non-linear direct relationship between `Age` and the fitted mean on the log scale, as seen in Figure 1.

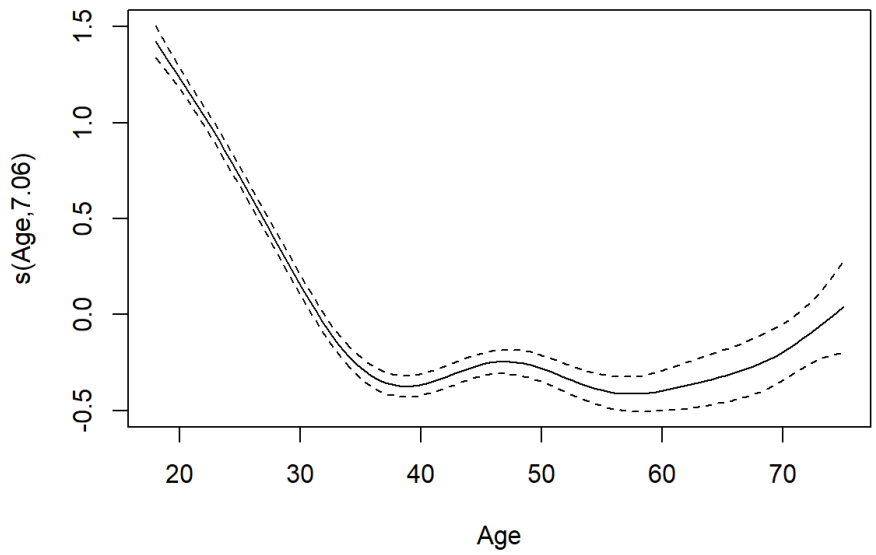


FIGURE 1. Baseline portfolio: fitted smooth effect $s(\texttt{Age})$ in the Tweedie mean equation.

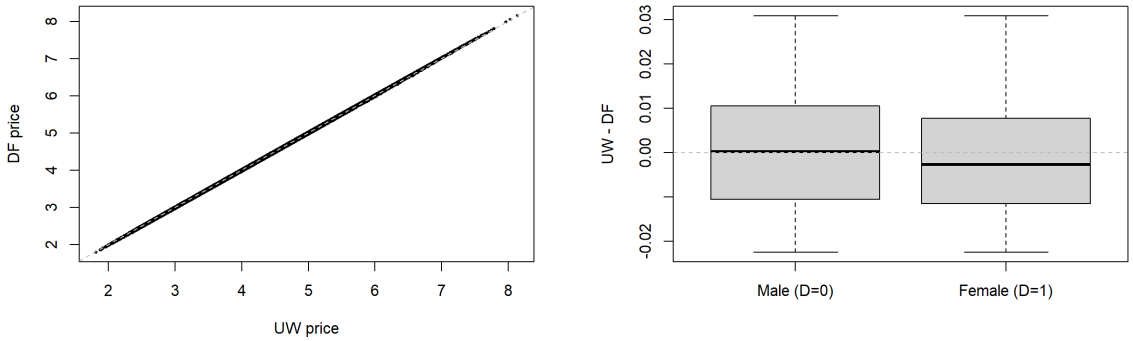
Unawareness and discrimination-free prices. We combine the fitted classifier $\hat{h}(\boldsymbol{x}) = \widehat{\mathbb{P}}(D = 1 \mid \boldsymbol{X} = \boldsymbol{x})$ and the fitted subgroup best-estimate surfaces $\hat{\mu}(\boldsymbol{x}, 0)$ and $\hat{\mu}(\boldsymbol{x}, 1)$ to obtain the estimated unawareness price,

$$\widehat{\mu}(\boldsymbol{x}) = (1 - \hat{h}(\boldsymbol{x}))\, \hat{\mu}(\boldsymbol{x}, 0) + \hat{h}(\boldsymbol{x})\, \hat{\mu}(\boldsymbol{x}, 1).$$

Furthermore, we calculate a discrimination-free price by replacing the conditional weight $\hat{h}(\boldsymbol{x})$ with the unconditional portfolio share $\hat{p} = \widehat{\mathbb{P}}(D = 1) = 0.3657$:

$$\widehat{\mu}^*(\boldsymbol{x}) = (1 - \hat{p})\, \hat{\mu}(\boldsymbol{x}, 0) + \hat{p}\, \hat{\mu}(\boldsymbol{x}, 1).$$

Figure 2a shows that $\widehat{\mu}$ and $\widehat{\mu}^*$ are very close for individual policies. The deviation $\widehat{\mu}(\boldsymbol{x})-\widehat{\mu}^*(\boldsymbol{x})$ isolates the effect of using an inferred weight $\hat{h}(\boldsymbol{x})$ rather than a fixed portfolio weight $\hat{p}$. In this portfolio, this deviation is small and centred around zero, as seen in Figure 2b. Moreover, $\widehat{\mu}(\boldsymbol{x}) - \widehat{\mu}^*(\boldsymbol{x})$ does not substantially differ across genders.

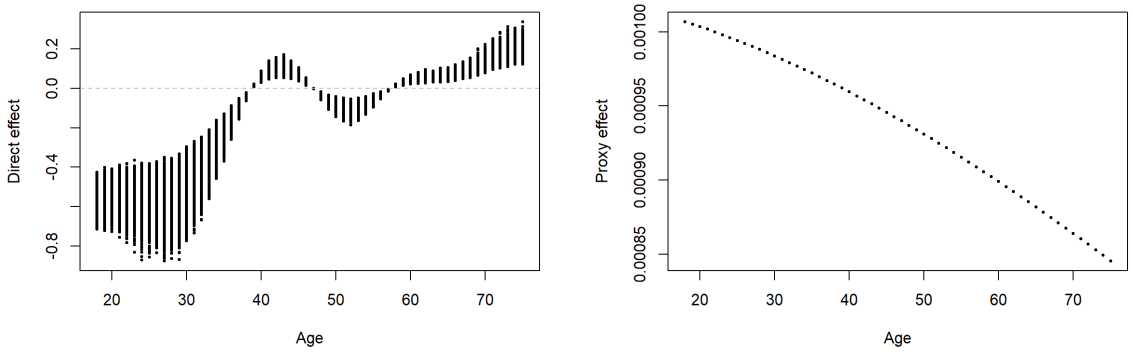


(a) UW price versus DF benchmark.

(b) Distribution of $\widehat{\mu}(\boldsymbol{x}) - \widehat{\mu}^*(\boldsymbol{x})$ by protected group.

FIGURE 2. Comparison of unawareness and discrimination-free prices.

Derivative-based direct and proxy effects. The comparison of unawareness and discrimination-free prices indicates a low level of proxy discrimination in the given dataset. We now calculate the direct and proxy effects of the `Age` variable, according to Definition 1. In the fitted model, the proxy effect of `Age`, $\mathrm{PE}_1(\boldsymbol{x})$, is driven by the `Age`-sensitivity of the classifier $\partial_{x_1}\hat{h}(x_1)$, and the subgroup gap $\hat{\mu}(\boldsymbol{x},1) - \hat{\mu}(\boldsymbol{x},0)$. The direct effect $\mathrm{DE}_1(\boldsymbol{x})$ is linked to the derivative of the smooth age term $s(\texttt{Age})$ in the mean equation. More precisely, under the log link, this derivative enters the direct effect on the price scale through the fitted mean functions $\hat{\mu}(\boldsymbol{x},0)$ and $\hat{\mu}(\boldsymbol{x},1)$, combined according to Definition 1.



(a) Direct effect of `Age`.

(b) Proxy effect of `Age`.

FIGURE 3. Effect of `Age` on the direct and proxy pathways.

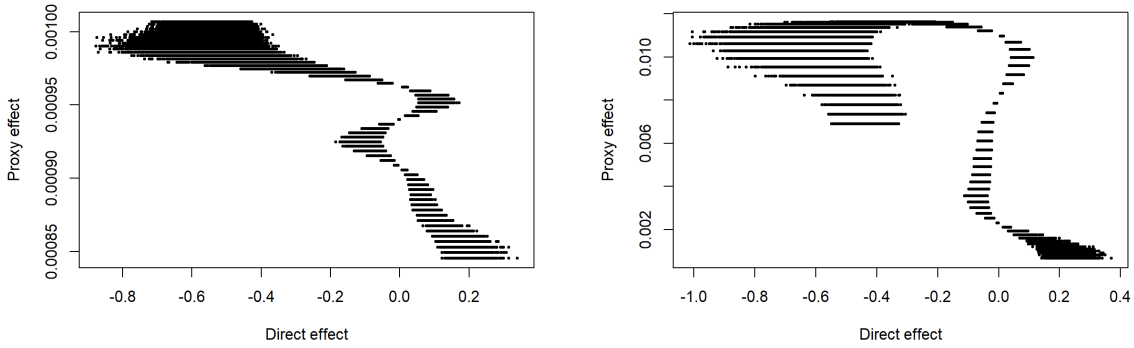


FIGURE 4. Direct versus proxy effects under the baseline model (left) and the distorted model (right).

The direct and proxy effects are plotted against `Age` in Figure 3. The proxy effect remains positive and decreases with `Age`. In absolute value, however, it remains substantially smaller than the direct effect throughout the age range considered. The left panel of Figure 4 shows the relationship between the direct and proxy effects under the baseline model. The proxy effect is consistently small relative to the direct effect, confirming that variation in the predicted premium is driven primarily by the direct channel.

Dependence stress test. The proxy effects observed in this dataset are weak. For that reason, we illustrate the behaviour of the diagnostics introduced in the paper on a distorted dataset, where the dependence between `Age` and D is made more pronounced. Specifically, we construct a synthetic dataset by redrawing the protected attribute from a logistic model with age coefficient $\tilde{\alpha}_1 = -0.1$, compared with the baseline estimate $\hat{\alpha}_1 = -0.0104$. Thus, the age slope in the distorted classifier is about ten times larger in absolute value, meaning that the inferred probability of $D = 1$ changes much more sharply with `Age` than in the original data. At the same time, we keep the best-estimate model fixed and preserve approximately the marginal distribution of D. The construction of the distorted dataset is described in Appendix A.2.

Comparing the two panels of Figure 4, the main effect of this distortion is a clear amplification of the proxy channel: in the distorted model, proxy effects are about ten times larger than in the baseline model. Hence, strengthening the dependence between `Age` and D translates directly into materially larger proxy effects, as intended. Nevertheless, even under this stress scenario, the direct component continues to dominate the overall premium sensitivity.

Proxy effects versus price sensitivity. Because the direct and proxy effects can differ substantially in both scale and sign across policyholders, separate summaries of each component may not be sufficient for screening economic relevance. For example, there may be policyholder scenarios where proxy effects dominate direct effects, but if in those scenarios the overall price sensitivity in `Age` is low, then the proxy effects are still not material. Furthermore, even when the global effects of

proxy discrimination are low, it is of interest to identify specific, potentially vulnerable, policyholder profiles.

We construct a diagnostic plot that combines the comparison of proxy and direct effects with the overall price sensitivity. Specifically, the horizontal axis shows the total price sensitivity,

$$\frac{\mathrm{DE}_1(\boldsymbol{x}) + \mathrm{PE}_1(\boldsymbol{x})}{\mu(\boldsymbol{x})},$$

interpreted as the local proportional change in the unawareness price for a one-unit increase in `Age` (X_1), while the vertical axis shows the proxy-to-direct ratio,

$$\frac{\mathrm{PE}_1(\boldsymbol{x})}{|\mathrm{DE}_1(\boldsymbol{x})|},$$

which measures the proxy effect relative to the magnitude of the direct effect. This joint display distinguishes observations with large proxy shares but negligible total sensitivity, from observations where both the price sensitivity and the proxy contribution to it may be economically material.

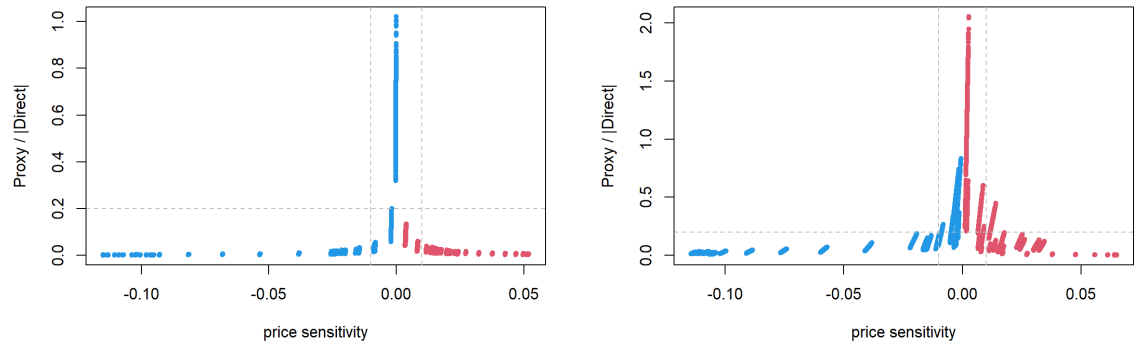
The results are shown in Figure 5a–5b for, respectively, the baseline model and the distorted model (using the dependence stress discussed above). Points are coloured according to the sign of the direct effect. To aid interpretation, we introduce simple screening thresholds for the two plotted quantities ($\pm 1\%$ for price sensitivity and 20% for the proxy-to-direct ratio), which provide a visual benchmark for identifying observations that may warrant closer scrutiny. Observations outside those bands – i.e., where both price sensitivity (in absolute value) and the proxy-to-direct ratio are high – are cause for concern.

In Figure 5a, most observations have small proxy-to-direct ratios, but there is also a noticeable concentration of policies with total price sensitivity close to zero and comparatively large proxy shares, including some values well above the horizontal screening line. In Figure 5b, more observations cross the horizontal materiality threshold and in some cases these correspond also to a non-negligible price sensitivity of $> 1\%$. These are precisely the policyholder scenarios where the potential impact on individual policyholders may need to be investigated.

3. Categorical Variables

In Section 2.3 we defined direct and proxy effects through derivatives of the function $\nu(\boldsymbol{x}, \boldsymbol{x}')$ in a continuous covariate space. In insurance applications, however, permitted rating factors are frequently categorical, such that infinitesimal perturbations (and hence classical derivatives) are not meaningful. In this section we adapt the framework of Section 2.3 by, first, embedding categorical profiles into a continuous space and, second, defining attributions to the direct and proxy channels of price changes resulting from level changes in categorical covariate values.

3.1. **Embedding into a continuous space.** For simplicity of exposition throughout this section, we assume that all components of the permitted covariate vector $\boldsymbol{X} = (X_1, \ldots, X_k)$ are categorical. Let X_j take values in a finite set of levels $\{c_{j1}, \ldots, c_{jl_j}\}$, where l_j denotes the number of levels of X_j. In standard regression modelling, X_j is represented via $l_j - 1$ dummy variables (excluding a



(a) Baseline model

(b) Distorted model (dependence stress)

FIGURE 5. Policyholder-level sensitivity diagnostic in the baseline and high-dependence scenarios.

reference category). This creates a challenge for gradient-based sensitivity analysis, since derivatives with respect to discrete variables are not well defined.

A natural approach is to embed $\boldsymbol{X}$ into a continuous latent space via a transformation

$$\boldsymbol{Z} := T(\boldsymbol{X}) \in \mathbb{R}^m,$$

where T is chosen so that the predictive information in $\boldsymbol{X}$ is preserved. For categorical data, one standard choice is Multiple Correspondence Analysis (MCA) (Abdi and Valentin, 2007). MCA starts from the full dummy-variable (one-hot) representation of the categorical variables and constructs orthogonal continuous coordinates that summarise the main association structure among category levels, in a way analogous to PCA for continuous data.

Here, however, we do not use MCA as a dimension-reduction device. Instead, we retain all non-redundant MCA components, i.e. the full score space corresponding to the rank of the indicator matrix. In this case, the transformation T is information-preserving on the observed categorical design space and can be viewed as a reparameterization of the original categorical design. In particular, when the pricing and inferential models are linear in the full dummy-variable representation of $\boldsymbol{X}$, the same fitted linear predictor can be expressed equivalently in terms of the MCA coordinates $\boldsymbol{z} = T(\boldsymbol{x})$. Since $\boldsymbol{Z}$ is an affine transformation of the original variables $\boldsymbol{X}$, given a linear prediction model like a GLM, we obtain the same predicted values given either $\boldsymbol{X} = \boldsymbol{x}$ or $\boldsymbol{Z} = \boldsymbol{z} = T(\boldsymbol{x})$.

With slight abuse of notation, we therefore write

$$\mu(\boldsymbol{z}) := \mathbb{E}[Y \mid \boldsymbol{Z} = \boldsymbol{z}], \qquad \mu(\boldsymbol{z}, d) := \mathbb{E}[Y \mid \boldsymbol{Z} = \boldsymbol{z},\, D = d], \qquad h(\boldsymbol{z}) := \mathbb{P}(D = 1 \mid \boldsymbol{Z} = \boldsymbol{z}),$$

bearing in mind that, in the present setting, these quantities are the same fitted objects as $\mu(\boldsymbol{x})$, $\mu(\boldsymbol{x}, d)$, and $h(\boldsymbol{x})$, expressed in the embedded coordinate system. We then define

$$\nu(\boldsymbol{z}, \boldsymbol{z}') := \mathbb{E}\big[\mu\big(\boldsymbol{z}, g(\boldsymbol{z}', U)\big)\big]\,, \tag{17}$$

so that $\mu(\boldsymbol{z}) = \nu(\boldsymbol{z}, \boldsymbol{z})$. In the binary case $D \in \{0, 1\}$, this reduces to

$$\nu(\boldsymbol{z}, \boldsymbol{z}') = \mu(\boldsymbol{z}, 0) + h(\boldsymbol{z}')\big(\mu(\boldsymbol{z}, 1) - \mu(\boldsymbol{z}, 0)\big).$$

Hence sensitivity analysis of $\nu(\boldsymbol{z}, \boldsymbol{z}')$ can be carried out along the lines of Section 2.3, provided the relevant functions are differentiable in their continuous arguments.

While MCA is a convenient and transparent embedding for categorical inputs, it is not the only option. Any embedding T that yields a continuous representation and preserves the information needed for pricing and inference can be used with the attribution framework developed below. For example, in the context of neural networks, embedding layers may be used; see Section 4.7 of Wüthrich and Merz (2023).

3.2. **Counterfactual policyholder profiles.** The embedding in Section 3.1 addresses differentiability, but derivatives of $\nu(\boldsymbol{z}, \boldsymbol{z}')$ can remain hard to interpret for categorical inputs: derivatives describe responses to infinitesimal perturbations, whereas meaningful changes in $\boldsymbol{X}$ correspond to finite moves between category labels. We therefore work with counterfactual profiles, understood as structured what-if perturbations of an observed profile.

Specifically, we select a policyholder with observed profile $\boldsymbol{x}$, and construct a perturbed profile $\tilde{\boldsymbol{x}}$ by changing the level of exactly one categorical variable while holding all others fixed. Mapping both profiles to the continuous space yields

$$\boldsymbol{z} = T(\boldsymbol{x}), \qquad \tilde{\boldsymbol{z}} = T(\tilde{\boldsymbol{x}}). \tag{18}$$

The associated price change can be written in $\boldsymbol{Z}$-space as

$$\mu(\tilde{\boldsymbol{x}}) - \mu(\boldsymbol{x}) = \mu(\tilde{\boldsymbol{z}}) - \mu(\boldsymbol{z}) = \nu(\tilde{\boldsymbol{z}}, \tilde{\boldsymbol{z}}) - \nu(\boldsymbol{z}, \boldsymbol{z}).$$

Our goal is to attribute this finite change to the direct and proxy channels.

A natural first step is a first-order Taylor expansion around $(\boldsymbol{z}, \boldsymbol{z})$. Since the perturbed profile $\tilde{\boldsymbol{z}}$ enters both arguments of ν, the same displacement $\Delta\boldsymbol{z} := \tilde{\boldsymbol{z}} - \boldsymbol{z}$ drives both channel contributions. This yields the approximation

$$\nu(\tilde{\boldsymbol{z}}, \tilde{\boldsymbol{z}}) - \nu(\boldsymbol{z}, \boldsymbol{z}) \approx \big[\nabla_{\boldsymbol{z}}\nu(\boldsymbol{z}, \boldsymbol{z})\big]^\top \Delta\boldsymbol{z} \;+\; \big[\nabla_{\boldsymbol{z}'}\nu(\boldsymbol{z}, \boldsymbol{z})\big]^\top \Delta\boldsymbol{z}. \tag{19}$$

While the approximation (19) offers a decomposition into direct and proxy contributions, it is inaccurate in the typical case when ν is non-linear.

To obtain an exact decomposition for finite profile changes, we use the Aumann-Shapley attribution (Billera and Heath, 1978). A simplified exposition in our context is as follows. Define the interpolation path

$$\boldsymbol{z}(\gamma) := \boldsymbol{z} + \gamma(\tilde{\boldsymbol{z}} - \boldsymbol{z}), \qquad \gamma \in [0, 1]. \tag{20}$$

Under differentiability of ν, the fundamental theorem of calculus gives

$$\mu(\tilde{\boldsymbol{z}}) - \mu(\boldsymbol{z}) = \nu(\tilde{\boldsymbol{z}}, \tilde{\boldsymbol{z}}) - \nu(\boldsymbol{z}, \boldsymbol{z}) = \int_0^1 \frac{\mathrm{d}}{\mathrm{d}\gamma}\nu(\boldsymbol{z}(\gamma), \boldsymbol{z}(\gamma))\,\mathrm{d}\gamma.$$

Applying the chain rule in turn yields,

$$\frac{\mathrm{d}}{\mathrm{d}\gamma}\nu(\boldsymbol{z}(\gamma),\boldsymbol{z}(\gamma)) = \sum_{j=1}^{m}(\tilde{z}_j - z_j)\,\partial_{z_j}\nu(\boldsymbol{z}(\gamma),\boldsymbol{z}(\gamma)) + \sum_{j=1}^{m}(\tilde{z}_j - z_j)\,\partial_{z'_j}\nu(\boldsymbol{z}(\gamma),\boldsymbol{z}(\gamma)).$$

This motivates the exact decomposition of price changes defined below.

Definition 2. Consider a given policyholder with baseline and perturbed profiles $\boldsymbol{x}$ and $\tilde{\boldsymbol{x}}$, a continuous embedding T, and the quantities defined in (17), (18), (20). Assume that ν is differentiable at $(\boldsymbol{z}(\gamma),\boldsymbol{z}(\gamma))$, for all $\gamma \in [0,1]$. We define the direct and proxy effects of the profile change $\boldsymbol{x} \to \tilde{\boldsymbol{x}}$ as

$$\mathrm{DE}(\boldsymbol{z},\tilde{\boldsymbol{z}}) = \sum_{j=1}^{m}(\tilde{z}_j - z_j)\int_0^1 \partial_{z_j}\nu(\boldsymbol{z}(\gamma),\boldsymbol{z}(\gamma))\,\mathrm{d}\gamma,$$

$$\mathrm{PE}(\boldsymbol{z},\tilde{\boldsymbol{z}}) = \sum_{j=1}^{m}(\tilde{z}_j - z_j)\int_0^1 \partial_{z'_j}\nu(\boldsymbol{z}(\gamma),\boldsymbol{z}(\gamma))\,\mathrm{d}\gamma.$$

The direct and proxy effects in Definition 2 represent the Aumann–Shapley attributions of the price change obtained by grouping contributions according to the first and second arguments of ν. By construction, we have

$$\mu(\tilde{\boldsymbol{z}}) - \mu(\boldsymbol{z}) = \mathrm{DE}(\boldsymbol{z},\tilde{\boldsymbol{z}}) + \mathrm{PE}(\boldsymbol{z},\tilde{\boldsymbol{z}}).$$

Hence, $\mathrm{DE}(\boldsymbol{z},\tilde{\boldsymbol{z}})$ measures the part of the price change attributable to modifying the profile through the best-estimate surface, while $\mathrm{PE}(\boldsymbol{z},\tilde{\boldsymbol{z}})$ measures the part attributable to the induced change in the conditional distribution of the protected attribute along the same path. These finite-change quantities play the same conceptual roles as the local derivative-based effects $\mathrm{DE}_i(\boldsymbol{x})$ and $\mathrm{PE}_i(\boldsymbol{x})$ introduced in Definition 1, but apply to finite level changes rather than infinitesimal perturbations.

3.3. **Case Study II: Categorical variables.** We now illustrate the categorical-variable framework on a European insurance claim frequency dataset with eleven categorical rating factors – for reasons of data protection both the line of business and the meaning of the covariates – except gender – are obscured. The aim is to operationalise the direct/proxy decomposition from Section 3.2 when meaningful profile changes correspond to level-to-level moves, and to show how an MCA encoding enables the attributions in a continuous coordinate system.

The dataset contains policyholder exposures, a claim count outcome, and a protected attribute D (gender). Throughout, we treat D as prohibited variable for pricing and ask whether the permitted categorical rating factors carry proxy information about D, and whether the fitted unawareness prices are subject to material proxy discrimination.

3.3.1. *Data overview.* The dataset includes eleven categorical covariates $(X_1, X_2, \ldots, X_{11})$. We also have a binary gender variable D (where 0 represents men and 1 represents women) and a discrete variable Y, representing the number of claims made by a policyholder. Exposure is measured in years. An overview of these variables is provided in Table 1.

Table 2 shows that, although the total is roughly balanced, the distribution across Y differs by gender. Cells for $Y = 3, 4$ are based on very small counts and should be interpreted cautiously, and most policyholders have zero claims, with counts declining as Y increases. These statistics indicate

TABLE 1. Overview of Variables in the Insurance Claim dataset.

Variable	Description	Type	Levels / Range
`observation`	Unique ID	Numeric	1 to 172572 (13 missing rows)
`D`	Gender (Protected Attribute)	Binary	0 = Man, 1 = Woman
`Y`	Claim Count	Integer	0 to 4
`duration`	Exposure (years)	Numeric	0.0027 to 5.0000
`X1 – X11`	Categorical Risk Factors	Factor	Various

TABLE 2. Outcome Y by Gender D (counts with row %).

Y	D = 0 (Men)	D = 1 (Women)	Total
0	83,390 (50.4%)	81,977 (49.6%)	165,367
1	3,209 (47.6%)	3,534 (52.4%)	6,743
2	192 (46.4%)	222 (53.6%)	414
3	11 (40.7%)	16 (59.3%)	27
4	5 (62.5%)	3 (37.5%)	8
Total	86,807	85,752	172,559

that women have a slightly higher average claim count, partly due to longer exposure durations. When normalized by exposure, the claim frequency is $0.0471/0.880 \approx 0.0535$ for women and $0.0420/0.795 \approx 0.0528$ for men.

TABLE 3. Summary statistics by Gender.

Gender (D)	Mean Claims (Y)	Std. Dev.	Mean Exposure (Years)
Male	0.0420	0.215	0.795
Female	0.0471	0.227	0.880

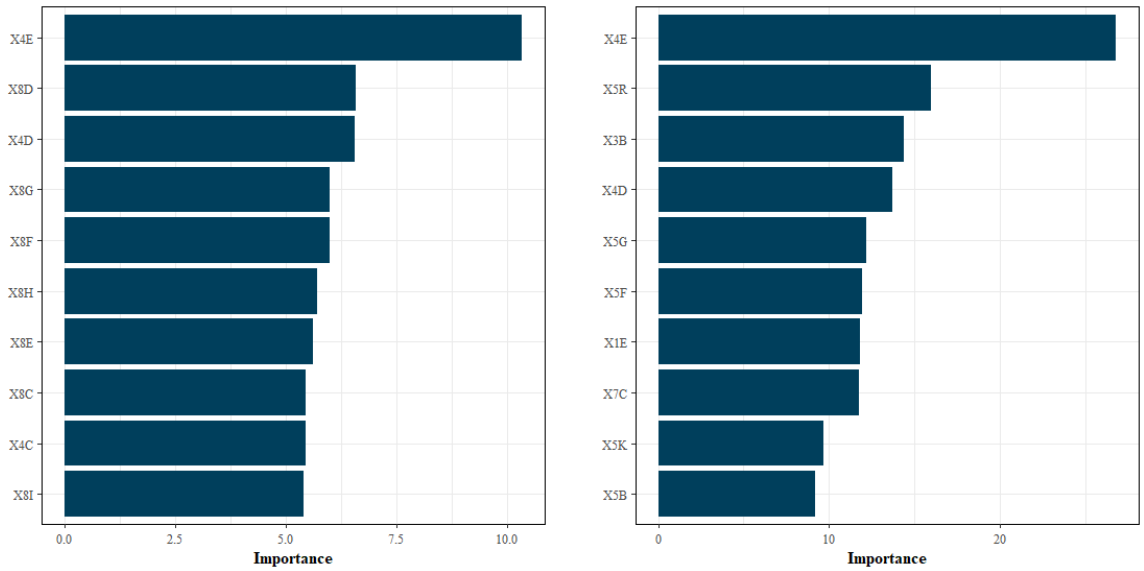
3.3.2. *Model estimation.* We estimate two fitted objects required for the direct/proxy decomposition in Section 3.2: (i) a best-estimate claim-frequency model $\mu(\boldsymbol{x}, d)$, and (ii) a classifier $h(\boldsymbol{x})$.

3.3.3. *Best-estimate Model.* We first fit a quasi-Poisson regression for the claim count Y with log link and exposure offset:

$$\log \mu(\boldsymbol{x}, d) \;=\; \beta_0 + \beta_D d \;+\; \sum_{j=1}^{11} \sum_{r \in \mathcal{R}_j} \beta_{jr}\, \mathbb{1}\{x_j = c_{jr}\} \;+\; \log(\texttt{duration}), \tag{21}$$

where the categorical variables X_j are encoded via dummies, and `duration` enters as an offset. Here $\mathcal{R}_j$ denotes the set of levels of X_j excluding a reference category. We refer to (21) as the best-estimate model because it uses the full information set $(\boldsymbol{X}, D)$.

In this model the coefficient on gender is $\hat{\beta}_D = 0.0745$ ($p = 0.013$), implying that, holding other factors fixed, women have $\exp(0.0745) - 1 \approx 7.7\%$ higher expected claim frequency than men. This supports the outcome-relevance condition for proxy discrimination: $\mu(\boldsymbol{x}, 1) \neq \mu(\boldsymbol{x}, 0)$ for some $\boldsymbol{x}$. Figure 6a shows the permutation-based variable importance plot (VIP; Breiman (2001); Molnar (2022)) for the fitted frequency model (top levels only). The most influential level is X_{4E}, followed by several additional levels of X_4 and X_8. Thus, at the original categorical level, X_4 emerges as one of the most important rating factors for expected claims frequency. The estimated coefficient of X_{4E} is $\hat{\beta}_{X_{4E}} = -0.881$ ($p < 10^{-16}$).



Best-estimate model (quasi-Poisson). Logistic regression for D.

FIGURE 6. Permutation-based variable importance (VIP).

3.3.4. *Classifier.* We fit a logistic regression model to predict gender from the same rating factors $X_1, \ldots, X_{11}$:

$$h(\boldsymbol{x}) = \mathbb{P}(D = 1 \mid \boldsymbol{X} = \boldsymbol{x}) = \text{logit}^{-1}\Big(\alpha_0 + \sum_{j=1}^{11} \sum_{r \in \mathcal{R}_j} \alpha_{jr}\, \mathbb{1}\{x_j = c_{jr}\}\Big). \tag{22}$$

The classifier attains an in-sample accuracy of 0.583 and an AUC of 0.620. Both exceed the 0.5 baseline of random guessing, indicating that D is weakly predictable from $\boldsymbol{X}$. We do not report the full regression results here, but only note that many levels are statistically significant. In particular, X_{4E} has a substantial positive effect ($\hat{\alpha}_{X_{4E}} = 0.911,\ p < 10^{-16}$) on predicting the policyholder being a woman. The corresponding VIP in Figure 6b again ranks X_{4E} at the top, with other levels of X_4 also appearing prominently. Hence X_4 is influential not only for claims frequency, but also for

predicting the protected attribute. Together with $\hat{\beta}_D > 0$ in the frequency model, this creates scope for proxy contributions even when D is omitted from pricing.

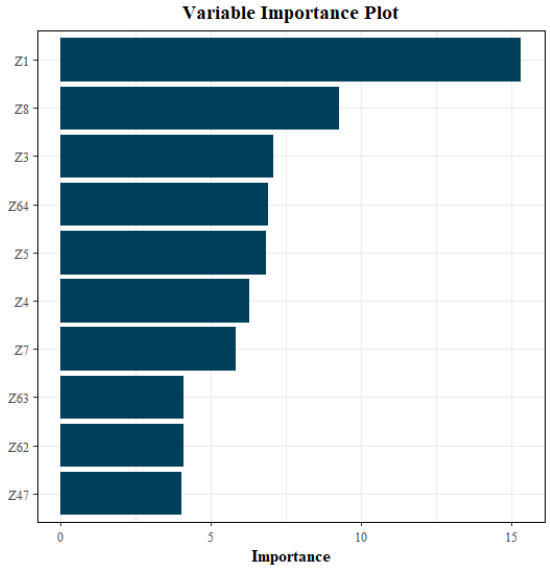
A subtle point is that association patterns can be directionally mixed. In our data, gender has a positive impact on the claims frequency ($\hat{\beta}_D > 0$), while X_{4E} has a negative direct effect on claims frequency ($\hat{\beta}_{X_{4E}} < 0$), and a positive effect on the probability of being a woman ($\hat{\alpha}_{X_{4E}} > 0$). Thus X_{4E}, on the one hand, reduces expected frequency directly, while on the other, increases expected frequency indirectly via the proxy path. This indicates that any proxy-discriminatory effects in this setting will suppress, rather than increase, demographic price disparities. This further motivates decompositions that keep track of both direct and proxy pathways separately.

3.3.5. *Encoding categorical variables.* To enable gradient-based attribution for categorical inputs, we apply Multiple Correspondence Analysis (MCA) (Abdi and Valentin, 2007) to $\{X_1, \ldots, X_{11}\}$. We retain 65 components (denoted $Z_1, \ldots, Z_{65}$), equal to the number of non-redundant one-hot indicators. By construction, keeping all non-redundant components recovers 100% of the total association, so no information from the original categories is discarded.

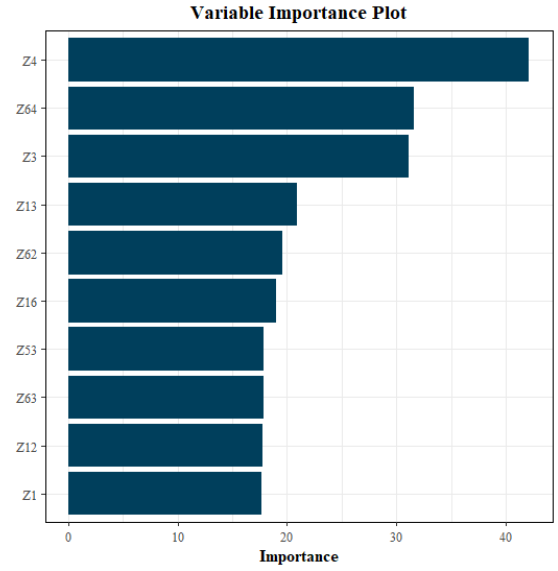
We then fit a quasi-Poisson regression for claims frequency (best-estimate model) using $\boldsymbol{Z}$ and D:

$$\log \mathbb{E}[Y \mid \boldsymbol{Z} = \boldsymbol{z}, D = d] = \beta_0 + \beta_D d + \sum_{r=1}^{65} \beta_r z_r + \log(\texttt{duration}).$$

The gender coefficient remains $\hat{\beta}_D = 0.0745$ ($p = 0.013$). Permutation importance for the projected frequency model is shown in Figure 7a. The most influential MCA coordinates are Z_1, Z_8, Z_3, and Z_{64}.



Best-estimate model fitted on MCA-projected features ($\boldsymbol{Z}$).



Logistic regression for D fitted on MCA-projected features ($\boldsymbol{Z}$).

FIGURE 7. Permutation-based variable importance (VIP).

We also fit a logistic regression for D on the same 65 scores:

$$\mathbb{P}(D = 1 \mid \boldsymbol{Z} = \boldsymbol{z}) = \text{logit}^{-1}\Big(\alpha_0 + \sum_{r=1}^{65} \alpha_r z_r\Big).$$

This classifier's in-sample accuracy and AUC mirror those obtained when fitting the classifier on $\boldsymbol{X}$. Permutation importance is shown in Figure 7b, where the dominant variables are Z_4, Z_{64}, Z_3, and Z_{13}.

To identify the MCA coordinates most closely associated with X_4, we examine the contribution and diagnostics in Appendix B, Figs. 15 and 16. These show that the variation associated with X_4, and in particular with the influential level X_{4E}, is well captured by Dimensions Z_3 and Z_4. This is also consistent with the projected-space VIPs.

3.3.6. *Quantifying direct and proxy effects.* We now calculate the direct and proxy effects of Definition 2, by constructing counterfactual profiles for a randomly chosen sub-sample of 1,000 policyholders (we use 'counterfactual' in the sense of structured perturbations of observed profiles). For each policyholder, we change one covariate level at a time while keeping the rest of the profile unchanged, thereby creating a baseline $\boldsymbol{x}$ and a perturbed profile $\tilde{\boldsymbol{x}}$.

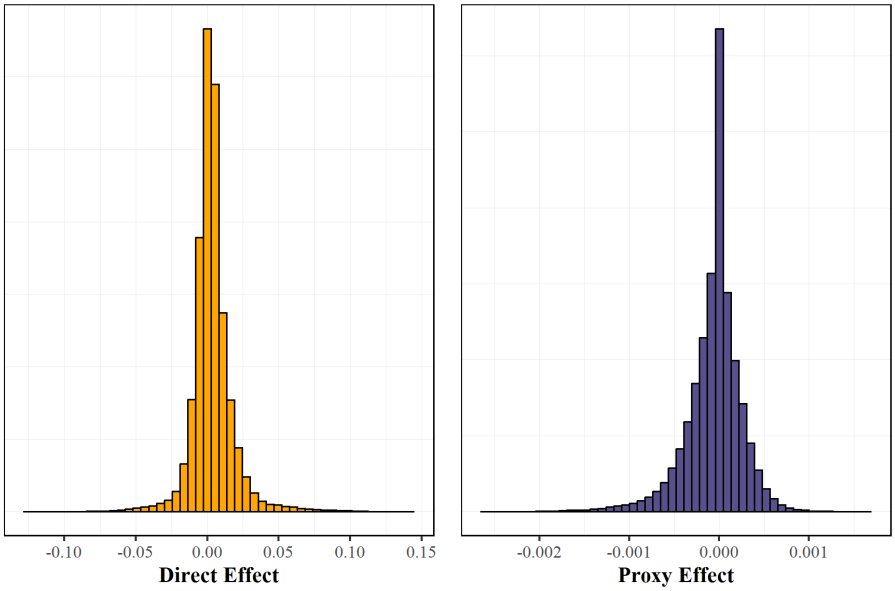


FIGURE 8. Distribution of direct and proxy components across counterfactual level changes.

Figure 8 summarises the distributions of the direct and proxy effects across all counterfactuals. Overall, direct effects dominate: most mass lies within roughly ± 0.03 (95th percentile ≈ 0.0277), with a left tail to larger decreases. Benchmarking to the mean premium $\bar{\mu} = 0.0574$, the median absolute direct change is 0.00551 (about 9.6% of $\bar{\mu}$). Proxy effects are over an order of magnitude smaller: the median absolute proxy change is 1.38×10^{-4} (0.24% of $\bar{\mu}$), and the 95th percentile is 6.37×10^{-4}

(1.11% of $\bar{\mu}$). Hence, for this portfolio, most one-level changes primarily affect predicted premiums through the direct pathway, with proxy contributions acting as small modifiers.

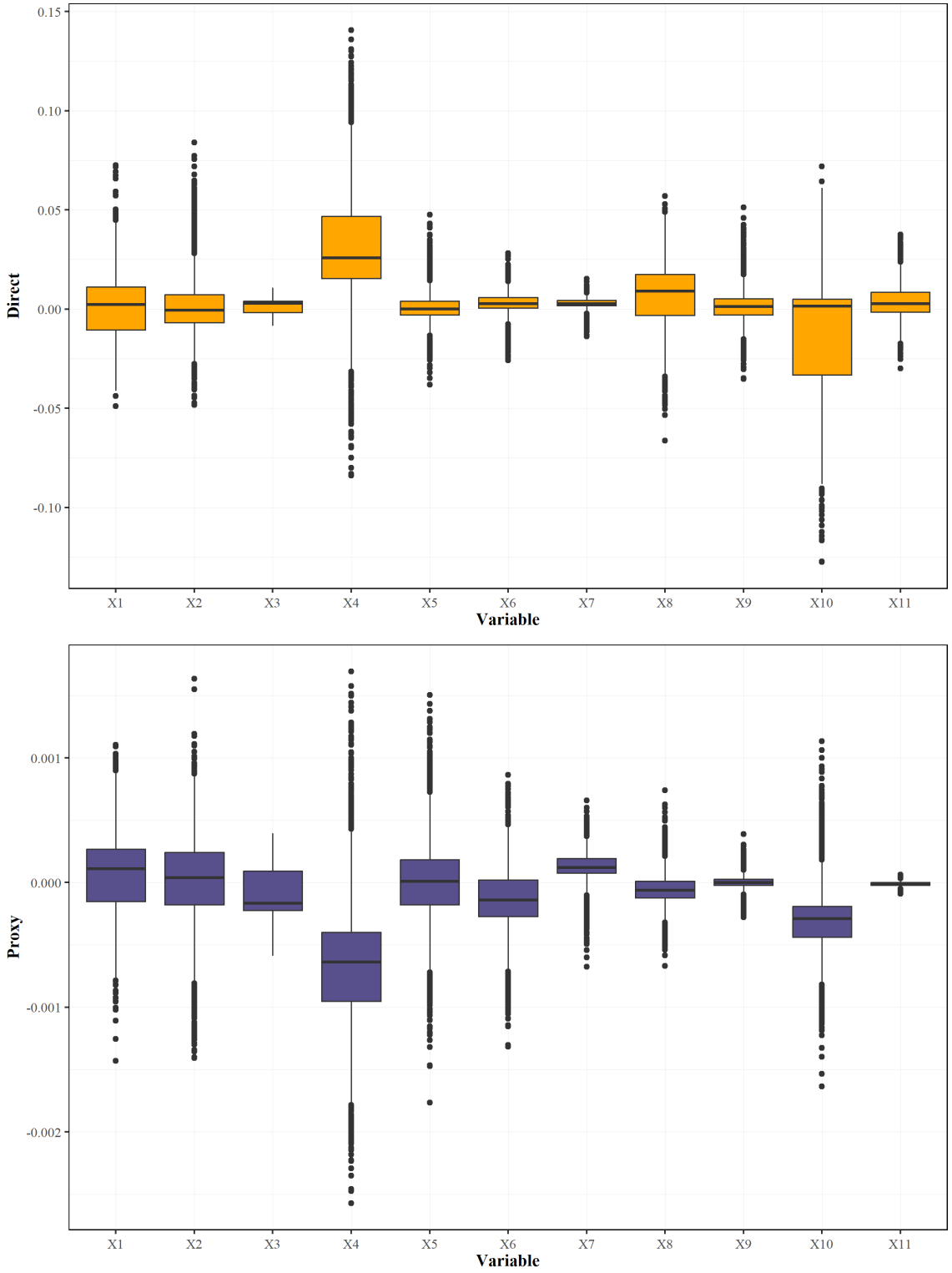


FIGURE 9. Distribution of direct and proxy components by variable across counterfactual level changes.

Figure 9 summarises the distribution of the decomposition components across covariates. The direct component varies substantially by variable, indicating that some covariates have a much stronger effect on predicted premiums than others. In particular, X_4 stands out as one of the variables with the largest and most dispersed direct contributions, suggesting that it plays an important role in the fitted frequency model. The proxy component is much smaller in absolute magnitude throughout. Again, X_4 is among the clearest cases where the proxy contribution shows noticeable dispersion, indicating that changes in this covariate can also affect premiums through the inferred gender probability.

We now complement the portfolio-wide summaries with local, case-by-case diagnostics. Individual counterfactuals reveal which variable-level changes drive a policyholder's premium and how the change decomposes into direct and proxy components. We illustrate the approach on a representative policyholder (ID 100).

Figure 10 provides a complete ranking of admissible level changes for this individual. The largest impacts are concentrated in a few covariates (notably X_{10} and X_4), while many other variables contribute smaller shifts, illustrating how risk information can be dispersed across categorical inputs.

Figure 11 shifts the focus to fairness by reporting, for each admissible level change of Observation 100, the proxy share of the premium change. By proxy share, we mean the percentage of the total decomposed effect attributable to the proxy component, computed using the absolute direct and proxy contributions, so that opposite signs do not cancel. Each bar corresponds to one admissible level change in a categorical covariate, for example, a move such as $X_5 : O \to K$, with all remaining covariates held fixed at their observed values. High proxy shares do not necessarily correspond to large monetary effects: some level changes have small exact impacts but a comparatively larger proxy share, while the largest premium changes typically exhibit small proxy shares. Taken together, Figures 10 and 11 provide two complementary perspectives: the first highlights the size of the premium changes, while the second highlights the relative importance of the proxy pathway.

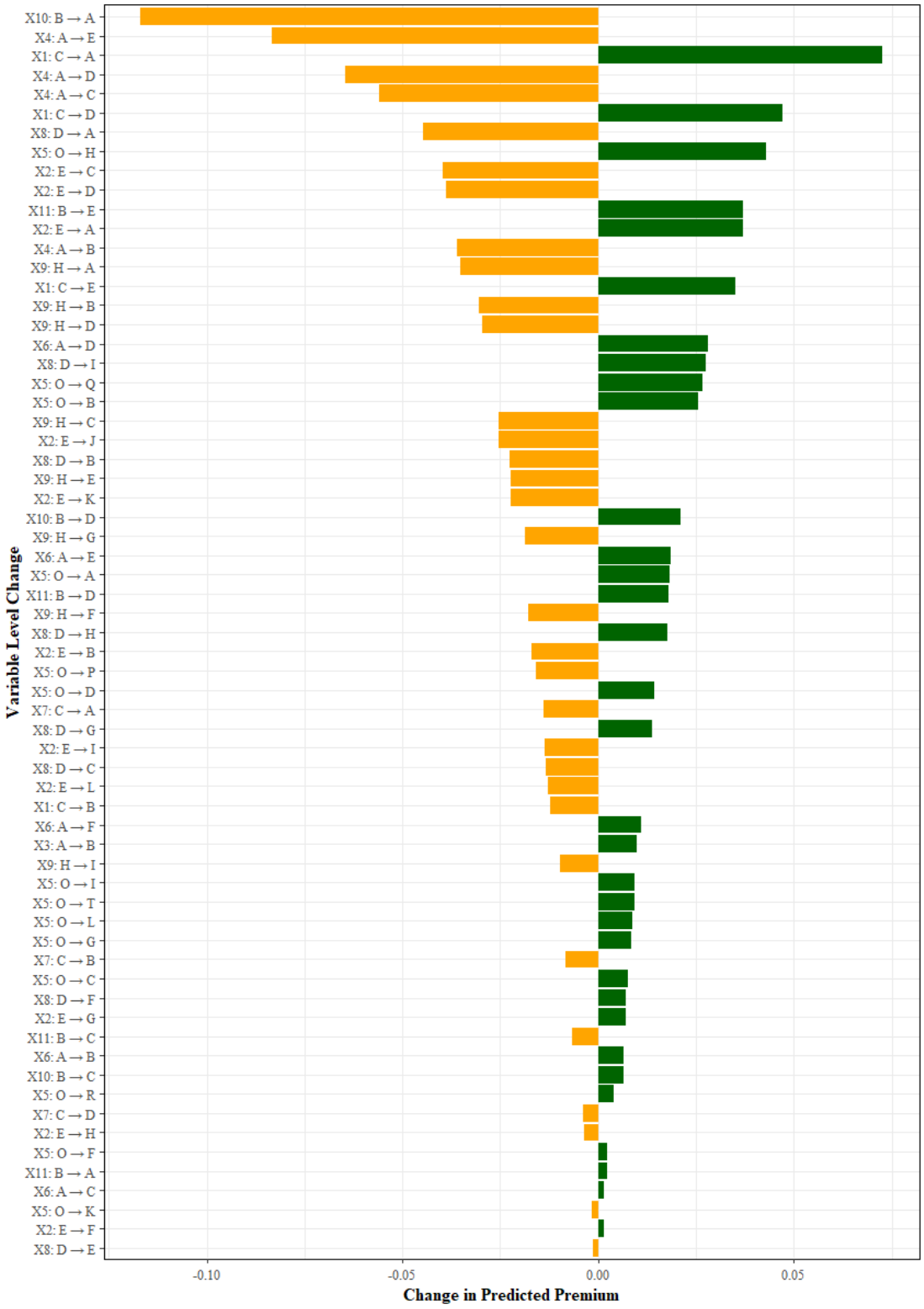


FIGURE 10. Impact of individual variable-level changes on the predicted premium for Observation 100 (exact changes).

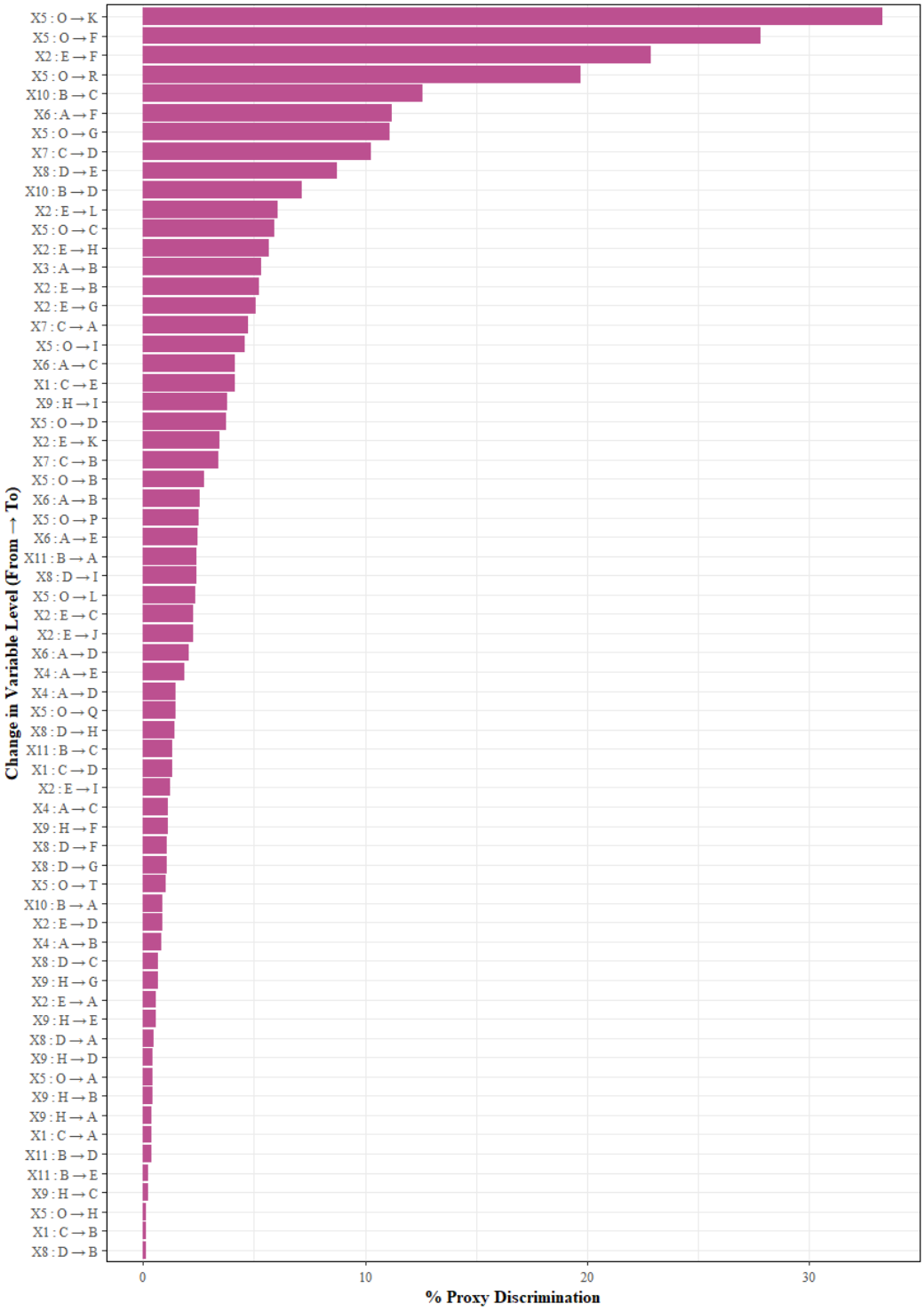


FIGURE 11. Proxy share of the premium change for Observation 100 across admissible level changes.

3.3.7. *ALE diagnostics for selected latent directions.* Since the MCA representation $\boldsymbol{Z}$ is high-dimensional, we do not visualize effects for all components. Instead, following the discussion in Section 3.3.5, we focus on latent directions that are both substantively interpretable and important in the fitted models. In particular, X_4 is one of the most influential original rating factors in both pricing and classification,

and the MCA diagnostics show that its variation is represented primarily by dimensions Z_3 and Z_4. This motivates the ALE analysis along these two latent coordinates.

We compute Accumulated Local Effects (ALE) (Apley and Zhu, 2020) for the bivariate predictor $\nu(\boldsymbol{z}, \boldsymbol{z}')$ with respect to $\boldsymbol{z}_i$ (direct effect) or $\boldsymbol{z}'_i$ (proxy effect), for $i = 1, \ldots, m$. The calculation takes place on the duplicated input $(\boldsymbol{Z}, \boldsymbol{Z}' = \boldsymbol{Z})$. Figure 12 shows that for both Z_3 and Z_4, the direct-channel ALE is substantially larger in magnitude than the proxy-channel ALE, indicating that variation in ν along these latent directions is driven mainly by the direct pricing channel. For both components, the direct ALE increases monotonically across the displayed range, while the proxy ALE decreases monotonically. The effect is especially pronounced for Z_4, whose direct-channel ALE exhibits the largest increase in magnitude among the two coordinates shown, whereas its proxy-channel ALE remains comparatively small and decreasing. Overall, it suggests that the latent directions most closely associated with X_4 affect predictions primarily through the direct channel, with only a limited contribution from the proxy channel. Insofar, the conclusions are consistent with the direct and proxy effects calculated in Section 3.3.6.

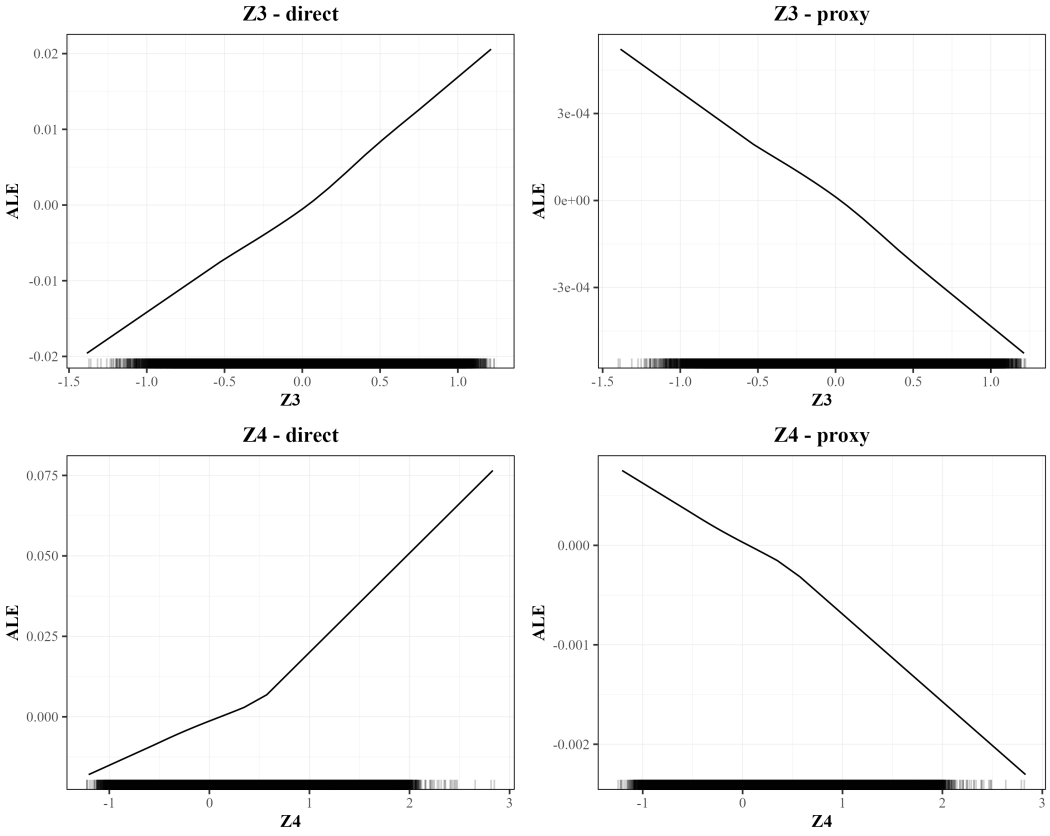


FIGURE 12. Channel-specific ALE diagnostics for the selected MCA components Z_3 and Z_4.

4. Conclusion

This paper developed a diagnostic differential framework for measuring proxy discrimination in insurance pricing. The starting point is a two-argument pricing functional that separates the role of permitted covariates in pricing into two channels: a direct channel, through their effect on expected claims, and a proxy channel, through their association with the protected attribute. Under differentiability, this construction yields local derivative-based measures – termed direct and proxy effects – that make it possible to quantify how small changes in policyholder characteristics affect premiums through each channel separately.

We illustrated the framework in two case studies. In the first, with a continuous covariate setting, we showed how the premium sensitivity can be decomposed into the direct and proxy effects. In the second, where the pricing inputs are largely categorical, we combined the framework with an MCA embedding and counterfactual level changes, allowing discrete premium differences to be analyzed in a continuous latent space. This made it possible to examine proxy effects not only at the portfolio level, but also at the individual policyholder level and for specific variable-level transitions. The general embedding approach is transferable to broader problems of sensitivity analysis in the presence of categorical variables.

Across the empirical analyses, the evidence points to weak but systematic proxy signals. In particular, even when gender is excluded from the pricing inputs, the remaining covariates retain modest ability to predict the protected attribute, indicating that proxy information is encoded in the covariate structure of the portfolio. The derivative-based and counterfactual decompositions show that these proxy pathways are generally small in magnitude relative to the direct pricing effects, but they are widespread and can be identified clearly for particular variables and profile changes. In this sense, the results suggest that proxy discrimination in the portfolio is limited in size, yet still detectable and interpretable.

The monitoring tools we proposed enable an analyst to quantify the materiality – not merely the presence – of proxy discrimination. From the perspective of materiality, the empirical studies produce a negative result, given the small size of measured effects. Such a negative finding is in itself useful for the relevant portfolio holders, since it demonstrates that potentially costly changes in the pricing process are unnecessary. This finding should not be generalised: the dependence of gender on permitted covariates is both line-of-business and portfolio-specific. Moreover, the discussion of other types of sensitive characteristics in the literature demonstrates that the strength of proxy discrimination is highly context dependent. Ethnicity has been shown to produce strong proxy-discriminatory effects in a motor insurance portfolio (Lindholm et al., 2024*a*, 2026), while there is a longstanding discussion on the extent to which credit scores, frequently used as pricing covariates in insurance, encode information on ethnicity and financial deprivation (e.g. Hurlin et al., 2026).

In this paper, we deliberately focused on simple and transparent regression models in order to preserve interpretability and make the direct and proxy channels analytically tractable. Predictive accuracy was therefore not the primary modelling objective. A natural direction for future work

is to extend the proposed framework to more flexible differentiable machine-learning models, such as neural networks, which also permit categorical variables to be embedded through representations learned jointly with the pricing model, thus refining the pre-computed embedding strategy adopted here.

References

Abdi, H. and Valentin, D. (2007), Multiple correspondence analysis, *in* N. J. Salkind, ed., 'Encyclopedia of Measurement and Statistics', SAGE Publications, Thousand Oaks, CA, pp. 651–657.

Apley, D. W. and Zhu, J. (2020), 'Visualizing the effects of predictor variables in black box supervised learning models', *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **82**(4), 1059–1086.

Araiza Iturria, C. A., Hardy, M. and Marriott, P. (2024), 'A discrimination-free premium under a causal framework', *North American Actuarial Journal* **28**(4), 801–821.

Barocas, S., Hardt, M. and Narayanan, A. (2023), *Fairness and Machine Learning: Limitations and Opportunities*, The MIT Press.
URL: *https://fairmlbook.org/*

Barocas, S. and Selbst, A. D. (2016), 'Big data's disparate impact', *California Law Review* **104**(3), 671–732.

Billera, L. J. and Heath, D. C. (1978), 'Allocation of income to members of a cooperative: A game-theoretic approach', *Journal of Business* **51**(1), 37–50.

Breiman, L. (2001), 'Random forests', *Machine Learning* **45**(1), 5–32.

Charpentier, A. (2022), *Insurance, Biases, Discrimination and Fairness*, CRC Press.

Chevalier, D. and Côté, M.-P. (2025), 'From point to probabilistic gradient boosting for claim frequency and severity prediction', *European Actuarial Journal* **15**(3), 707–752.

Côté, O., Côté, M.-P. and Charpentier, A. (2025*a*), 'A fair price to pay: Exploiting causal graphs for fairness in insurance', *Journal of Risk and Insurance* **92**(1), 33–75.

Côté, O., Côté, M.-P. and Charpentier, A. (2025*b*), 'A scalable toolbox for exposing indirect discrimination in insurance rates', *Casualty Actuarial Society Working Paper* .

Dutang, C. and Charpentier, A. (2024), *CASdatasets: Insurance datasets*. R package version 1.2-0.
URL: *https://dutangc.github.io/CASdatasets/*

European Council (2004), 'Council Directive 2004/113/EC implementing the principle of equal treatment between men and women in the access to and supply of goods and services', *Official Journal of the European Union* **L 373**, 37–43.

Frees, E. W. and Huang, F. (2023), 'The discriminating (pricing) actuary', *North American Actuarial Journal* **27**(1), 2–24.

Huang, F. and Pesenti, S. M. (2025), 'Marginal fairness: Fair decision-making under risk measures'. SSRN working paper; version dated May 24, 2025.
URL: *https://ssrn.com/abstract=5266857*

Hurlin, C., Pérignon, C. and Saurin, S. (2026), 'The fairness of credit scoring models', *Management Science* **72**(1), 406–425.

Lindholm, M., Richman, R., Tsanakas, A. and Wüthrich, M. V. (2022), 'Discrimination-free insurance pricing', *ASTIN Bulletin: The Journal of the IAA* **52**(1), 55–89.

Lindholm, M., Richman, R., Tsanakas, A. and Wüthrich, M. V. (2024*a*), 'A multi-task network approach for calculating discrimination-free insurance prices', *European Actuarial Journal* **14**(2), 329–369.

Lindholm, M., Richman, R., Tsanakas, A. and Wüthrich, M. V. (2024*b*), 'What is fair? proxy discrimination vs. demographic disparities in insurance pricing', *Scandinavian Actuarial Journal* **2024**(9), 935–970.

Lindholm, M., Richman, R., Tsanakas, A. and Wüthrich, M. V. (2026), 'Sensitivity-based measures of discrimination in insurance pricing', *European Journal of Operational Research* .

Mara, T. A. and Tarantola, S. (2012), 'Variance-based sensitivity indices for models with dependent inputs', *Reliability Engineering & System Safety* **107**, 115–121.

Marra, G. and Radice, R. (2024), *GJRM: Generalised Joint Regression Modelling*. R package version 0.2-6.7.

Marra, G. and Radice, R. (2025), *Copula Additive Distributional Regression Using R*, CRC Press.

Miao, K. and Pesenti, S. (2026), 'Discrimination-insensitive pricing', *arXiv preprint arXiv:2603.16720* .

Molnar, C. (2022), *Interpretable Machine Learning: A Guide for Making Black Box Models Explainable*, Independently published.
URL: *https://christophm.github.io/interpretable-ml-book/*

Pesenti, S. M., Millossovich, P. and Tsanakas, A. (2021), 'Cascade sensitivity measures', *Risk Analysis* **41**(12), 2392–2414.

Pesenti, S. M., Millossovich, P. and Tsanakas, A. (2025), 'Differential quantile-based sensitivity in discontinuous models', *European Journal of Operational Research* **322**(2), 554–572.

Ponnet, J., Van Oirbeek, R. and Verdonck, T. (2021), 'Concordance probability for insurance pricing models', *Risks* **9**(10), 178.

Rüschendorf, L. and de Valk, V. (1993), 'On regression representations of stochastic processes', *Stochastic Processes and their Applications* **46**(2), 183–198.

Tschantz, M. C. (2022), What is proxy discrimination?, *in* 'Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency', pp. 1993–2003.

Wüthrich, M. V. and Merz, M. (2023), *Statistical Foundations of Actuarial Learning and Its Applications*, Springer.

Xin, X. and Huang, F. (2024), 'Antidiscrimination insurance pricing: Regulations, fairness criteria, and models', *North American Actuarial Journal* **28**(2), 285–319.

Appendix A. Case Study I (pg15training) supplementary material

A.1. **Tables.**

Table 4. Logistic regression (logit link) for D on Age

Term	Estimate	Std. Error	z-value	p-value	Sig.
(Intercept)	-0.1252976	0.0199996	-6.265	3.73e-10	***
Age	-0.0104209	0.0004659	-22.365	$< 2 \times 10^{-16}$	***

Notes: Model: $\text{logit}\big(\mathbb{P}(D = 1 \mid \text{Age})\big) = \alpha_0 + \alpha_1 \,\text{Age}$. Significance codes: *** $p < 0.001$, ** $p < 0.01$, * $p < 0.05$, . $p < 0.1$. Null deviance = 131,323 on 99,999 d.f.; residual deviance = 130,817 on 99,998 d.f.; AIC = 130,821. Binomial dispersion parameter taken to be 1. Fisher scoring iterations: 4.

Table 5. Tweedie best-estimate model (mean equation, μ): parametric terms

Term	Estimate	Std. Error	z-value	p-value	Sig.
(Intercept)	2.9408418	0.0546701	53.793	$< 2 \times 10^{-16}$	***
D	-0.3959950	0.0257070	-15.404	$< 2 \times 10^{-16}$	***
Density	0.0072780	0.0003921	18.560	$< 2 \times 10^{-16}$	***
Bonus	0.0093939	0.0002052	45.781	$< 2 \times 10^{-16}$	***
Occupation: Housewife	0.1065564	0.0319823	3.332	0.000863	***
Occupation: Retired	-0.4136644	0.0795754	-5.198	2.01×10^{-7}	***
Occupation: Self-employed	-0.0619577	0.0360160	-1.720	0.085381	.
Occupation: Unemployed	0.3877791	0.0327239	11.850	$< 2 \times 10^{-16}$	***
Group1	0.0687998	0.0026110	26.350	$< 2 \times 10^{-16}$	***
Group2: M	0.1490439	0.0592960	2.514	0.011952	*
Group2: N	0.3474689	0.0578627	6.005	1.91×10^{-9}	***
Group2: O	0.1933814	0.0697102	2.774	0.005536	**
Group2: P	0.2276631	0.0684465	3.326	0.000881	***
Group2: Q	-0.2327501	0.0531588	-4.378	1.20×10^{-5}	***
Group2: R	-0.4491401	0.0832080	-5.398	6.75×10^{-8}	***
Group2: S	0.0720728	0.0758095	0.951	0.341751	
Group2: T	0.0270115	0.0714415	0.378	0.705362	
Group2: U	-0.0678741	0.0636232	-1.067	0.286056	

Notes: Link for μ: log. Exposure handled via offset log(`Exppdays`/365). Baseline categories are the omitted levels for `Occupation` and `Group2`. Significance: *** $p < 0.001$, ** $p < 0.01$, * $p < 0.05$, . $p < 0.1$.

TABLE 6. Tweedie best-estimate model (mean equation, μ): smooth term

Smooth term	edf	Ref.df	χ^2	p-value
s(Age)	7.057	8.036	2514.0	$< 2 \times 10^{-16}$

TABLE 7. Tweedie best-estimate model (dispersion equation, σ): parametric terms

Term	Estimate	Std. Error	z-value	p-value	Sig.
(Intercept)	5.5607437	0.0310548	179.062	$< 2 \times 10^{-16}$	***
D	0.1254853	0.0146106	8.589	$< 2 \times 10^{-16}$	***
Density	-0.0020250	0.0002281	-8.877	$< 2 \times 10^{-16}$	***
Bonus	-0.0058307	0.0001252	-46.575	$< 2 \times 10^{-16}$	***
Occupation: Housewife	-0.2000925	0.0185356	-10.795	$< 2 \times 10^{-16}$	***
Occupation: Retired	0.7599327	0.0420680	18.064	$< 2 \times 10^{-16}$	***
Occupation: Self-employed	0.1207279	0.0202361	5.966	2.43×10^{-9}	***
Occupation: Unemployed	0.0136857	0.0191991	0.713	0.475950	
Group1	-0.0333010	0.0015127	-22.014	$< 2 \times 10^{-16}$	***
Group2: M	-0.0541571	0.0340257	-1.592	0.111460	
Group2: N	-0.0428171	0.0329598	-1.299	0.193920	
Group2: O	-0.0122893	0.0384486	-0.320	0.749250	
Group2: P	-0.0303685	0.0377081	-0.805	0.420610	
Group2: Q	0.0622814	0.0300688	2.071	0.038330	*
Group2: R	0.1401476	0.0479719	2.921	0.003480	**
Group2: S	0.0574890	0.0413719	1.390	0.164660	
Group2: T	-0.0638536	0.0394146	-1.620	0.105220	
Group2: U	0.0158455	0.0354058	0.448	0.654490	

Notes: Link for σ: log. Baseline categories are the omitted levels for `Occupation` and `Group2`. Significance: *** $p < 0.001$, ** $p < 0.01$, * $p < 0.05$, . $p < 0.1$.

TABLE 8. Tweedie best-estimate model (dispersion equation, σ): smooth term

Smooth term	edf	Ref.df	χ^2	p-value
s(Age)	7.024	8.022	460.6	$< 2 \times 10^{-16}$

TABLE 9. Tweedie best-estimate model (power parameter equation, ν)

Term	Estimate	Std. Error	z-value	p-value	Sig.
(Intercept)	-0.005096	0.012028	-0.424	0.672	

A.2. **Dependence stress test: construction of the distorted portfolio.** The purpose of the stress test in Case Study I is to strengthen the dependence between `Age` and the protected attribute D, while otherwise preserving the baseline fitted specification used to generate the response. The distorted portfolio is constructed in two steps: first, we

re-simulate D from a steeper age-based mechanism; second, conditional on the distorted labels, we generate Y using the baseline fitted Tweedie specification evaluated on the distorted data.

In the baseline portfolio, the inference mechanism is estimated by the logistic regression

$$h(x) = \mathbb{P}(D = 1 \mid \texttt{Age} = x) = \text{logit}^{-1}(\hat{\alpha}_0 + \hat{\alpha}_1 x).$$

To strengthen the dependence between `Age` and D, we keep the observed `Age` values fixed and redraw the protected attribute using a distorted logistic rule

$$\tilde{h}(x) = \text{logit}^{-1}(\tilde{\alpha}_0 + \tilde{\alpha}_1 x),$$

where we set $\tilde{\alpha}_1 = -0.1$, which is larger in absolute value than the baseline estimate $\hat{\alpha}_1 \approx -0.0104$, such that the inferred probability of $D = 1$ changes more steeply with `Age`. For each observation i, we then simulate

$$\tilde{D}_i \sim \text{Bernoulli}\big(\tilde{h}(\texttt{Age}_i)\big), \qquad i = 1, \ldots, n.$$

This leaves the observed age profile unchanged, but strengthens the association between `Age` and D.

Figure 13 compares the baseline fitted probabilities $\hat{h}(\texttt{Age})$ with the distorted fitted probabilities $\widehat{\tilde{h}}(\texttt{Age})$ obtained after refitting the logistic classifier on the distorted labels.

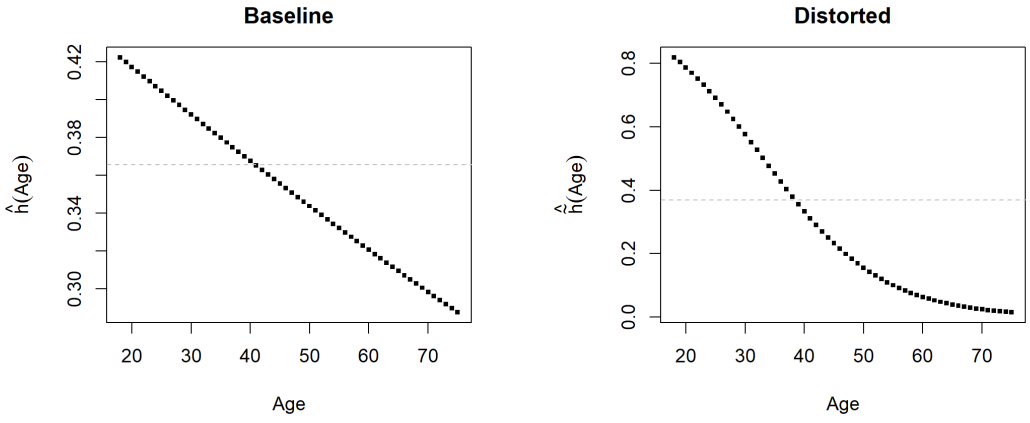


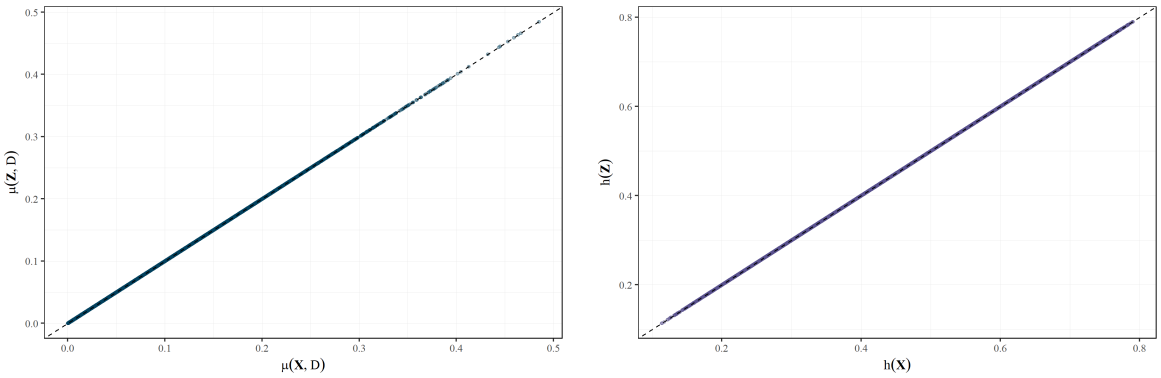
FIGURE 13. Baseline and distorted inferred group probabilities, $\hat{h}(\texttt{Age})$ and $\widehat{\tilde{h}}(\texttt{Age})$, respectively. In each panel, the dashed line indicates the corresponding unconditional share of $D = 1$.

After redrawing D, we generate new claim amounts from the baseline fitted Tweedie model evaluated on the distorted dataset. For each observation i, we recompute the fitted mean $\hat{\mu}_i$ and fitted dispersion $\hat{\phi}_i$ by plugging the distorted label $\tilde{D}_i$ into the baseline fitted mean and dispersion submodels, while keeping the fitted Tweedie power parameter $\hat{p}$ fixed. We then simulate

$$\tilde{Y}_i \mid (\boldsymbol{X}_i, \tilde{D}_i) \sim \text{Tw}(\hat{\mu}_i, \hat{\phi}_i, \hat{p}), \qquad i = 1, \ldots, n.$$

The full estimation pipeline is then re-run on the distorted portfolio: the classifier for $D \mid \texttt{Age}$ is refitted, the best-estimate Tweedie model for $Y \mid \boldsymbol{X}, D$ is refitted, and the resulting fitted models are used to reconstruct the unawareness and discrimination-free prices.

Appendix B. Case Study II supplementary figures



Claim predictions $\mu(\boldsymbol{X}, D)$ versus $\mu(\boldsymbol{Z}, D)$.

Classifier predictions $h(\boldsymbol{X})$ versus $h(\boldsymbol{Z})$.

Figure 14. Comparison of models in the original feature space and the embedded space.

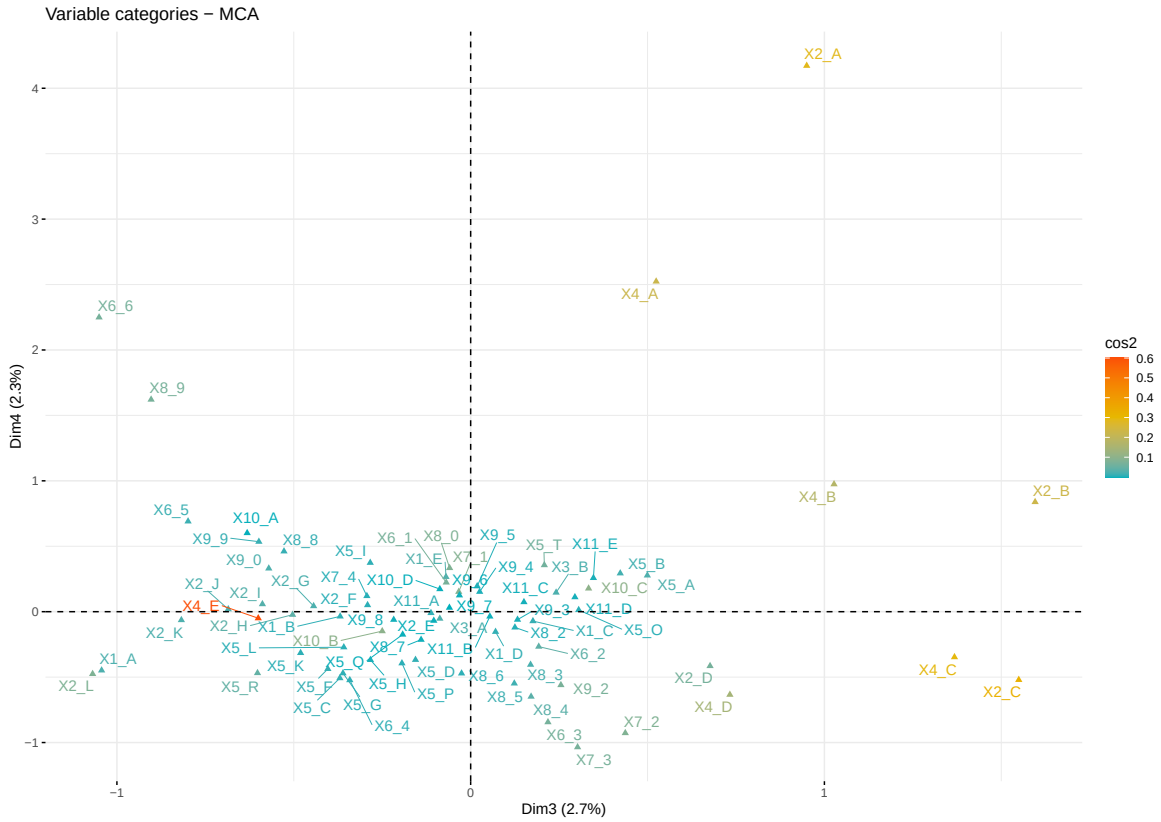


FIGURE 15. Squared cosine (quality of representation) for category levels on Dimensions 3–4. Warmer colours indicate levels that are well represented by these two dimensions. X_{4E} is best represented, with X_{4A} and X_{4C} also strongly represented. Some levels may appear visually close to an axis yet have a low squared cosine; this indicates that they are better explained by other MCA dimensions rather than by Dimensions 3–4.

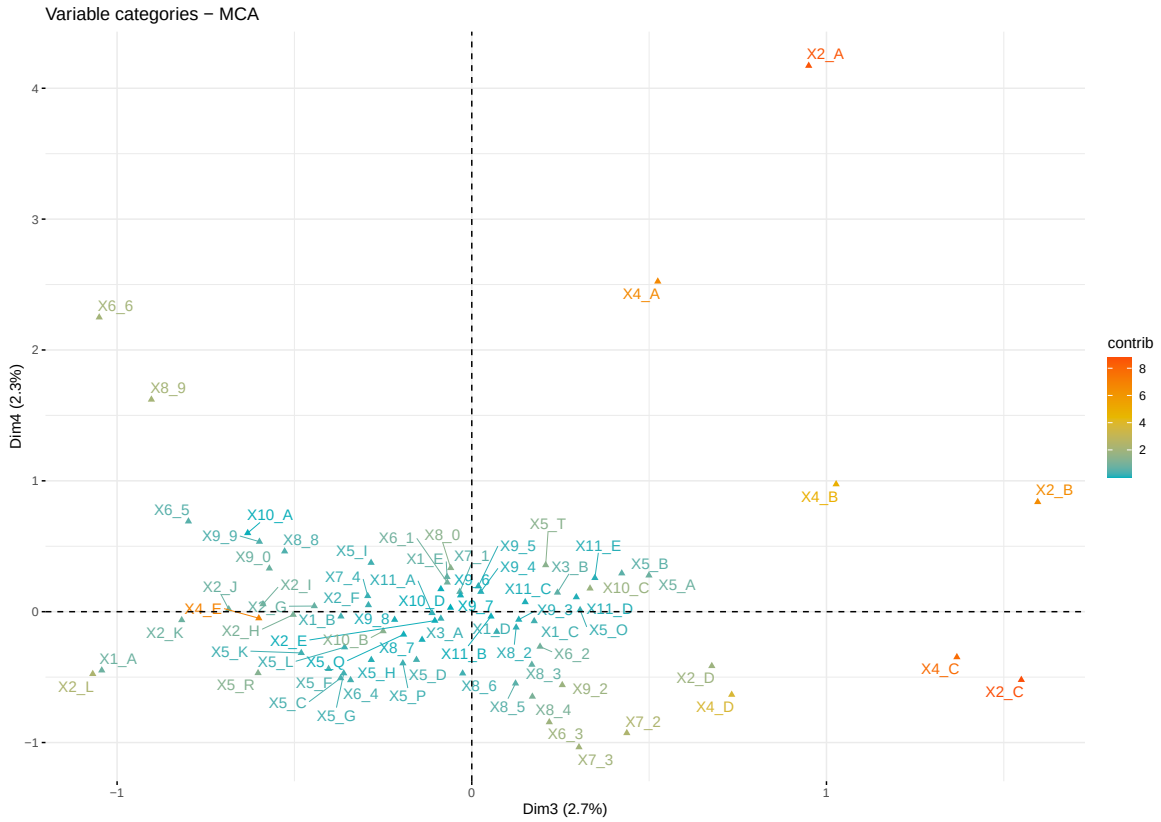


FIGURE 16. Contributions of category levels to constructing Dimensions 3–4. Warmer colours highlight levels that most strongly define these axes. X_{4C}, X_{4A}, and X_{4E} contribute heavily, which, together with their high squared cosine values, shows that key X_4 levels both shape and are well captured by Z_3 and Z_4.