An Optimization-Based Algorithm for Fair and Calibrated Synthetic Data Generation

Jan Pablo Burgard, João Vitor Pamplona, Maria Eduarda Pinheiro

Abstract. For agent based micro simulations, as used for example for epidemiological modeling during the COVID-19 pandemic, a realistic base population is crucial. Beyond demographic variables, health-related variables should also be included. In Germany, health-related surveys are typically small in scale, which presents several challenges when generating these variables. Specifically, strongly imbalanced classes and insufficient observations within sensitive groups necessitate the use of advanced synthetic data generation methods. To address these challenges, we present a method formulated as a mixed-integer linear optimization model designed to create health variables based on class probabilities. This model incorporates fairness by considering the class distribution across sensitive groups as constraints. Furthermore, we prove that the proposed model possesses unimodularity properties and present a preprocessing technique. This allows us to generate data for large populations, such as Germany's population of over 80 million. Our numerical tests, using one of the largest German Health Survey (GEDA), demonstrate that our approach yields better classification results than a standard random forest when considering different ages as sensitive groups.

1. Introduction

The role of machine learning in real-world applications has expanded significantly. For example, machine learning techniques are now widely employed in diverse fields such as agriculture (Attri et al. 2023), earthquake ground motion prediction (Joshi et al. 2023), and cloud computing (Zhang et al. 2024). A prominent application is synthetic data generation, which uses a small sample to create larger, augmented datasets, increasingly replacing or supplementing real-world datasets in various domains (Figueira and Vaz 2022; Murtaza et al. 2023). For example, given a small dataset of customer purchases, including customer ID, product, date, and amount it is possible

Date: October 2, 2025.

Key words and phrases. Machine Learning, Fairness, Mixed Integer Linear Optimization, Health Data.

to generate a larger synthetic dataset, creating new customer profiles with diverse purchasing patterns. Lu et al. (2024) provide a survey on machine learning techniques for synthetic data generation. While this survey covers a wide range of approaches, none of them specifically consider optimization models. The combination of optimization and machine learning, however, is a rapidly expanding research area, as evidenced by the survey Sun et al. (2020).

Moreover, to enhance the representativeness and robustness of this synthetic data, employing a variety of techniques is beneficial, leading to improved model performance. In the context of classification, an ensemble method combines predictions from different machine learning approaches, often including decision trees, to produce more accurate and robust results. This approach enhances the model's ability to generalize when classifying unlabeled data, that is, the data whose labels are unknown.

Generating features for population-representative data requires protecting sensitive groups, a key concern in fair machine learning. For instance, considering older individuals as a sensitive group, protecting them from false negatives in disease diagnoses, such as comorbidities related to severe COVID-19 infections, is important because they have a much higher risk of hospitalization and death. Various fairness metrics address this need. For binary classification, Zafar et al. (2017) introduced metrics and constraints to equalize the proportions of correct and incorrect classifications between two sensitive groups. Burgard and Pamplona (2024a,b) extend these constraints to Support Vector Machines and Logistic Regression for heterogeneous data. For further metrics and surveys on fair machine learning, see Besse et al. (2022), Caton and Haas (2024), and Miron et al. (2020). Specifically, within health data, maximizing accuracy within a sensitive group, such as the elderly, is prioritized over accuracy for the general population. Carrizosa et al. (2025) considers this fairness measure, proposing a Mixed Integer Linear Programming (MILP) model for ensemble training to maximize sensitive group accuracy, focusing on labeled data. In this work, we also address this fairness measure, but within the context of unlabeled data, where classifications are unknown.

Furthermore, in some cases, such as demographic surveys, labeled data is designed to reflect the entire population, necessitating population representativeness. Therefore, it is crucial to maintain the labeled data's class distributions within the unlabeled data, that represents the entire population, and calibrate its classifications. This requires imposing constraints on the classifier to match the total number of elements in each class. In other applications, the total amount of elements in each class within a population is already available, e.g., when external sources provide this information. Burgard et al. (2025b) propose a MILP model to aggregate this calibration information on a random forest for binary classification, while in this work we consider this calibration in the multiclass setting.

We propose a fair and calibrated ensemble method that imposes cardinality constraints on the classification of the unlabeled data, for some sensitive groups. More specifically, our method is a mixed-integer linear programming (MILP) model that incorporates the probabilities of each unlabeled point belonging to each class and the cardinality information of each group. Our main theoretical result shows that the constrained matrix of the proposed MILP model is unimodular. Consequently, to find its solution, it is sufficient to solve a linear programming (LP) model, which has a lower computational cost than the original MILP, allowing us to solve the problem for large instances. Furthermore, we present a theoretical result that leads to a preprocessing technique to fix certain points in specific classes before solving the optimization problem, thereby reducing computational cost.

The last main contribution of this work is the creation of health data variables for the entire Germany, extending the Gesyland dataset, a research project initiated by Burgard et al. (2025a). The SARS-CoV-2 pandemic highlighted the critical need for population health data, including morbidity, cardiac complications, respiratory illnesses, and smoking habits. However, access to this detailed information is often restricted due to privacy regulations and confidentiality concerns. Consequently, fully representative, publicly available health datasets are limited, with some information available only through surveys like the German Health Update, GEDA (Robert Koch Institute 2025), per example. Therefore, the creation of synthetic health datasets that accurately emulate population characteristics is highly beneficial. Survey data can be utilized to expand upon available information and create comprehensive synthetic health datasets. These synthetic datasets are highly beneficial for applications such as agent based microsimulations Ponge et al. (2023).

This paper is organized as follows. In Section 2 we present the optimization problem and the theoretical results regarding unimodularity. After that, the preprocessing technique to fix the class of some point that represents a population unit is discussed in Section 3. There we also present our algorithm that combines this technique and the optimization problem. Numerical results that show the benefices of our approach are reported in Section 4 and in Section 5 we present a computational study on the creation of health information for Germany. Finally, we concluded in Section 6.

2. Calibrated Model

Assume $X_u = \{x^1, \dots, x^m\} \subset \mathbb{R}^d$ represents the unlabeled data with m points $x^i \in \mathbb{R}^d$. Let $X_\ell \subset \mathbb{R}^d$ be the labeled dataset, where the class label of each data point is known. We are interested in solving a classification problem with K classes. That is, we want to assign each point x^i, $i \in [1, m] := \{1, \dots, m\}$ a class $k \in \mathcal{K} := [1, K]$. Besides that, consider $\mathcal{S}_1, \cdots, \mathcal{S}_J$, to be J disjoint subsets of the indices $[1, m]$, that represent J sensitive groups with $[1, m] = \cup_{j=1}^{J} \mathcal{S}_j$. That is, if for instance, $\mathcal{S}_1 = \{1, 3, 4\}$, the unlabeled points x^1, x^3, x^4 belong to the sensitive group 1. In terms of fairness, these groups can represent, for instance, groups of different genders or groups with different population ages. This can also account for the various intersections of characteristics. For example, when considering age and education level, we observe categories such as older individuals with high education, older individuals with low education, younger individuals with high education, and younger individuals with low education.

Our approach is based on the fact that for each $k \in \mathcal{K}$ and $j \in \mathcal{J} := [1, J]$, we know the total number of unlabeled points $\lambda_k^j \in \mathbb{N}$ of the sensitive group $\mathcal{S}_j$ in class k. The probability p_k^i of the unlabeled point x^i belonging to class k is also known, with $\sum_{k \in \mathcal{K}} p_k^i = 1$. The matrix of all these probabilities is denoted by $P = [p^1 \dots p^m] \in \mathbb{R}^{K \times m}$.

The calibrated information λ can be generated using the class distribution of the labeled data X_ℓ. That is, for each class k, we take the proportion of points X_ℓ classified as k and multiplying by the total number of points in the unlabeled data X_u. The information λ also may be available from external sources, e.g., census data or sample surveys. The probabilities p_k^i can be obtained, for example, using an ensemble method that combines predictions from various supervised learning algorithms. Each algorithm might consider a different subset and/or different features of the labeled data X_ℓ to classify each unlabeled point, enabling independent predictions.

Hence, our objective is to classify each unlabeled point $x^i, i \in [1, m]$, such that the number of points in each sensitive group $\mathcal{S}_j$, $j \in \mathcal{J}$, classified as class k is as close as possible to λ_k^j , for each $k \in \mathcal{K}$, taking into account the probabilities p_k^i. In what follows, z_k^i denotes the assignment of x^i to the class k ($z_k^i = 1$ if x^i is classified as k and 0 otherwise).

Moreover, in several applications, if the probability p_k^i is zero for some $i \in [1, m]$, $k \in \mathcal{K}$, it can be a problem if the point x^i is classified as k. This happens, for example, in the case of a disease diagnosis. For this, we assume the following assumption.

Assumption 1. *If, for some $i \in [1, m]$, $k \in \mathcal{K}$, the probability p_k^i of the point x^i belonging to class k is zero, then x^i cannot be classified as k.*

Consider $\mathcal{F}_i = \{k \in \mathcal{K} : p_k^i = 0\}$, for each $i \in [1, m]$, and $\mathcal{G}_k = \{i \in [1, m] : p_k^i = 0\}$, for each $k \in \mathcal{K}$. Then, by Assumption 1, given an point x^i $z_k^i = 0$, for each $k \in \mathcal{F}_i$. On the other hand, given a class $k \in \mathcal{K}$ $z_k^i = 0$, for each $i \in \mathcal{G}_k$

To achieve our goal, we want to find optimal parameters $\eta, \beta \in \mathbb{R}^{K \times J}$ and $z \in \{0, 1\}^{K \times m}$ that solve the optimization problem

$$\min_{\eta,\beta,z} \quad -\frac{1}{m}\sum_{i=1}^{m}\sum_{k\in\mathcal{K}\setminus\mathcal{F}_i} p_k^i z_k^i + \frac{B}{2}\sum_{k\in\mathcal{K}}\sum_{j\in\mathcal{J}}(\eta_k^j + \beta_k^j) \tag{P1a}$$

$$\text{s.t.} \quad \sum_{k\in\mathcal{K}\setminus\mathcal{F}_i} z_k^i = 1, \quad i \in [1, m], \tag{P1b}$$

$$\sum_{i\in\mathcal{S}_j\setminus\mathcal{G}_k} z_k^i - \eta_k^j + \beta_k^j = \lambda_k^j, \quad k \in \mathcal{K}, \quad j \in \mathcal{J}, \tag{P1c}$$

$$z_k^i \in \{0, 1\}, \quad k \in \mathcal{K}\setminus\mathcal{F}_i, \quad i \in [1, m], \tag{P1d}$$

$$\eta_k^j, \beta_k^j \geq 0, \quad k \in \mathcal{K}, \quad j \in \mathcal{J}, \tag{P1e}$$

and $z_k^i = 0$, for each $i \in \mathcal{F}^i$, $k \in \mathcal{K}$. Note that the binary variables z_k^i are the classification variables that model whether the unlabeled point x^i is classified as class k ($z_k^i = 1$) or not ($z_k^i = 0$). Constraint (P1b) ensures that the point x^i is classified to exactly one class k such that $p_k^i \neq 0$. Moreover, for each $k \in \mathcal{K}$, $j \in \mathcal{J}$ $\sum_{i\in\mathcal{S}_j\setminus\mathcal{G}_k} z_k^i$ represents the number of unlabeled points in the sensitive group j classified as k.

Since we have a minimization problem, the variables η_k^j and β_k^j represent the number of points in group j over- and under-classified as k, compared to λ_k^j, respectively. In fact, consider (z^*, η^*, β^*) being an optimal solution of Problem (P1). Constraints (P1c), (P1e) and the second term of objective function in (P1a) lead to:

$$\sum_{i\in\mathcal{S}_j\setminus\mathcal{G}_k} z_k^{i\,*} = \lambda_k^j \implies \eta_k^{j\,*} = \beta_k^{j\,*} = 0,$$

$$\sum_{i\in\mathcal{S}_j\setminus\mathcal{G}_k} z_k^{i\,*} > \lambda_k^j \implies \eta_k^{j\,*} = \sum_{i\in\mathcal{S}_j\setminus\mathcal{G}_k} z_k^{i\,*} - \lambda_k^j,\ \beta_k^{j\,*} = 0,$$

$$\sum_{i\in\mathcal{S}_j\setminus\mathcal{G}_k} z_k^{i\,*} < \lambda_k^j \implies \eta_k^{j\,*} = 0,\ \beta_k^{j\,*} = \lambda_k^j - \sum_{i\in\mathcal{S}_j\setminus\mathcal{G}_k} z_k^{i\,*}.$$

Hence, the smaller the values of η and β, the better the calibrated information is satisfied. If $\eta_k^{j*} = \beta_k^{j*} = 0$, exactly λ_k^j points in sensitive group j are classified as k. These variables are included because, in some cases, it may be unfeasible to have exactly λ_k^j points classified as k. For example, this can happen when the number of points in sensitive group j for which $p_{ik} \neq 0$ is less than λ_k^j. Hence, the objective function in (P1a) is a compromise between classifying each point according to its probabilistic assignment, and satisfying the calibrated information for each class in each sensitive group as closely as possible. Furthermore, the penalty parameter $B > 0$ aims to control the importance of the calibration λ, respectively, The greater the value of B, the more important it is to classify according to the calibrated information λ.

Problem (P1) is a mixed-integer linear programming (MILP) which involves a binary variable for each point-class combination. Consequently the problem becomes increasingly difficult to solve as the number of unlabeled points (m) and the number of classes (K) increase. The following theorem shows that the binary variables z_k^i can be relaxed to continuous variables, simplifying the problem, without compromising the guarantee of the binary values z_k^i in the optimal solution.

Theorem 1. *Consider the linear problem*

$$\min_{\eta,\beta,z} \quad -\frac{1}{m}\sum_{i=1}^{m}\sum_{k\in\mathcal{K}\backslash\mathcal{F}_i} p_k^i z_k^i + \frac{B}{2}\sum_{k\in\mathcal{K}}\sum_{j\in\mathcal{J}}(\eta_k^j + \beta_k^j) \tag{P2a}$$

$$s.t. \quad (\text{P1b}), (\text{P1c}), (\text{P1e}), \tag{P2b}$$

$$z_k^i \geq 0, \quad k \in \mathcal{K}\backslash\mathcal{F}_i, \quad i \in [1, m]. \tag{P2c}$$

All extreme points of the feasible set of this problem are integers. Particularly, there is a solution (z^*, η^*, β^*) to Problem (P2) such that $z_k^{i\,*}$ is binary for each $k \in \mathcal{K}\backslash\mathcal{F}_i$, $i \in [1, m]$.

Proof. Note that Problem (P2) is in standard linear programming form. Let A be the constraint matrix of this problem For illustration, with $m = 4$, $\mathcal{S}_1 = \{1, 2\}$, $\mathcal{S}_2 = \{3, 4\}$ and $K = 2$, $p_2^1 = 0$, A is given by

$$A = \begin{array}{c} \begin{array}{ccccccccccccccc} z_1^1 & z_1^2 & z_2^2 & z_1^3 & z_2^3 & z_1^4 & z_2^4 & \eta_1^1 & \eta_1^2 & \eta_2^1 & \eta_2^2 & \beta_1^1 & \beta_1^2 & \beta_2^1 & \beta_2^2 \end{array} \\ \left[\begin{array}{ccccccccccccccc} 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 & 0 & 0 & -1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 1 & 0 & 0 & -1 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & -1 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 1 & 0 & 0 & 0 & -1 & 0 & 0 & 0 & 1 \end{array}\right] \end{array} \begin{array}{l} \left.\begin{array}{l} \\ \\ \\ \\ \end{array}\right\}\text{(P1b)} \\ \left.\begin{array}{l} \\ \\ \\ \\ \end{array}\right\}\text{(P1c)} \end{array} .$$

For each $k \in \mathcal{K}$ and $j \in \mathcal{J}$, the column of A corresponding to η_k^j (respectively, β_k^j) has exactly one non-zero entry, a -1 (respectively, 1), due to their single appearance in Constraint (P1c).

Moreover, each variable z_k^i, $k \in \mathcal{K} \backslash \mathcal{F}_i$, $i \in [1, m]$, appears in exactly two Constraints, (P1b) and (P1c), each with the coefficient 1. Therefore, each column of A corresponding to z_k^i contains two non-zero entries, one in the row block corresponding to constraint (P1b) and the other in the block corresponding to (P1c).

From the properties described above, it follows that A is a totally unimodular matrix (Theorem in Appendix of Heller and Tompkins (1957)). Hence, by Schrijver (1986), all extreme point of the polytope defined by Constraints (P2b)-(P2c) are integers, which means there is a solution (z^*, η^*, β^*) to Problem (P2) that is integer.

Since Constraints (P1b) and (P2c) ensure that, for each $k \in \mathcal{K} \backslash \mathcal{F}_i$, $i \in [1, m]$, $0 \leq z_k^{i\,*} \leq 1$ holds, $z_k^{i\,*}$ is binary for each $k \in \mathcal{K} \backslash \mathcal{F}_i$, $i \in [1, m]$. Now, instead of solving the MILP Problem (P1) it is enough to solve the LP Problem (P2) that has a lower computational cost.

□

3. Optimal Fair Calibrated Algorithm

3.1. **Preprocessing.** Note that Assumption 1 prevents points with zero probability of class assignment from being classified. Hence, if a point has probability 1 of belonging to a class, this point can be classified before solving Problem (P2) and the calibrated information must be updated. It is what is shown in the Proposition below.

Proposition 1. *Under Assumption 1, for each $i \in [1, m]$, if $p_{\bar{k}}^i = 1$ for some $\bar{k} \in \mathcal{K}$, then every feasible point (z, η, β) of Problem (P2) satisfies $z_{\bar{k}}^i = 1$.*

Moreover, for each $k \in \mathcal{K}$ and $j \in \mathcal{J}$ consider

$$\phi_k^j := \{i \in \mathcal{S}_j : p_k^i = 1\}.$$

Then, $|\phi_k^j|$ points in the sensitive group j are already classified as k and due to the calibration information λ_k^j, $\min\{0, \lambda_k^j - |\phi_k^j|\}$ points in the sensitive group j must be classified as k.

Proof. Note that $\sum_{k\in\mathcal{K}} p_k^i = 1$. Hence, $p_{\bar{k}}^i = 1$ implies $p_{\hat{k}}^i = 0$, for all $\hat{k} \in \mathcal{K}\backslash\{\bar{k}\}$. That is, $\mathcal{F}_i = \mathcal{K}\backslash\{\bar{k}\}$. Thus, because of Assumption 1, Constraint (P1b) becomes $z_{\bar{k}}^i = 1$. That is, x^i is classified as $\bar{k}$.

This means that for each $k \in \mathcal{K}$ and $j \in \mathcal{J}$, $|\phi_k^j|$ points are already classified as k. If $|\phi_k^j| \geq \lambda_k^j$, due to the calibrated information λ_k^j, no remaining point $x^i \in X_u\backslash(\cup_{k\in\mathcal{K}}\phi_k)$ such that $i \in \mathcal{S}_j$ must be classified as k. On the other hand, if $|\phi_k^j| < \lambda_k^j$, only $\lambda_k^j - |\phi_k^i|$ remain points in the sensitive group j must be classified as k. This update is presented in Step 6 of Algorithm 1. □

Algorithm 1 integrates all previously proposed ideas, from preprocessing to the optimization problem itself. Phases 1, 2 and 3 are explained in Proposition 1, Assumption 1 and Theorem 1, respectively.

Algorithm 1: Optimal Fair Calibrated Algorithm (OFCA)

```
0. Input : $P \in \mathbb{R}^{K\times m}$, $\lambda \in \mathbb{N}^{K\times J}$, $\mathcal{S} \subset \mathbb{R}^d$, $\mathcal{C} = \emptyset \in \mathbb{R}$, $\mathcal{F} = \emptyset \in \mathbb{R}^m$,
           $\mathcal{G} = \emptyset \in \mathbb{R}^K$, $C > 0$.

1. Preprocessing (Proposition 1) :
for $j \in \{1, \cdots, J\}$ do
    for $i \in \mathcal{S}_j$ do
        if $\max_{k\in\mathcal{K}}\{p_k^i\} = 1$ then
            Find $\bar{k} = \arg\max_{k\in\mathcal{K}} p_k^i$ and set $\mathcal{C} \leftarrow \mathcal{C} \cup \{i\}$.
            if $\lambda_{\bar{k}}^j > 0$ then
                Set $\lambda_{\bar{k}}^j \leftarrow \lambda_{\bar{k}}^j - 1$.
            end
        end
    end
end

2. Fixing probabilities (Assumption 1) :
for $i \in \{1, \dots, m\} \backslash \mathcal{C}$ do
    for $k \in \{1, \dots, K\}$ do
        if $p_k^i = 0$ then
            Set $\mathcal{F}_i \leftarrow \mathcal{F}_i \cup \{k\}$, $\mathcal{G}_k \leftarrow \mathcal{G}_k \cup \{i\}$, and $z_k^{i*} := 0$.
        end
    end
end

3. Solving the optimization problem (Theorem 1) :
```

$$
\begin{aligned}
\min_{\eta,\beta,z} \quad & -\frac{1}{m}\sum_{i=1}^{m}\sum_{k\in\mathcal{K}\backslash\mathcal{F}_i} p_k^i z_k^i + \frac{B}{2}\sum_{k\in\mathcal{K}}\sum_{j\in\mathcal{J}}(\eta_k^j + \beta_k^j) \\
\text{s.t.} \quad & \sum_{k\in\mathcal{K}\backslash\mathcal{F}_i} z_k^i = 1, \quad i \in [1,m]\backslash\mathcal{C}, \\
& \sum_{i\in\mathcal{S}_j\backslash(\mathcal{C}\cup\mathcal{G}_k)} z_k^i - \eta_k^j + \beta_k^j = \lambda_k^j, \quad k \in \mathcal{K}, \quad j \in \mathcal{J}, \\
& z_k^i \geq 0, \quad k \in \mathcal{K}\backslash\mathcal{F}_i, \quad i \in [1,m]\backslash\mathcal{C}, \\
& \eta_k^j, \beta_k^j \geq 0, \quad k \in \mathcal{K}, \quad j \in \mathcal{J}.
\end{aligned}
$$

to compute z^*, η^*, β^*.

4. Numerical Tests

In this section, we conduct numerical tests using only real-world data from GEDA to demonstrate the effectiveness of the proposed methods.

The GEDA dataset refers to the German Health Update surveys, which are part of a larger effort to monitor the health status and behaviors of the German population. GEDA is a series of cross-sectional surveys designed to collect comprehensive health data across various dimensions, including general health, lifestyle factors, social determinants of health, and the prevalence of chronic diseases. The GEDA surveys are primarily conducted through telephone interviews, where trained interviewers contact individuals from a representative sample of the German population. These telephone surveys allow researchers to gather self-reported health data directly from participants in a structured format. This method ensures that the data collected is from a diverse group of people across different regions, age groups, and socioeconomic backgrounds, providing a comprehensive picture of the country's overall health status. The GEDA dataset contains approximately 20 000 data points.

Moreover, when dealing with health-related variables, it is important to consider ages as sensitive groups. Additionally, since we are creating an entire population, we want to maintain the same labeled data class distribution as in GEDA.

Our numerical test has been implemented in `Julia 1.11` (Bezanson et al. 2017). Initially, we want to demonstrate the statistical importance of the cardinality constraint and our preprocessing technique to speed up the solution process. To this, we conduct numerical tests using only the GEDA dataset.

Each variable with two classes described in Sections 4.1 was considered as the label to be generated, i.e. $K = 2$. We consider $\mathcal{S}_1$ the indices of points representing individuals under 60 years of age (84.37 % of the the dataset) and $\mathcal{S}_2$ as the indices of points representing individuals over or equal 60 years of age (15.63 % of the the dataset).

Table 1 explains the information utilized for the variables generation. For each variable we created 100 samples with 5 % of the data being labeled.

Formula	Variables used in prediction (See descriptions in Appendix A)
Labels	Possible outputs of the classification process
Percentage	Percentage of labels in the GEDA dataset
Percentage in $\mathcal{S}_1$	Percentage of labels for sensitive group $\mathcal{S}_1$ in the GEDA dataset.
Percentage in $\mathcal{S}_2$	Percentage of labels for sensitive group $\mathcal{S}_2$ in the GEDA dataset.

Table 1. Description of information utilized for variable generation.

Afterward, we generated the probabilities p_k^i. For it, we used the `DecisionTree` package (Sadeghi et al. 2022) with the following parameters:

Random Forest Parameter	**Value**
Number of Trees	$\max(20 \times K, 100)$
Portion Sample	63%
Tree Depth	$\min(2 \times K \times V, 50)$

Table 2. Random Forest Parameters

where V is the number of variables used in the prediction, i.e., the cardinality of the information *Formula* in the tables in Section 4.1. We then compared the following approaches:

- MV: Random Forest with majority vote.
- P1: Problem (P1) with $B = 2$.
- OFCA as described in Algorithm 1 with $B = 2$.

We used `JuMP` (Lubin et al. 2023) and `HiGHS` (Huangfu and Hall 2018) to solve Problem (P1) and the one presented in Algorithm 1 (OFCA). For both P1 and OFCA, the value $B = 1$ was selected to give equal importance to the probabilistic assignments and the classification respecting the calibrated information.

For each variable, we first evaluate the classification concerning the calibrated information λ. This evaluation examines the classification performance of each approach within each sensitive group j specifically focusing on over- and underclassification. For a given approach $\mathbb{A}$ let $\theta(\mathbb{A})_j^k$ represent the number of points in the sensitive group j classified as k by this approach. We then compute the cardinality error rate with respect to the calibration λ:

$$\text{CER} := \frac{\sum_{k\in\mathcal{K}} |\lambda_k^j - \theta(\mathbb{A})_k^j|}{2|S_j|} \in [0,1], \tag{1}$$

where lower values indicate a better classification.

Note that CER can be greater than zero even for OFCA. This occurs for two reasons. First, the probability of a point in a sensitive group j being classified into a class k might be equal to 1 for more than λ_k^j points. In this case, because of Proposition 1 more than λ_k^j points are classified as class k. Second, the objective function present in OFCA balances two goals: classifying each point based on its probabilistic assignment and satisfying the calibrated information. Since the calibrated information is a soft goal and not a strict requirement, the model may classify more or fewer points than the required λ_k^j for a given sensitive group j and class k.

Moreover, for each approach, since the true labels of all points are known in the simulation, we categorize them into four distinct categories: true positive (TP) or true negative (TN) if the point is classified correctly in classes 1 or 2, respectively, as well as false positive (FP) if the point is misclassified in the class 1 and as false negative (FN) if the point is misclassified in the class 2. Since the proportion of the points in each class is imbalanced (as will be seen in the next sections), the measure "accuracy" alone can misrepresent classification performance. Consequently, we calculate the F1-score as a more balanced measure of classification (Chicco and Jurman 2020). The F1-score is harmonic mean of its precision and recall, with a higher value indicating better overall performance in correctly identifying positive instances while minimizing both false positives and false negatives. It is given by

$$\text{F1} := \frac{2\text{TP}}{2\text{TP} + \text{FP} + \text{FN}} \in [0,1], \tag{2}$$

different from CER, higher value indicates a better classification. Since Problem (P1) and OFCA solve the same problem, we only compare the cardinality error rate and F1-score of MV and OFCA.

Finally, we evaluate the different methods based on their runtime. To facilitate a comparison of these runtimes we consider all the 700 instances and utilize empirical cumulative distribution functions (ECDFs). Specifically, for $\mathcal{A}$ being a set of approaches and for $\mathcal{P}$ being a set of problems, we denote by $t_{p,a} \geq 0$ the run time of the approach $a \in \mathcal{A}$ applied to the problem $p \in \mathcal{P}$ in seconds. With these notations, the performance profile of approach a is the graph of the function $\gamma_a : [0,\infty) \to [0,1]$ given by

$$\gamma_a(\sigma) = \frac{1}{|\mathcal{P}|}|\{p \in \mathcal{P} \colon t_{p,a} \leq \sigma\}|.$$

4.1. **Description of the variables.** The following tables present the description of the variables.

KHlip12A (Elevated blood cholesterol within the last 12 months)	
Formula	age5B, gender, MIcitizen, KAgfka, KAgfmk
Labels	“1” = Yes, “2” = No
Percentage	“1” = 21.27 %, “2” = 78.73 %
Percentage in $\mathcal{S}_1$	“1” = 19.89 %, “2” = 81.11 %
Percentage in $\mathcal{S}_2$	“1” = 34.28 %, “2” = 65.62 %

KHhyp12 (Hypertension: Within the last 12 months)	
Formula	age5B, gender, MIcitizen, KAgfka, KAgfmk
Labels	“1” = Yes, “2” = No
Percentage	“1” =30.04 %, “2” = 69.96 %
Percentage in $\mathcal{S}_1$	“1” = 25.60 %, “2” = 74.40 %
Percentage in $\mathcal{S}_2$	“1” = 54.05 %, “2” = 45.95 %

KHsa12 (Stroke: Within the last 12 months)	
Formula	age5B, gender, KHlip12A, KHhyp12
Labels	“1” = Yes, “2” = No
Percentage	“1” =1.69 %, “2” = 98.31 %
Percentage in $\mathcal{S}_1$	“1” = 1.24 %, “2” = 98.76 %
Percentage in $\mathcal{S}_2$	“1” = 4.14 %, “2” =95.86 %

PKdep12 (Depression: Within the last 12 months)	
Formula	age5B, gender, MIcitizen, SDisced11z, KAgfka
Labels	“1” = Yes, “2” = No
Percentage	“1” = 9.11 %, “2” = 90.89 %
Percentage in $\mathcal{S}_1$	“1” = 9.46 %, “2” = 90.54 %
Percentage in $\mathcal{S}_2$	“1” = 7.18 %, “2” = 92.82 %

KHdiabB12 (Diabetes: Within the last 12 months)	
Formula	age5B, gender, KHhyp12, SDisced11z, PAbmiB_k2
Labels	“1” = Yes, “2” = No
Percentage	“1” = 9.08 %, “2” = 90.92 %
Percentage in $\mathcal{S}_1$	“1” = 7.37 %, “2” = 92.63 %
Percentage in $\mathcal{S}_2$	“1” = 18.30 %, “2” = 81.70 %

KHab12 (Asthma: Within the last 12 months)	
Formula	age5B, gender, KAgfka, RCstatE
Labels	“1” = Yes, “2” = No
Percentage	“1” = 7.83 %, “2” = 92.17 %
Percentage in $\mathcal{S}_1$	“1” = 7.92 %, “2” = 92.08 %
Percentage in $\mathcal{S}_2$	“1” = 7.27 %, “2” = 92.63 %

KHcb12B (Chronic bronchitis: Within the last 12 months)	
Formula	age5B, gender, KAgfka, RCstatE
Labels	“1” = Yes, “2” = No
Percentage	“1” =5.66 %, “2” = 94.34 %
Percentage in $\mathcal{S}_1$	“1” = 4.87 %, “2” = 95.13 %
Percentage in $\mathcal{S}_2$	“1”= 9.97 %, “2” = 90.03 %

KHdge12 (Arthrose: Within the last 12 months)	
Formula	age5B, gender, KAgfka, KAgfmk
Labels	“1” = Yes, “2” = No
Percentage	“1” = 19.44 %, “2” = 80.56 %
Percentage in $\mathcal{S}_1$	“1” = 15.79 %, “2” = 84.21 %
Percentage in $\mathcal{S}_2$	“1” = 39.25 %, “2” = 60.75 %

4.2. **Results.** As already mentioned, for the cardinality error rate (CER), a lower value is preferable, while for the F1 metric, a higher value is desired. Besides that, the box in the boxplots depicts the range of the medium 50 % of the values; 25 % of the values are below and 25 % are above the box.

The results are presented in Figure 1, where OFCA$_j$ and MV$_j$ indicate the performance of OFCA and MV (respectively) in the sensitive group j. With respect to all the 8 variables presented in this section, Figure 1 (left) shows that our approach prioritizes respecting the calibration information during classification, for all variables almost 100 % of the cases has cardinality error rate (CER) equal to 0, as indicated by OFCA_1 and y OFCA_2. The majority vote has a higher CER, mainly for the sensitive group of older individuals, as indicated by MV_2. Figure 1 (right) shows that OFCA has a higher value of F1 score than MV when compared in each sensitive group. This means that considering the calibration information, as our approach does, improves the F1 score of each sensitive group.

Figure 2 displays the Empirical Cumulative Distribution Functions (ECDFs) for the measured run times. As expected, MV is the fastest algorithm, since it not consider any optimization model. Notably, OFCA outperforms P1, solving all instances in almost 0.50 seconds, while P1 required 0.60 seconds. This demonstrates the effectiveness of the preprocessing technique. We also expect OFCA to be significantly faster than P1 as the number of points increases.

5. Construction of health variables in a georeferenced synthetic German population

In this section, we continue to develop the ideas proposed in Burgard et al. (2025a), where demographic variables are created for a synthetic, georeferenced population of Germany, the Gesyland.eu. From Burgard et al. (2025a), we derive our initial variables, which include age, gender, nationality, educational level (ISCED), and the federal state in which the individual resides. These variables serve as the foundation for the synthetic population. By utilizing this dataset in conjunction with the German Health Update (GEDA) data, we can apply the ideas proposed in this work. Specifically, we aim to demonstrate the impact of considering the total number of points in each class relative to each sensitive group.

In this part of the computational study, we focus on the ongoing development of the Gesyland synthetic population. Specifically, we generate the same variables for the Gesyland dataset that were used in the numerical studies mentioned above, employing the GEDA as a training set and utilizing the distribution information derived from the GEDA to generate the calibration information λ. Note that the Gesyland dataset contains approximately 85 million data points. Hence, to reduce computational cost, all variables were generated separately for each federal state within Germany, using the Algorithm 1 (OFCA). That is, for each federal state, we use random forests to calculate probabilities p_k^i and compute the information λ, which were required to match the distribution of each individual region. Since we do not know the true labels, we cannot measure the quality of the classification

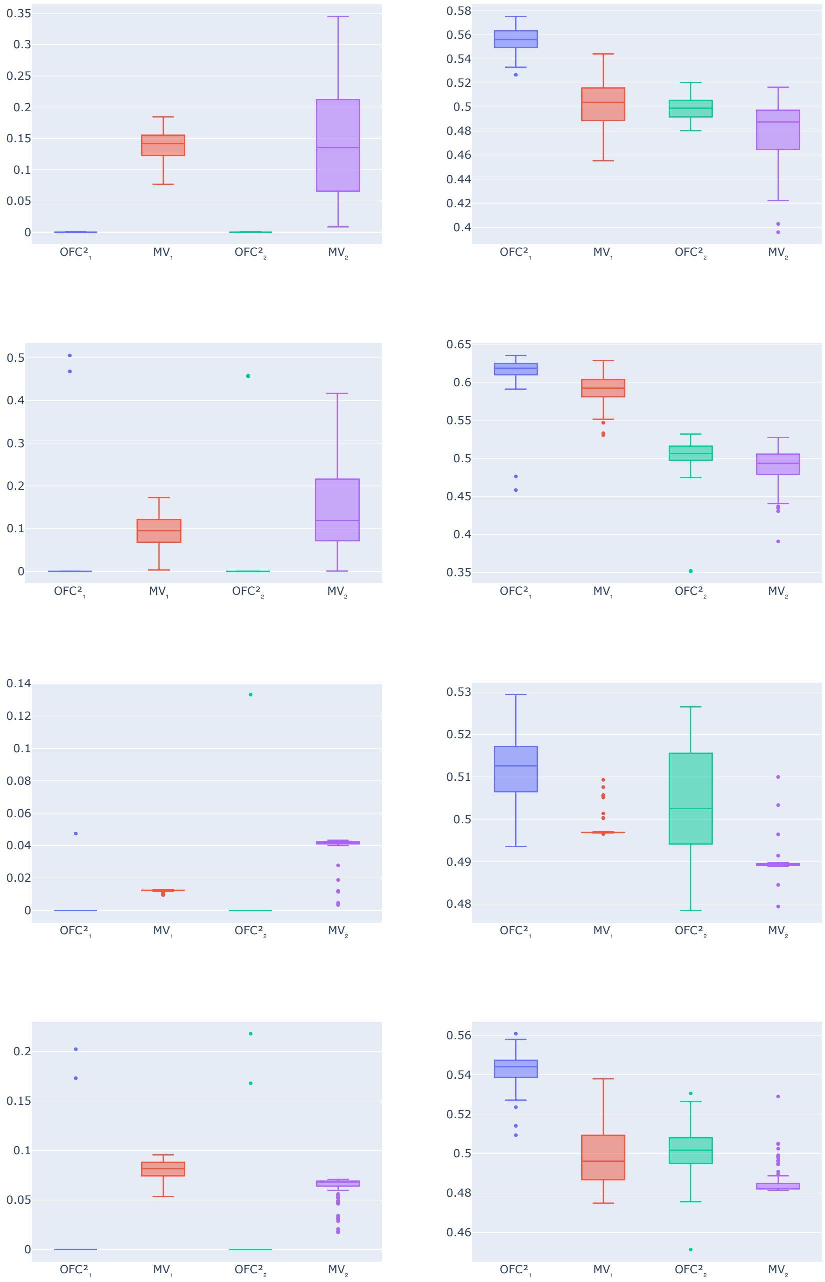


FIGURE 1. Comparison of CER (left) and F1 (right); see (1) and (2). The First row: KHlip12A, second row: KHhyp12, third row: KHsa12, last row: PKdep12.

with the F1 score, as we do for the validation numerical study described in Section 4. Hence, we only present the distributions over the classes for the GEDA and for the Gesyland. We focus on displaying the results for the variables that used fairness in the classification process, that is, the same

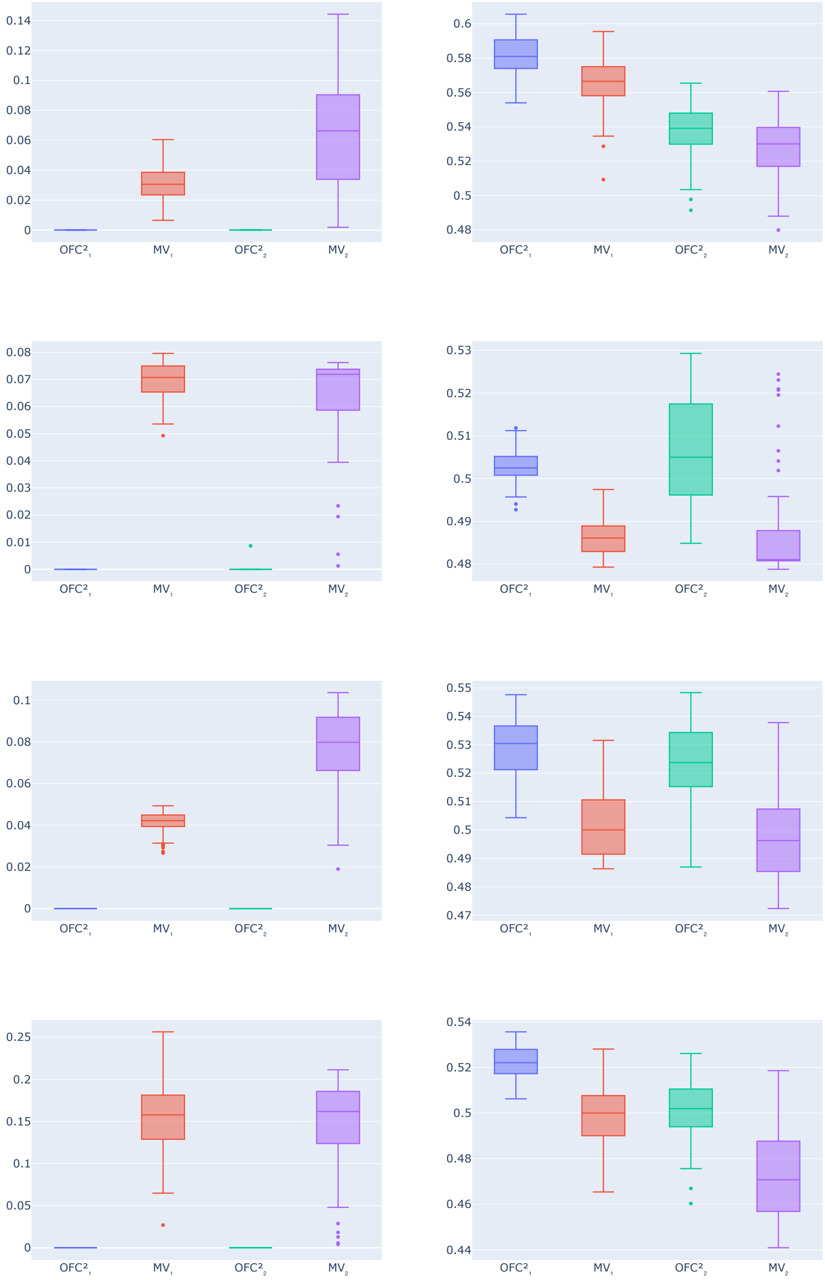


Figure 1. (continued) First row: KHdiabB12, second row: KHab12, third row: KHcb12B, last row: KHdge12.

as presented in Section 4. All other generated variables can be found in Appendix A.

Table 3 displays the percentage of points labeled 1 (with the remainder labeled 2) in the GEDA and Gesyland datasets for the variables already mentioned. In the Gesyland, these variables were generated using OFCA. The

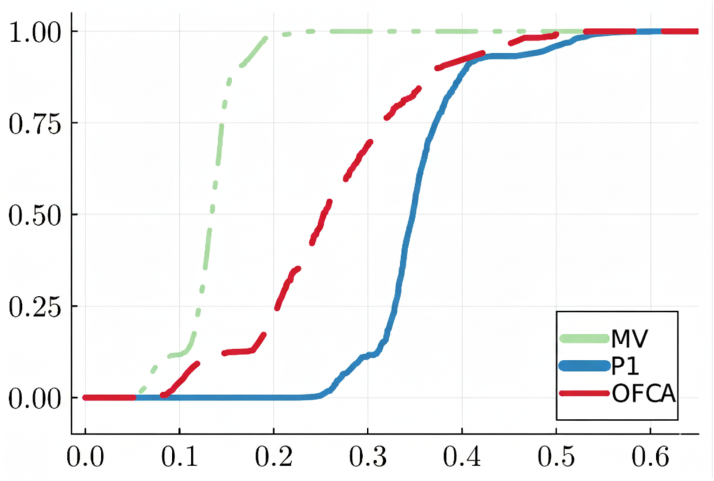


FIGURE 2. ECDFs for run times (in seconds).

Variable	Sensitive Group $\mathcal{S}_1$		Sensitive Group $\mathcal{S}_2$	
	GEDA	Gesyland	GEDA	Gesyland
KHlip12A	6.93	6.91	20.07	20.08
KHhyp12	22.59	22.60	55.70	55.70
KHsa12	1.21	1.12	4.26	3.96
PKdep12	12.16	12.15	7.89	7.93
KHdiabB12	6.93	6.91	20.07	20.08
KHab12	8.21	8.20	7.42	7.33
KHcb12B	5.42	5.42	10.85	10.71
KHdge12	13.45	13.49	40.00	39.92

TABLE 3. Percentage of points with labaled 1.

observed similarity in these percentages suggests that incorporating distribution as an optimization constraint positively impacts population similarity.

Nevertheless, a slight difference in the distributions of the variables can still be observed in Table 3. In the next section, we will further explore these discrepancies.

Conditional Distribution. We now examine the conditional distributions of variables $KHdiabB12$ and $KHhyp12$ in terms ofage and gender; information for other variables can be found in Appendix B.

As detailed in Appendix A, the $age5B$ variable ranges from 1 to 16, representingage in 5-year intervals. For instance, an $age5B = 1$ indicates anage between 0 and 5 years, and so on.

A potential concern was that enforcing cardinality constraints might significantly degrade these conditional distributions; however, our observations indicate that this degradation does not occur. As observed in Figure 3, the

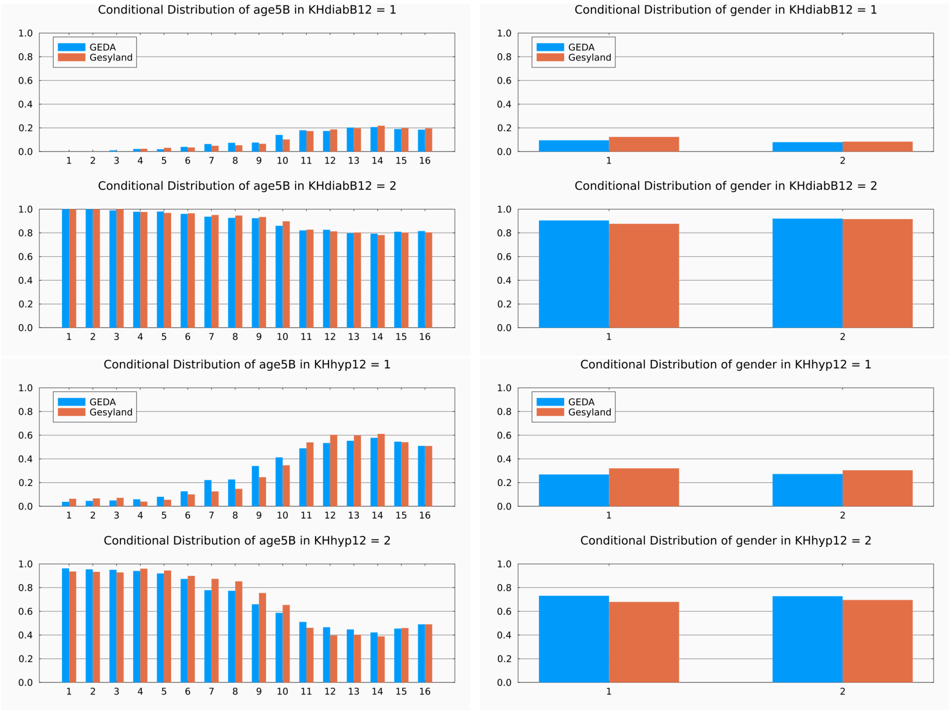


Figure 3. Conditional distribution of variables in respect of age and gender.

conditional distributions, while exhibiting minor differences, remain robust and logical.

These minor differences arise from the fact that our initial variables, derived from the Gesyland dataset, are generated based on the Mikrozensus dataset from Herwig and Schimpl-Neimanns (2013). Consequently, the populations differ in terms ofage, gender, and educational level (ISCED). We reiterate that this is not a limitation but rather a consequence of employing distinct datasets in the creation of a synthetic population.

6. Conclusion

In many research fields, it is of importance to have a close to reality population file. Two major fields are agent based simulation in epidemiology and traffic simulations. Depending on the field of application, some subgroups of the population can be impactful for the outcome of the study. These sensitive parts of the population have to be generated with extra care.

To address this, we present a fair and calibrated classifier based on a MILP model that incorporates these distributions as constraints to ensure fairness across sensitive groups.

We also proved the unimodularity of the proposed model, allowing it to be solved as a linear programming (LP) problem. To further reduce the computational cost, we introduce a preprocessing technique that fixes the label of points with evident class assignments before solving the optimization model.

Our numerical results demonstrate that the proposed model achieves a higher F1 score and more accurately reflects the correct class distribution compared to a random forest with majority voting.

Furthermore, considering the German Health Update (GEDA) as a demographic survey, we used the proposed algorithm to generate health variables for a georeferenced synthetic German population of nearly 85 million individuals. The numerical results show that the created variables exhibit a distribution similar to GEDA.

An interesting future research question would be how to extend the proposed method to allow for time series data. E.g. creating variables, that show the disease progression accounting for household cluster effects over time.

Acknowledgements

The authors thank the German Federal Ministry of Education and Research (BMBF) for their support of this research project and the "OptimAgent Project".

The authors also thank the DFG for their support within RTG 2126, "Algorithmic Optimization".

References

Attri, I., L. Awasthi, and T. Sharma (July 2023). "Machine learning in agriculture: a review of crop management applications." In: *Multimedia Tools and Applications* 83, pp. 1–41. DOI: 10.1007/s11042-023-16105-2.

Besse, P., E. del Barrio, P. Gordaliza, J.-M. Loubes, and L. Risser (Apr. 2022). "A Survey of Bias in Machine Learning Through the Prism of Statistical Parity." In: *The American Statistician* 76.2, pp. 188–198. DOI: 10.1080/00031305.2021.195.

Bezanson, J, A Edelman, S Karpinski, and V. B. Shah (2017). "Julia: A fresh approach to numerical computing." In: *SIAM review* 59.1, pp. 65–98. DOI: 10.1137/14100067.

Burgard, J. P. and J. V. Pamplona (2024a). *Fair Generalized Linear Mixed Models*. arXiv: 2405.09273 [cs.LG].

– (2024b). *FairML: A Julia Package for Fair Classification*. arXiv: 2412.01585 [cs.LG]. URL: https://arxiv.org/abs/2412.01585.

Burgard, J. P., J. V. Pamplona, and S. Shams (2025a). *Gesyland: A Georeferenced Synthetic Population for Germany*. URL: https://gesyland.eu/en/.

Burgard, J. P., M. E. Pinheiro, and M. Schmidt (2025b). "Mixed-Integer Linear Optimization for Cardinality-Constrained Random Forests." In: *Optimization Letters*. DOI: 10.1007/s11590-025-02191-8.

Carrizosa, E., K. Kurishchenko, and D. Romero Morales (2025). "On enhancing the explainability and fairness of tree ensembles." In: *European Journal of Operational Research*. DOI: https://doi.org/10.1016/j.ejor.2025.01.008.

Caton, S. and C. Haas (Apr. 2024). "Fairness in Machine Learning: A Survey." In: *ACM Comput. Surv.* 56.7. DOI: 10.1145/3616865.

Chicco, D. and G. Jurman (2020). "The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation." In: *BMC genomics* 21, pp. 1–13.

Figueira, A. and B. Vaz (2022). "Survey on Synthetic Data Generation, Evaluation Methods and GANs." In: *Mathematics* 10.15. DOI: 10.3390/math10152733.

Heller, I. and C. B. Tompkins (1957). "14. An Extension of a Theorem of Dantzig's." In: *Linear Inequalities and Related Systems*. Princeton: Princeton University Press, pp. 247–254. DOI: doi:10.1515/9781400881987-015.

Herwig, A. and B. Schimpl-Neimanns (2013). *Mikrozensus Scientific Use File 2010: Dokumentation und Datenaufbereitung*. Vol. 2013/10. GESIS-Technical Reports. Mannheim: GESIS - Leibniz-Institut für Sozialwissenschaften, p. 25.

Huangfu, Q. and J. J. Hall (2018). "Parallelizing the dual revised simplex method." In: *Mathematical Programming Computation* 10.1, pp. 119–142.

Joshi, A., B. Raman, K. M. Chalavadi, and L. R. Cenkeramaddi (Oct. 2023). "Application of a new machine learning model to improve earthquake ground motion predictions." In: *Natural Hazards* 120. DOI: 10.1007/s11069-023-06230-4.

Lu, Y., M. Shen, H. Wang, X. Wang, C. van Rechem, T. Fu, and W. Wei (2024). *Machine Learning for Synthetic Data Generation: A Review*. arXiv: 2302.04062 [cs.LG]. URL: https://arxiv.org/abs/2302.04062.

Lubin, M, O Dowson, J. D. Garcia, J Huchette, B Legat, and J. P. Vielma (2023). "JuMP 1.0: Recent improvements to a modeling language for mathematical optimization." In: *Mathematical Programming Computation*. DOI: 10.1007/s12532-023-00239-3.

Miron, M., S. Tolan, E. Gómez, and C. Castillo (2020). *Addressing multiple metrics of group fairness in data-driven decision making.* arXiv: 2003.04794 [cs.LG]. URL: https://arxiv.org/abs/2003.04794.

Murtaza, H., M. Ahmed, N. F. Khan, G. Murtaza, S. Zafar, and A. Bano (2023). "Synthetic data generation: State of the art in health care domain." In: *Computer Science Review* 48, p. 100546. DOI: https://doi.org/10.1016/j.cosrev.2023.100546.

Ponge, J., D. Horstkemper, B. Hellingrath, L. Bayer, W. Bock, and A. Karch (2023). "Evaluating parallelization strategies for large-scale individual-based infectious disease simulations." In: *2023 Winter Simulation Conference (WSC).* IEEE, pp. 1088–1099.

Robert Koch Institute (2025). *GEDA German Health Update.* URL: https://www.rki.de/EN/Topics/Noncommunicable-diseases/Health-surveys/Studies/geda-german-health-update.html (visited on 05/05/2025).

Sadeghi, B., P. Chiarawongse, K. Squire, D. C. Jones, A. Noack, C. St-Jean, R. Huijzer, R. Schätzle, I. Butterworth, Y.-F. Peng, and A. Blaom (Nov. 2022). *DecisionTree.jl - A Julia implementation of the CART Decision Tree and Random Forest algorithms.* Version 0.11.3. DOI: 10.5281/zenodo.7359268. URL: https://doi.org/10.5281/zenodo.7359268.

Schrijver, A. (1986). *Theory of linear and integer programming.* USA: John Wiley & Sons, Inc.

Sun, S., Z. Cao, H. Zhu, and J. Zhao (2020). "A survey of optimization methods from a machine learning perspective." In: *IEEE Transactions on Cybernetics* 50.8, pp. 3668–3681. DOI: 10.1109/TCYB.2019.2950779.

Zafar, M. B., I Valera, M Gomez-Rodriguez, and K. P. Gummadi (2017). "Fairness Constraints: Mechanisms for Fair Classification." In: *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics.* Ed. by A. Singh and J. Zhu. Vol. 54. Proceedings of Machine Learning Research. PMLR, pp. 962–970.

Zhang, Y., B. Liu, Y. Gong, J. Huang, J. Xu, and W. Wan (2024). "Application of Machine Learning Optimization in Cloud Computing Resource Scheduling and Management." In: CIBDA '24. Wuhan, China: Association for Computing Machinery, 171–175. DOI: 10.1145/3671151.3671183. URL: https://doi.org/10.1145/3671151.3671183.

Appendix A. Variables Description

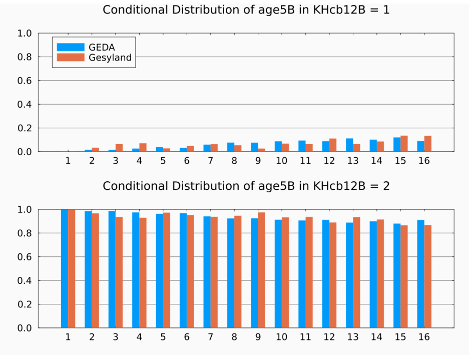
This section details all variables obtained in this study for the entire German synthetic population (approximately 85 million individuals), beyond those already presented within the paper's scope. These variables are listed in Table 4 and can be found at Gesyland.eu.

Table 4. Synthetic population health variables.

Variable	Description
age5B	Age of the individual (5-year intervals)
gender	Gender of the individual
MIcitizen	Nationality of the individual
SDisced11z	Educational level (ISCED) of the individual
KAspodauC_k	Total weekly sports in minutes
KAgfka	Adherence to WHO aerobic recommendations
KAgfmk	Adherence to WHO muscle strengthening recommendations
PAbmiB_k2	BMI group detail
KHbbmyo12	Myocardial infarction/chronic complaints: Within the last 12 months
BBmyo12	Chronic complaints after heart attack: Within the last 12 months
KHkhk12	Coronary Heart Disease/Angina Pectoris: Within the last 12 months
AKechi47	Harmful alcohol consumption (ECHI 47; $>$ 20/40 g/day)
KHlz12	Cirrhosis: Within the last 12 months
RCstatE	Smoker
KHalgi112	Allergies: Within the last 12 months

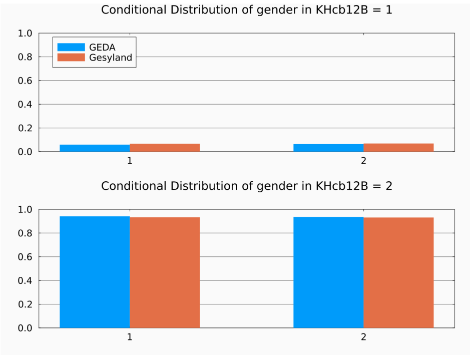
Appendix B. Conditional distribution os variables

In this section we provide the conditional distributions of the remaining variables. Further details and explanations for each of these variables are provided in Appendix A.



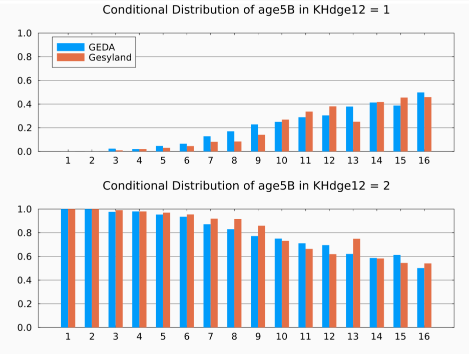
Conditional Distribution of age5B in KHcb12B = 1

Conditional Distribution of age5B in KHcb12B = 2



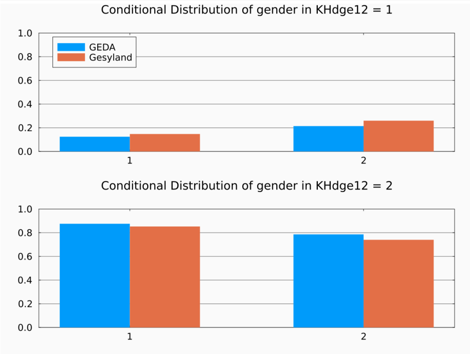
Conditional Distribution of gender in KHcb12B = 1

Conditional Distribution of gender in KHcb12B = 2



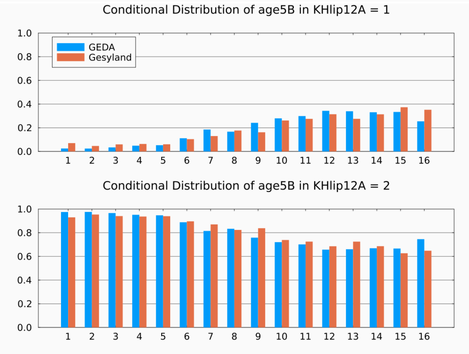
Conditional Distribution of age5B in KHdge12 = 1

Conditional Distribution of age5B in KHdge12 = 2



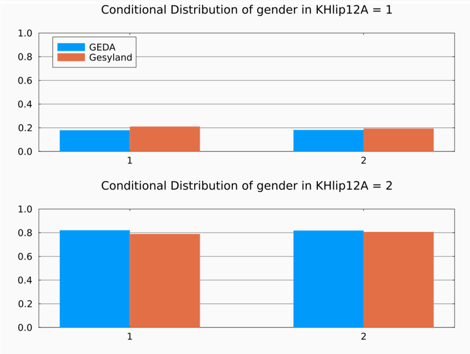
Conditional Distribution of gender in KHdge12 = 1

Conditional Distribution of gender in KHdge12 = 2



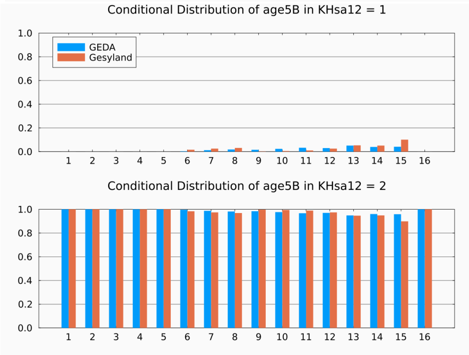
Conditional Distribution of age5B in KHlip12A = 1

Conditional Distribution of age5B in KHlip12A = 2



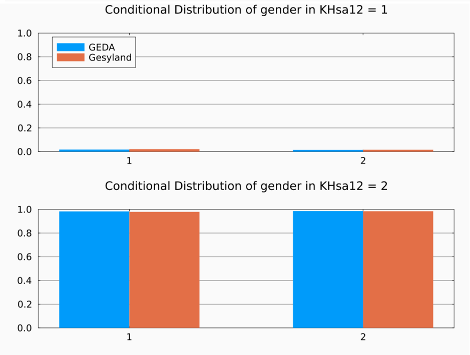
Conditional Distribution of gender in KHlip12A = 1

Conditional Distribution of gender in KHlip12A = 2



Conditional Distribution of age5B in KHsa12 = 1

Conditional Distribution of age5B in KHsa12 = 2



Conditional Distribution of gender in KHsa12 = 1

Conditional Distribution of gender in KHsa12 = 2

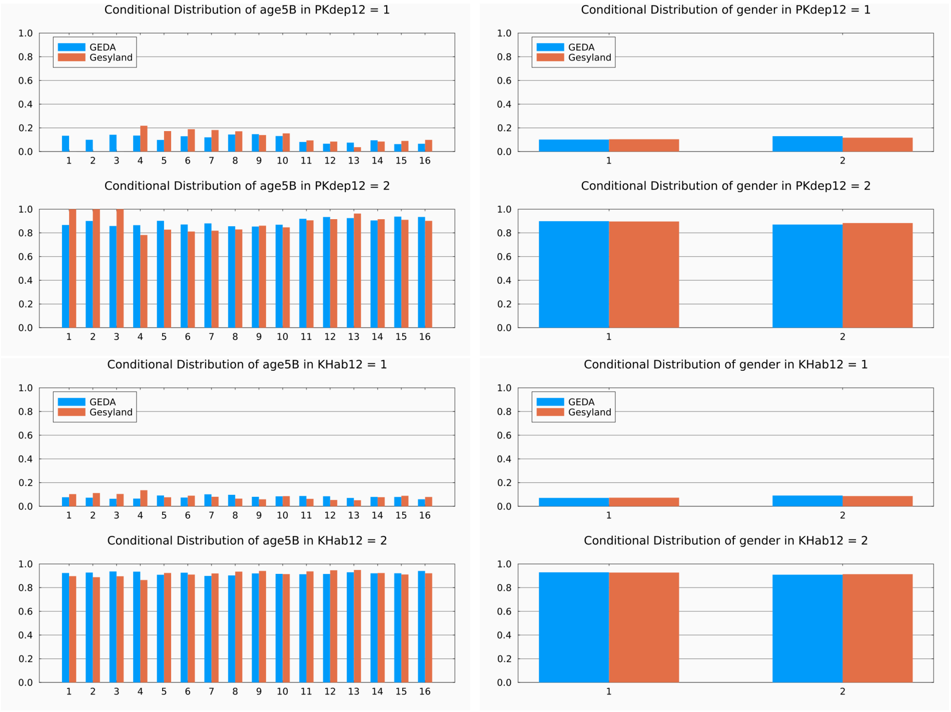


Figure 4. Conditional distribution of variables in respect of age and gender.

(J. P. Burgard, J.V.Pamplona) Trier University, Department of Economic and Social Statistics, Universitätsring 15, 54296 Trier, Germany

Email address: burgardj@uni-trier.de

Email address: pamplona@uni-trier.de

(M. E. Pinheiro) Trier University, Department of Mathematics, Universitätsring 15, 54296 Trier, Germany

Email address: pinheiro@uni-trier.de