Frimpong, Rejoice; Richiardi, Matteo

Working Paper

Machine learning regionalisation of input data for microsimulation models: An application of a hybrid GBM/IPF method to build a tax-benefit model for the Essex region in the UK

CeMPA Working Paper Series, No. CeMPA WP 9/25

Provided in Cooperation with:
University of Essex, Centre for Microsimulation and Policy Analysis (CeMPA)

Suggested Citation: Frimpong, Rejoice; Richiardi, Matteo (2025) : Machine learning regionalisation of input data for microsimulation models: An application of a hybrid GBM/IPF method to build a tax-benefit model for the Essex region in the UK, CeMPA Working Paper Series, No. CeMPA WP 9/25, University of Essex, Centre for Microsimulation and Policy Analysis (CeMPA), Colchester

This Version is available at:
https://hdl.handle.net/10419/339424

CeMPA Working Paper Series

CeMPA WP 9/25

Machine learning regionalisation of input data for microsimulation models: An application of a hybrid GBM / IPF method to build a tax-benefit model for the Essex region in the UK.

Rejoice Frimpong

Matteo Richiardi

August 2025

CeMPA
Centre for Microsimulation and Policy Analysis

Machine learning regionalisation of input data for microsimulation models: An application of a hybrid GBM / IPF method to build a tax-benefit model for the Essex region in the UK.

Rejoice Frimpong[1,2] and Matteo Richiardi[2]

[1] Essex County Council

[2] Centre for Microsimulation and Policy Analysis (CeMPA), University of Essex

Abstract

Development of microsimulation models often requires reweighting some input dataset to reflect the characteristics of a different population of interest. In this paper we explore a machine learning approach whereas a variant of decision trees (Gradient Boosted Machine) is used to replicate the joint distribution of target variables observed in a large commercially available but slightly biased dataset, with an additional raking step to remove the bias and ensure consistency of relevant marginal distributions with official statistics. The method is applied to build a regional variant of UKMOD, an open-source static tax-benefit model for the UK belonging to the EUROMOD family, with an application to the Greater Essex region in the UK.

1. Introduction

This paper describes a novel approach to reweighting the input population of a microsimulation model. The specific application is a regional variant of UKMOD, an open-source tax-benefit model for the UK belonging to the EUROMOD family (Richiardi et al., 2021), but the approach is of more general interest.[1] UKMOD is based on representative input data for the UK at a macro-regional level (the 12 government office

[1] For instance, the methodology is being applied by the research team to the construction of a regional input population for a dynamic microsimulation model.

regions). We aim at building a regional variant of UKMOD for the Greater Essex region, to explore the distributional impact of the national tax-benefit policies on the Essex population.[2] The problem is similar to that of Panori et al. (2017), who developed a tax-benefit model for Athens, and Boscolo et al. (2025), who developed a NUTS-3 version of the EUROMOD tax-benefit model for Italy. However, these studies could only exploit information from marginal distributions to calibrate the model to the sub-regional level – from the census in the case of Panori et al., and from both the census and tax data in the case of Boscolo et al. – as the data providers typically limit the number of disaggregating variables for which statistics are made available, on any variable of interest.

As it is customary for aligning multivariate distributions (the original survey data, representative at the national or macro-regional level) to target marginal distributions (census and tax information at the sub-regional level), these studies used Iterative Proportional Fitting (IPF), also known as raking. For our exercise, in addition to census information and other aggregate information at the sub-regional level coming from the national statistical office, we can exploit a rich commercially available regional household dataset. Our target is therefore a joint distribution of individual and household characteristics for the Greater Essex region, rather than a set of marginal distributions. This additional information obviously improves the quality of the reweighting exercise, while at the same time suggesting to go beyond IPF. Following Zhao et al. (2017), we apply propensity score matching using a variant of decision trees, namely Gradient Boosted Machine. Marginal distributions from the target dataset are however different in significant ways from those coming from official statistics. We therefore apply a final step of raking to obtain an input population representative of the Greater Essex region, to the best available evidence.

Our exercise involves reweighting the original input data rather than constructing a fully synthetic population. However, the machine learning methods employed here are also used for building synthetic populations, see for instance Zhou et al. (2022).

[2] The term 'Greater Essex' refers to the administrative region and includes the historical county of Essex plus the unitary authorities of Southend-on-Sea and Thurrock.

Reweighting approaches are also common in the literature on income inequality. DiNardo, Fortin and Lemieux (1996) introduced a method to reweight the sample distribution to create counterfactual scenarios under alternative covariate compositions. Later work, including Recentred Influence Function (RIF) regressions, builds on this approach by linking distributional statistics to covariates using parametric weights (Fortin, Lemieux and Firpo, 2010). Sologon, Doorley and O'Donoghue (2023) apply these techniques within a microsimulation-decomposition framework to study the drivers of income inequality across countries and over time. Their approach uses parametric models to simulate counterfactual income distributions, isolating the contribution of demographic, labour market, and policy-related factors. While their framework provides valuable insights into policy and structural effects, it relies on conventional reweighting techniques that may be sensitive to model specification. In contrast, the Gradient Boosted Machine is general in terms of the type of data it can handle (e.g. numerical, categorical), fully non-parametric and therefore flexible to adapt to whatever non-linear features of the data.

The rest of the paper is structured as follows. Section 2 describes the data and key challenges. Section 3 outlines the methodological approach, including gradient boosted machines and propensity score estimation techniques. Section 4 provides an overview of the implementation pipeline. Section 5 details the data preparation process and model training procedures. Section 6 evaluates matching quality through overlap analysis, covariate balance diagnostics, and effective sample size calculations. Section 7 presents macro-validation results comparing UKMOD-Essex outputs against administrative benchmarks for employment and self-employment income. Section 8 provides some final discussion points.

2. Data

UKMOD input data are derived from the Family Resources Survey (FRS) and are representative of the UK population at the level of the former government office regions (GORs). In our application we use the 2022 dataset, comprising observations for 53,577 individuals in 25,045 households for the whole of the UK, of which 5,158 individuals in

2,352 households for the East of England.[3] More detailed geographical information is available in the secure version of the FRS, which could be used to build new input data for UKMOD, but sample sizes at the sub-regional level become too small. In contrast, Experian is a commercially available household-level dataset with a much larger effective sample size and broader local coverage, capturing nearly all households at the postcode level. The 2023 wave of the Experian data contains information on 1,861,043 individuals in 738,993 households for the Greater Essex region, practically a 100% cover of the Essex population.[4] Differently from the FRS, Experian is not based on a probability sample but rather compiled from administrative records, commercial sources, and modelled estimates; its variables may be categorised or imputed differently, and household definitions may not be fully consistent with the UKMOD input data. Even more importantly, relevant information that is requested by the tax-benefit model and is available in the FRS is not included in the Experian data. This makes it very difficult to construct input data for UKMOD from the Experian data.[5] Our approach therefore consists in using the standard (FRS-based) input data for UKMOD and reweight it to mimic the target joint distribution of the common variables between FRS and Experian. This ensures the dataset remains comprehensive and internally consistent, retaining all the variables that are required by the tax-benefit model.

In our application, we refer to UKMOD input data as 'control', and to Experian data for Greater Essex as 'treatment', The terminology originates from causal inference, a framework commonly used to compare outcomes between different groups. The treated group is to be understood as the reference or target population, and the control group is adjusted to resemble it. This framing allows for a straightforward application of established techniques such as propensity score estimation and inverse probability weighting.

[3] We also reweighted the 2023 UKMOD input dataset, with similar performance. Both datasets are publicly available, see Section 8.
[4] According to data from the Office for National Statistics (ONS), the population of Greater Essex in March 2023 comprised 1,841,192 individuals in 771,189 households.
[5] Using the Experian data as input for UKMOD would also have consequences in terms of how the data can be shared with users of the model, as the data is not publicly available.

Propensity scores quantify the likelihood that a household belongs to the treated group, conditional on its observed characteristics. These scores guide the reweighting process. Observations that more closely resemble the target group receive larger weights, while those less similar contribute less to the final estimates.

3. Methods

We use a reweighting strategy involving two-steps: estimating the likelihood of each household appearing in the target dataset (propensity score estimation) and adjusting the data using Inverse Probability Weighting (IPW). A Gradient Boosted Machine (GBM) model is used to estimate propensity scores, the probability that a given household would appear in the treated dataset, conditional on observed characteristics. The GBM model incorporates a wide range of covariates, including age, tenure, household size, employment status, equivalised income, and the presence of children, along with interaction terms.

Traditional models, such as logistic regression, assume linear, additive relationships between variables, an assumption that often overlooks important interactions or nonlinear patterns. Tree-based models such as Gradient Boosted Machines and Random Forests are well-suited to address this complexity. They capture higher-order interactions and nonlinearities directly from the data, producing more reliable propensity scores that reflect realistic patterns in household-level characteristics. In the following we briefly describe both methods, in order for the reader to gain an appreciation of the commonalities and differences, and the rationale behind our choice of GBMs.

Gradient Boosted Machines and Random Forests

Gradient Boosted Machines and Random Forests are nonparametric methods, meaning they do not assume a fixed functional form between covariates and treatment assignment, and are capable of capturing complex, nonlinear relationships and higher order interactions. Recent evidence shows that these models improve covariate balance and reduce bias compared to standard parametric approaches (Lee, Lessler and Stuart, 2010). GBM in particular has become a widely used method for propensity score

estimation in health and social science applications due to its strong performance in reducing imbalance (Leite et al., 2024).

Gradient Boosted models and Random Forests also provide a practical solution to the problem of missing data by incorporating surrogate splits, which enable the model to utilise incomplete cases without requiring imputation. This enables robust estimation without compromising the sample size.

Random Forest builds many decision trees on bootstrapped samples of the data and averages their predictions to improve accuracy and reduce variance (Breiman, 2001). It is relatively robust to overfitting, handles missing data through surrogate splits, and typically requires minimal tuning. GBM by contrast, builds trees sequentially, each one improving on the residuals of the previous, which makes it more sensitive to nuances in the data but also more prone to overfitting if not properly tuned (Friedman, 2001; Cortes, Mohri and Storcheus, 2019).

In this context, both models estimate the probability that an observation belongs to the treated group, which is then used to compute inverse probability weights. GBM is often preferred when the goal is to optimize prediction quality for weighting, because its sequential learning approach (boosting) is typically better at correcting systematic prediction errors than Random Forest's averaging approach (bagging).

However, Random Forest remains a strong alternative when interpretability, stability, or runtime is a concern. The choice between them ultimately depends on the balance between flexibility, interpretability, and computational cost required for the task at hand. Table 1 summarises the differences between the two methods.

Table 1: Comparison of Random Forest and Gradient Boosted Machines

Feature	**Random Forest**	**Gradient Boosted Machine (GBM)**	**Summary Comparison**
Tree Building Approach	Bagging (parallel tree building)	Boosting (sequential tree building)	Random Forest builds all trees independently and in parallel, so there's no learning from previous errors.

Feature	Random Forest	Gradient Boosted Machine (GBM)	Summary Comparison
			GBM learns from previous errors in a step-by-step manner, helping refine the match between UKMOD and Experian across iterations.
Learning Strategy	Independent trees	Trees built sequentially, each learning from the last	Random Forest can capture general patterns but is less effective at identifying complex multi-way interactions, particularly as the number of covariates increases. GBM Captures deeper, layered relationships, ideal for complex socioeconomic patterns
Performance Monitoring	Not available	Built in	Random Forest does not have built-in cross-validation or performance tracking during training. GBM provides detailed training diagnostics, including deviance plots and automatic stopping rules via cross-validation.
Scalability to Many Variables	Struggled as covariates increase	Handles many covariates well	GBM maintained covariate balance and matching performance with over 10 socioeconomic predictors.
Handling Interactions	Limited, often misses multi-way interactions	Excels at finding subtle 2-way, 3-way, 4-way interactions	GBM picked up nuanced patterns like “FT employment + no young children + large HH”
Final Matching Quality	Adjusted covariate balance worse than unadjusted (in places) when using more than 6 covariates	Adjusted balance consistently better than unadjusted	GBM consistently improved covariate balance across all variables, achieving lower Standardised Mean Differences (SMD) than Random Forest.

Propensity Score Estimation and Reweighting

Propensity score estimation relies on a small set of core assumptions that must hold for inverse probability weighting to produce valid and interpretable results.

Strong Ignorability. After conditioning on observed covariates, group assignment must be independent of potential outcomes. The propensity score model should include all key factors that influence both group assignment and outcomes. While this assumption is untestable, it can be supported through careful and comprehensive covariate selection.

Overlap or Common Support. There must be sufficient overlap in the distribution of propensity scores between the treated and control groups. If one group lacks observations in certain parts of the score range, weights can become unstable or extreme. This reduces the effective sample size and undermines comparability. Diagnostic checks and trimming are used to address these issues and ensure a reliable reweighting process.

Model Appropriateness. Nonparametric models such as GBM and Random Forest do not assume a fixed functional form, but they still depend on a well-specified set of covariates. The model must be flexible enough to capture nonlinear relationships and interactions, without introducing excess noise or instability. The objective is to generate accurate treatment probabilities that improve covariate balance, not necessarily to maximise predictive performance.

Once estimated, the propensity scores are used to derive inverse probability weights. To improve stability, these weights are stabilised and capped, preventing any single household from dominating the reweighted dataset. Additional trimming is applied to restrict the analysis to the region of common support, where treated and control observations overlap in their propensity scores. This reduces extrapolation and ensures comparability. Households with a low probability of being treated (i.e. low similarity to the benchmark population) receive down-weighted influence, while those more similar are up-weighted.[6] Differences in coverage between datasets also mean that some

[6] Adjusting weights to ensure that their total matches the population size is a standard practice in survey sampling. This technique, known as calibration estimation, involves scaling weights so that their sum aligns with known population totals, thereby enhancing the accuracy and representativeness of survey estimates (Henry and Valliant ;2015).

control units may not have close counterparts in the treated group. The weighting procedure accommodates this by assigning smaller weights to less similar units, and where no meaningful similarity exists, based on common support thresholds, those observations are downweighted to zero. This ensures that only observations with sufficient overlap contribute to final estimates, while preserving the full structure of the original dataset. Adjusted weights are scaled such that their total sum matches the total number of households in the target population, so that aggregate estimates based on the reweighted dataset remain meaningful and interpretable at the population level.

The reweighted dataset is then evaluated using several diagnostics. Standardised mean differences (SMD) is used to measure covariate balance before and after weighting, while effective sample size (ESS) is calculated to understand how much of the control sample meaningfully contributes to the weighted analysis. Finally, overlap is assessed to confirm that the reweighting process produced sufficient similarity between the treated and control groups across the full range of covariates.

A key consideration in propensity score estimation is the relative size of the treated and control groups. Significant disparities in sample sizes can introduce estimation challenges, reduce model efficiency, and affect the accuracy of resulting weights. Empirical evidence from King and Zeng (2001) shows that when the treated data is disproportionately large with respect to the control group (as in our case), under-sampling the treated group can improve estimation accuracy without sacrificing representativeness. Crump et al. (2009) similarly note that propensity score methods perform best when treated and control groups have sufficient overlap, and extreme imbalances can reduce this overlap, which in turn makes propensity score estimates less reliable. At the same time, retaining the full treated sample preserves its richness and descriptive power. Austin and Stuart (2015) caution against arbitrary sampling adjustments, emphasising that inverse probability weighting can be used to correct for imbalance while preserving full data integrity. In practice, the decision to under-sample or retain the full treated group involves trade-offs. Using a balanced subset may improve computational efficiency and statistical robustness, while a full treated sample ensures that rare subgroups are not excluded. In the application explored in this paper, where the

treated dataset contains over 700,000 households and the control group around 25,000, a balanced subsample is selected to match the employment-type distribution in UKMOD.

Rather than employing formal stratification across multiple demographic and socioeconomic variables, this study adopts a targeted employment-type matching approach for several methodological reasons. First, with a control sample of 25,000 UKMOD households, formal stratification across the full range of relevant covariates (age bands, tenure, household composition, employment status, and income) would create numerous thin cells with insufficient observations for reliable matching, particularly for rare household types. Second, the employment-focused sampling strategy directly addresses the primary policy-relevant heterogeneity between datasets while allowing the subsequent propensity score model to handle remaining covariate imbalances more flexibly through gradient boosting and interaction terms. This sequential approach, employment-based sampling followed by propensity score adjustment and population raking, achieves the distributional goals of stratification while maintaining methodological tractability and avoiding the over-stratification problems that would arise from attempting to balance all relevant dimensions simultaneously at the sampling stage. This approach preserves the representation of important subgroups while improving estimation performance in the propensity score model.

To explore different reweighting outcomes, we tested several sample configurations, including under-sampling the treated group, keeping original sizes, and using equal sample sizes from each dataset. Results were evaluated using covariate balance and effective sample size metrics. While full-sample models (with around 740,000 treated and 25,000 control) offered high statistical power, they performed poorly on balance, with only 6 out of 12 covariates falling within accepted thresholds, particularly tenure-related variables. Equal-size samples performed better across most characteristics, with all 12 covariates meeting standardised criteria. Although household size remained slightly imbalanced across all models, this configuration offered the best trade-off

between accuracy and robustness. Based on this finding, the equal-size approach was adopted for model estimation.

Visual diagnostics from the full sample support this decision. As shown in Appendix 1, the full sample showed poor common support between treated and control groups. Propensity score distributions remained misaligned even after trimming, matching, and weighting (Figures A1 to A4). The treated group dominates the upper end of the distribution, and matching did not fully resolve the imbalance. While IPW and raking improved overlap, differences in the underlying covariate structure remained. In comparison, the equal-size configuration (shown in section 4 and 5) produced better overlap and improved covariate balance. The covariate balance plot for the full sample (Figure A5) shows that several variables remained outside accepted thresholds, even after post-raking adjustments. These results confirm that the balanced subsample provides better common support and model performance.

After estimating propensity scores, weights are computed using IPW, and then stabilised, trimmed and calibrated through raking to match official marginal totals for age, employment and presence of children. After raking, all observations in the control dataset are retained, with final weights reflecting either their reweighted IPW-calibrated value or, where unmatched or trimmed, a default weight of 0.

4. Implementation Pipeline Overview

This section outlines the steps used to align the control dataset to the treated dataset using our machine learning-based reweighting approach. The process is designed to improve comparability between the two samples while retaining as much useful information as possible.

Data Preparation

Key variables are selected and standardised across datasets, including age, household size, number of children, tenure, and employment status. Disaggregated employment indicators and interaction terms (e.g. age × household size, retired × employed) are included to improve model flexibility. Equivalised income is residualised prior to

inclusion, to reduce its correlation with other covariates and preserve covariate balance during matching (see the next section).

Variable Selection and Feature Engineering for Balance

Variables are selected based on their relevance to treatment assignment and consistency across datasets. The core variables used in the model are:

- Age (categorised)
- Tenure type
- Household size
- Presence of children
- Number of children aged 0–4, 5–11, and 12–17
- Labour market activity (Student or Unemployed, Retired, Part time or Housewife, Employed Full Time)
- Equivalised income

Each variable is constructed to be consistent across sources. Age is grouped into standard brackets, tenure is recoded into harmonised categories, and labour market activity is converted into binary indicators. Variables are retained in their categorical or count form to preserve structure and avoid over-simplification.

To improve the model's ability to capture structural differences, a set of interaction terms is included:

- Age * Household size
- Age * Retirement status
- Household size * employment
- Children * Household size
- Has children * Housewife

These interactions help the model account for non-additive effects and improve balance in overlapping but distinct subgroups. All variables and interactions are used in both the matching and weighting steps.

Including equivalised income directly as a covariate disrupted covariate balance and reduces overlap between the treated and control groups as observed in earlier versions

of the model. This led to poor performance in key diagnostics, particularly for subgroups where income strongly correlated with other characteristics such as age and employment status. For example, the model overemphasised differences between retired households and working-age households simply due to income variation, even when other characteristics were similar. To address this, we switched to using residualised income. This involved regressing income on all other covariates in the model, such as age, household size, tenure, and employment status, and then including only the residuals from that regression in the propensity score model. This removed predictable, linear relationships between income and those variables, allowing income to contribute additional information without distorting the balance of the other covariates.

Propensity Score Estimation Using GBM

Using the prepared covariates, a Gradient Boosted Machine model is trained to estimate propensity scores. These represent the probability that a given household appears in the treated dataset, based on observed characteristics. The model incorporates interaction terms and is tuned to maximise predictive accuracy. After estimation, households with propensity scores falling outside the range shared by both groups are trimmed to ensure valid comparisons during reweighting.

Matching, Inverse Probability Weighting and Raking

Treated and control observations are matched using nearest-neighbour matching with a caliper restriction to ensure good matches. After matching, inverse probability weighting is used to assign weights to the control observations. These weights reflect how closely each control household resembles the treated sample and allow the reweighted control group to approximate the target distribution. To improve stability, the weights are stabilised and capped at the ninety-ninth percentile. Control households in the top one percent of the weight distribution are trimmed, and the remaining weights are renormalised. As a final step, the stabilised and trimmed weights are calibrated using raking to ensure alignment with official population totals. Raking adjusts the weights so that the distribution of key characteristics in the reweighted UKMOD data matches

known shares from the Greater Essex population. This includes age group, household employment status, the presence of children, and housing tenure. Raking improves representativeness and helps ensure that key household attributes are correctly reflected in the final weighted dataset.

Diagnostics and Validation

Covariate balance is assessed using Standardized Mean Differences (SMD) and Kolmogorov–Smirnov (KS) tests before and after weighting. Effective sample size (ESS) is used to assess how much of the control sample contributes meaningfully to the weighted analysis. Overlap is checked visually using density plots and propensity score distributions.

Output and Export

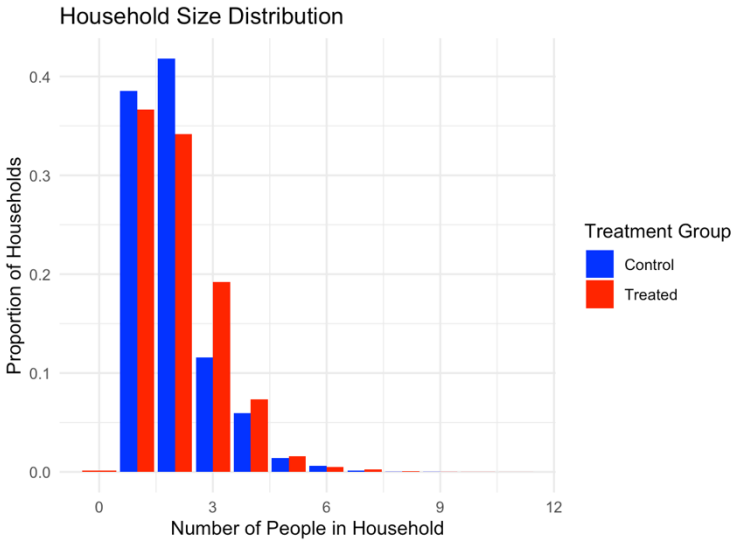
Final calibrated weights from the matched and reweighted sample are merged back into the original UKMOD dataset to enable consistent use in UKMOD simulations and further analysis. This ensures that the full household-level dataset remains intact, with updated weights applied to each observation based on their alignment with the treated sample. Observations that were trimmed or lacked common support are retained with a final weight of 0, allowing the complete dataset to be preserved while excluding non-aligned cases from influencing weighted estimates.

5. Data Preparation and Variable Selection

Covariate Comparison

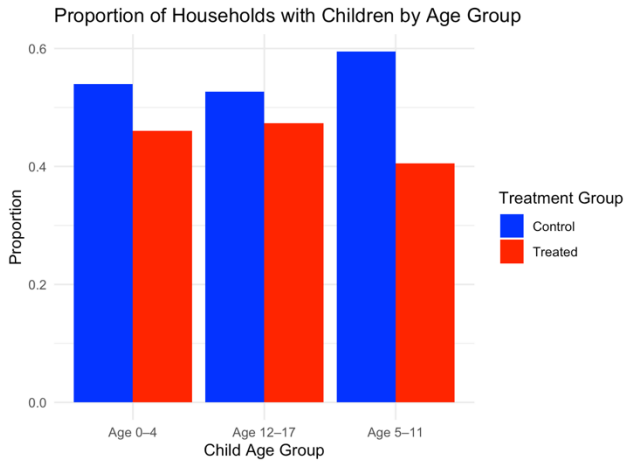
Before estimating propensity scores, it is important to examine how key household characteristics differ across the treated and control groups. The following plots compare the distributions of selected covariates using the full samples from each dataset. These comparisons highlight areas of overlap as well as imbalance.

Figure 1: Household size distribution, treated vs. controls



The distribution of household size shows notable structural differences between the two datasets (Figure 1). Smaller households (1–2 people) are more prevalent in the control group, whereas larger households (4 or more people) are more frequent in the treated group.

Figure 2: Household with children, treated vs. controls



Across all child age categories, a higher share of control households report having children compared to treated households. This difference is especially pronounced for children aged 5–11, where the gap is widest (Figure 2).

Figure 3: Household age distribution, treated vs. controls

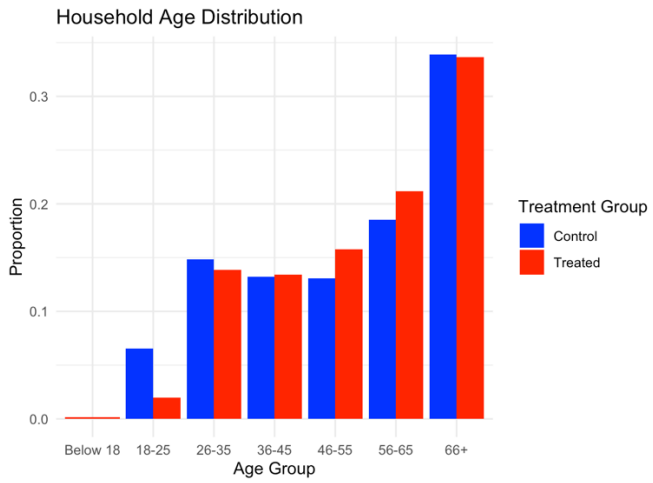
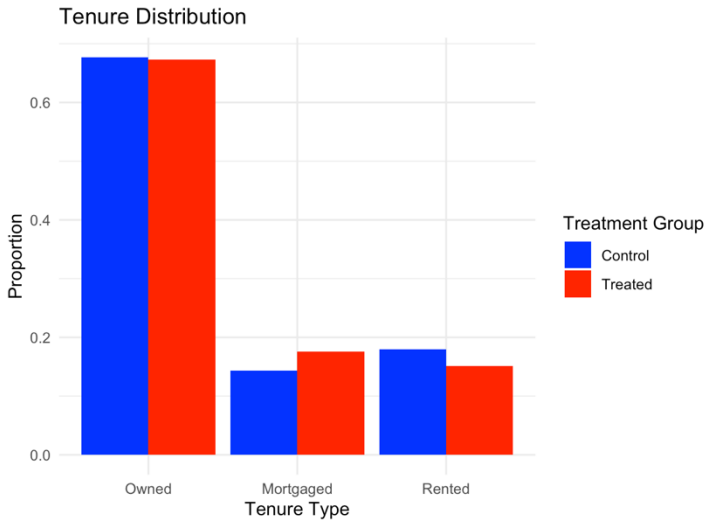


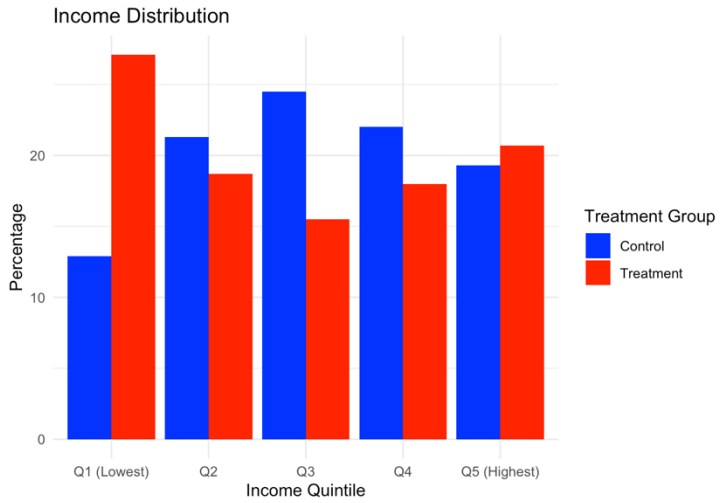
Figure 3 compares the age distribution in the two datasets. Both samples are concentrated in the older age brackets, but the control group contains a slightly larger share of households where the reference person is aged 66+. The treated group shows a more even distribution across working-age brackets (26–65), and (36-45).

Figure 4: Tenure distribution, treated vs. controls



Most households in both samples are owners (Figure 4), making up over two-thirds of each sample. The treated group has a slightly higher proportion of mortgaged households, while the control group shows a marginally larger share in the rental category. Overall, tenure patterns are broadly similar.

Figure 5: Income distribution, treated vs. controls



The income quintile distribution shows notable differences between the two datasets (Figure 5). The treated group (Experian) has a larger share of households in the lowest income quintile, while the control group (UKMOD) is more concentrated in the middle quintiles. These imbalances highlight the importance of including income in the reweighting process, while carefully handling its influence to preserve covariate balance.

Building the Dataset

Before estimating propensity scores and applying inverse probability weighting, the dataset is prepared to ensure consistent structure and formatting across all observations. This involves merging the treated and control groups into a single household-level dataset, with harmonised definitions for each covariate. Each row represents a household, and a binary treatment indicator identifies whether it belongs to the treated group (e.g., Experian, coded as 1) or the control group (e.g., UKMOD, coded as 0). Core covariates are cleaned, standardised, and aligned across sources.

- Household structure: household size, presence of children, number of children by age band
- Demographics: age group of household members, retirement status
- Economic activity: employment status, student/unemployed, part-time or full-time work, income

- Tenure: housing tenure (e.g., owned, mortgaged, rented)

Once the merged dataset is ready, a GBM model is trained to estimate the probability that each household belongs to the treated sample, based on observed covariates.

After training, each household is assigned a propensity score, which is used to compute inverse probability weights. The distribution of scores is then reviewed to ensure there is sufficient overlap between treated and control observations, which is essential for stable and interpretable weighting.

6. Propensity Score Estimation Using GBM

Training A GBM Model for Propensity Score Estimation

Once the dataset is pre-processed and structured, gradient boosting is used to estimate the propensity scores that underpin the inverse probability weighting process. A GBM classifier is trained using covariates that capture demographic structure, tenure, household composition, labour market status, and income. Interaction terms are also included to account for non-additive relationships, such as how employment status might interact with household size or retirement.

The model outputs a probability for each household representing the likelihood of being in the treated group, based on its observed characteristics. These probabilities form the basis for weighting control units to resemble the treated sample. Cross-validation is used to determine the optimal number of trees, and shrinkage and tree depth are tuned to maximise generalisation while controlling overfitting. GBM offers flexibility in capturing complex nonlinear patterns without imposing rigid parametric assumptions.

Visualising Decision Trees

Interpretation of decision trees is often more intuitive in Random Forests, where each tree directly outputs class probabilities and predicted labels at each node. These models are widely used in applied research and offer clear, interpretable structures: nodes represent splits on features (e.g., *Has_Children = 1*), and leaf nodes contain predicted outcomes and class proportions. A leaf node is simply the endpoint of a path through the

tree, a segment of the dataset where no further splits are made. In Random Forests, this leaf outputs the final prediction for households that fall into that segment.

Although our focus is on Gradient Boosted Machines (GBM), which build predictions sequentially rather than independently, we draw on Random Forests to help explain how GBM trees work. The differences between them are instructive, and understanding Random Forests first helps clarify GBM outputs.

While tree-based models such as GBM and Random Forest are primarily used here as flexible estimators of treatment assignment, inspecting individual trees provides valuable insights into the model behaviour and feature importance. Each decision tree maps out a sequence of binary splits based on household characteristics, offering a transparent view of how the model segments the data to distinguish between treated and control units.

Understanding Tree Components and Values

Decision trees from GBM and Random Forest models display information differently, reflecting their distinct approaches to prediction:

GBM Tree Components (Figure 6): Each node contains:

- Split Condition: The binary rule dividing households
- Cover: The number of households reaching this node during training
- Gain: How much this split improves predictive accuracy (higher values indicate more informative splits)
- Value: Numeric output at leaf nodes representing incremental adjustments

Random Forest Tree Components (Figure 7): Each node displays three lines of information:

- Top line: The predicted class at that node (0 = control/UKMOD, 1 = treated/Experian)
- Middle line: The class probability - proportion of treated households among all reaching this node (e.g., 0.29 = 29% treated)

- Bottom line: Node size - percentage of total training sample reaching this node (100% at root)

Root Node: Each decision tree starts with a root node containing the full dataset. In Random Forest, this shows 100% sample coverage with the overall class distribution. GBM root nodes are guided by residuals from previous trees, while Random Forests select splits based on purity measures. The split condition identifies the most discriminating feature, such as household size or employment status.

Internal Nodes: As trees grow, internal nodes partition data using conditions like "Has_Children = 1" or household size thresholds. In Random Forest trees, the class probability at each node shows the shifting balance between treated and control households as you move down branches. The node size indicates what proportion of the original sample reaches each decision point.

Leaf Nodes and Value Interpretation

The leaf nodes represent endpoints where final predictions are made, with fundamentally different interpretations:

GBM: Leaf nodes output small incremental adjustments (e.g., -0.030, 0.019) that contribute toward the final log-odds of treatment assignment. These values are summed across hundreds of trees and transformed through a logistic function to produce the final probability. Positive values increase treatment likelihood; negative values decrease it.

Random Forest: Leaf nodes show the final predicted class (control or treated) along with the exact probability estimate. A leaf showing 0.8 means 80% of households reaching this endpoint are treated. These probabilities are averaged across all trees in the forest to produce the final propensity score. The bottom number indicates what fraction of the total sample ends up in this particular leaf.

Comparing the Approaches

Figure 6 shows a GBM tree where Cover values track household flow and Gain measures split informativeness, while leaf values represent small incremental contributions that

accumulate across the ensemble. Figure 7 displays a Random Forest tree where node probabilities directly reflect the class distribution of households reaching each point, with color-coded predictions (orange = control, green = treated) and explicit sample proportions.

The fundamental distinction lies in how predictions are made:

- GBM builds knowledge incrementally through sequential trees correcting previous errors.
- Random Forest aggregates predictions across many independent trees, each trained on a random subset of the data.

Because Random Forest nodes report direct class probabilities, they often appear more intuitive. In contrast, GBM trees operate through accumulated small contributions, which can be harder to interpret at the node level but often yield more accurate results overall. In this study, GBM was chosen for its superior balance performance, but insights from Random Forests help illustrate how tree-based methods divide and classify households during propensity score estimation.

Figure 6: A single tree in the GBM

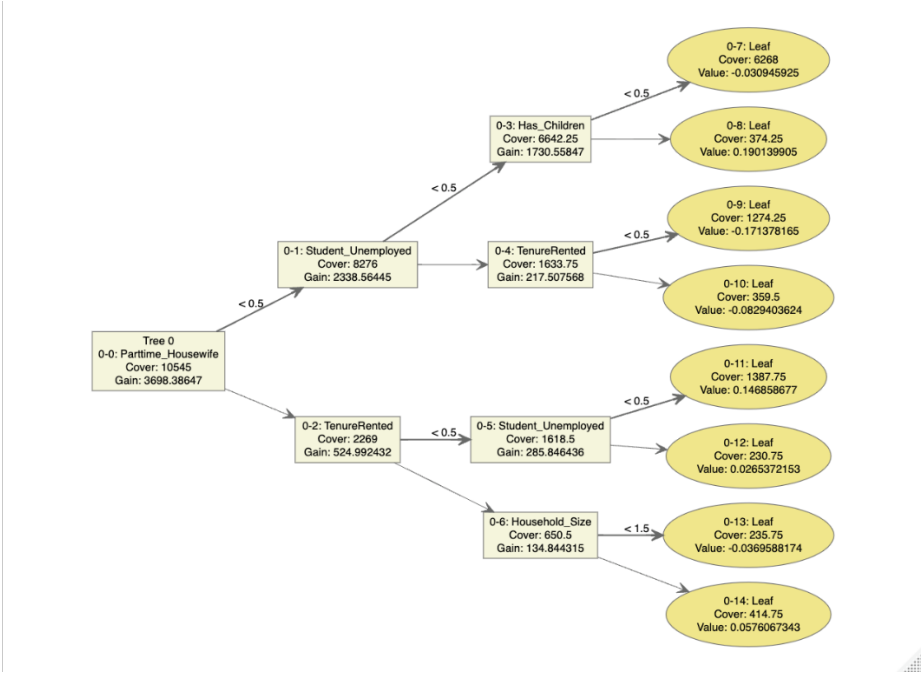
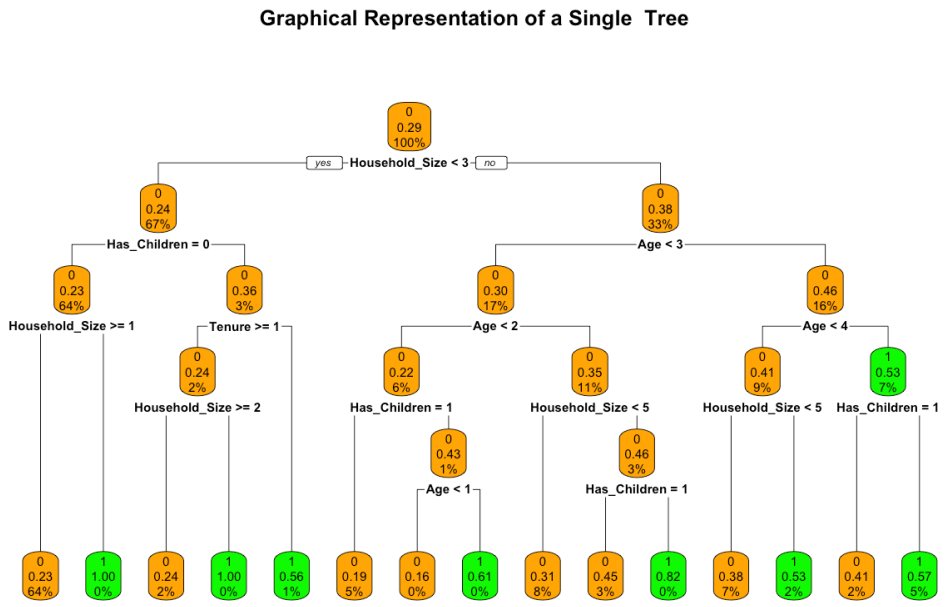


Figure 7: A single tree in the Random Forest



7. Matching, Weighting and Diagnostics

After implementing the matching, inverse probability weighting and raking steps described in section 4, it is essential to assess whether the reweighting process has improved comparability between the treated and control groups. The goal is to ensure that any observed differences reflect meaningful variation rather than underlying sample composition. Without proper diagnostics, residual bias may persist, even after careful modelling.

This section introduces three key diagnostics used to evaluate matching and reweighting quality:

- Overlap in propensity score distributions
- Covariate balance before and after weighting
- Weight Distribution before and After weighting
- Effective sample size (ESS)

Together, these tools provide a robust framework for judging whether the weighted control group offers a credible comparison to the treated sample.

Overlap Analysis: Propensity Score Distribution

Figure 8: Propensity Score Distribution Before matching

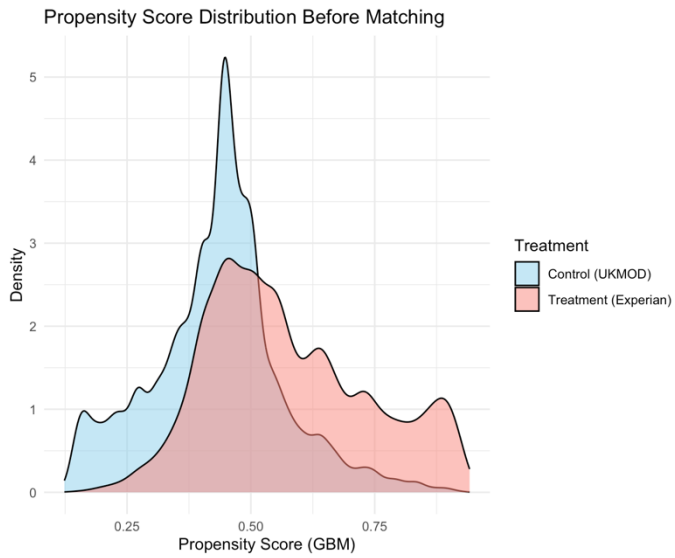
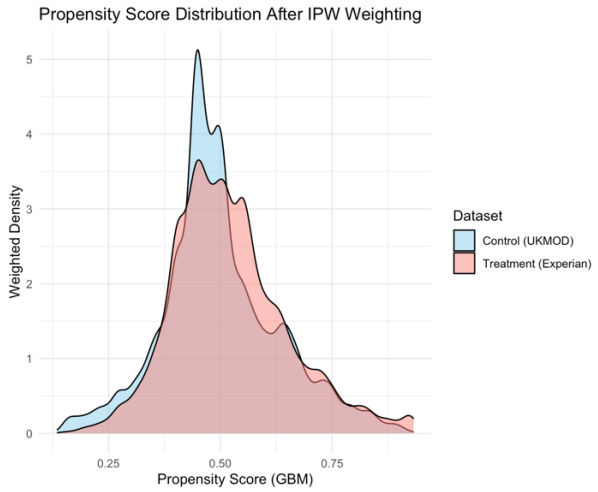


Figure 8 shows the estimated propensity score distributions for the control group (UKMOD, in blue) and the treated group (Experian, in pink) before any adjustments are applied. For IPW to work well, there needs to be sufficient overlap between these distributions, meaning control households must span a similar range of characteristics as those in the treated group. In this case, both curves align reasonably well through the central range, but the control group is more sharply peaked, while the treated group is more dispersed, particularly in the upper tail. These edge cases can reduce the effectiveness of reweighting and may require trimming or calibration in later steps to improve overlap and ensure stable weights.

Figure 9: Propensity Score Distribution After Matching and Weighting



After matching and applying inverse probability weighting, the distribution of propensity scores between the treated and control groups shows markedly improved alignment as shown in figure 9. The shaded areas now overlap more closely, especially across the central range of scores, indicating that the reweighted control group better mirrors the treated population. This improvement in overlap suggests that the weighting adjustments have effectively addressed initial disparities, reducing bias and increasing the validity of downstream comparisons. While some differences remain at the extremes, the overall fit confirms that the sample is now suitably aligned for further analysis.

Post-IPW Calibration Using Raking

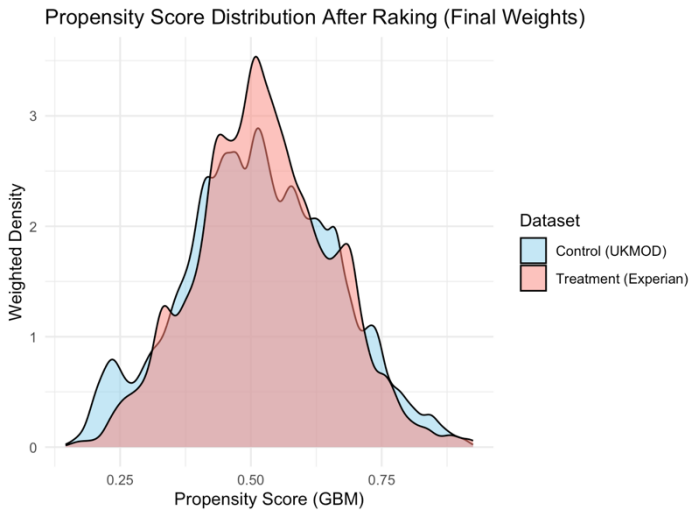
While inverse probability weighting substantially improves alignment between the control and treated groups by adjusting for observed covariates, it does not guarantee that weighted totals match known population margins. In practice, even after IPW, residual imbalances can remain, particularly for characteristics that are only weakly predicted by the propensity model or that interact in complex ways not fully captured by it. To address this, we apply the additional calibration step raking, or iterative proportional fitting.

Raking adjusts the IPW-derived weights so that the weighted UKMOD sample exactly reproduces known marginal distributions of key characteristics observed in the target population. We use official population estimates to do this. This ensures that the final reweighted sample not only resembles the target dataset in structure but also aligns with benchmarks for the Greater Essex population. The raking step targets several core dimensions like:

- Age group distributions (e.g., 18–25, 26–45, 46–65, 66+)
- Employment status (e.g., employed full-time, part-time, retired, unemployed)
- Household composition indicators such as the presence of children

Raking improves the representativeness and interpretability of the weighted UKMOD sample by constraining the final weights to meet these margins, see figure 10.

Figure 10: Propensity Score Distribution After Raking

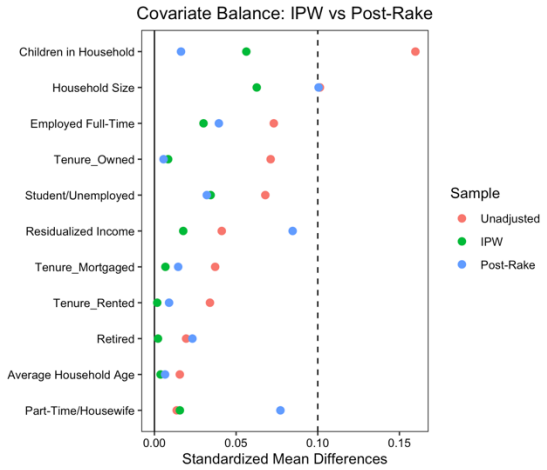


Covariate Balance Diagnostics

The next step assesses whether covariates are balanced between the treated and weighted control groups. Standardised Mean Differences (SMDs) are used to evaluate the difference in means across groups for each covariate, scaled by their pooled standard deviation. Post-weighting SMDs below 0.1 are typically considered acceptable.

Covariate balance checks as shown in figure 11 confirm the weighting process has neutralised confounding relationships between treatment assignment and covariates. In cases where multiple covariates remain imbalanced after weighting, the model may need to be re-estimated with additional interaction terms or alternate specifications.

Figure 11: Covariate Balance



Sample Retention and Effective Sample Size (ESS)

An important diagnostic is how much of the original sample is retained and how much of it meaningfully contributes after reweighting. Unlike matching methods, IPW retains all control observations, allowing the full dataset to be used in analysis. However, if very large weights are assigned to only a small subset of control units, the resulting estimates can become unstable and less reliable.

The Effective Sample Size (ESS) quantifies the amount of usable information remaining after weighting. A low ESS suggests that only a few high-weight observations are driving the results, even if many observations remain in the dataset. This undermines both

precision and generalisability. In contrast, a high ESS indicates that the reweighted sample distributes influence more evenly, preserving the richness of the original data while improving comparability.

In this case, the ESS remained high after applying IPW, indicating that most matched UKMOD households retained meaningful influence in the final weighted sample.

Final Output Data

The final output is a reweighted version of the UKMOD dataset, designed to mirror the demographic structure of Greater Essex as observed in the Experian data and official statistics. Each household in UKMOD is assigned a propensity score, which reflects the likelihood of resembling an Experian household based on characteristics such as age, tenure, household size, employment status, and presence of children. These scores are estimated using a gradient boosted model that includes interaction terms and a residualised version of equivalised income, to improve covariate balance and maintain overlap.

The propensity scores are then used to calculate inverse probability weights, which adjust the influence of each household in the control dataset. Households that more closely resemble the target population receive higher weights, while less similar observations are down weighted. The weights are stabilised and trimmed to prevent extreme values from distorting results and then further adjusted through a raking step.

The result is a synthetic microdata set that preserves the structure and richness of UKMOD while offering improved representativeness for the Greater Essex population.

8. Macro-Validation of UKMOD-Essex

UKMOD outputs for Greater Essex are validated by comparing income estimates to external official benchmarks. The current focus is on two market income components: employment and self-employment income. These are benchmarked against administrative and survey sources, including the Survey of Personal Incomes (SPI) and the Annual Survey of Hours and Earnings (ASHE).

The figures reflect reweighted UKMOD estimates, calibrated to match Greater Essex's population totals and income structure. Comparisons are made both in aggregate and across the distribution, covering recipient counts, average incomes, and percentiles. Each component uses the most recent available external data, typically for 2022 or 2022/23. Where relevant, the UKMOD estimates are filtered or broken down to match how populations are defined in external sources, such as SPI's focus on tax filers.

Employment Income

Earnings are a key exogenous input to UKMOD, so it's important to benchmark against the Survey of Personal Incomes. SPI is based on administrative records from HMRC and only includes individuals with income above the personal allowance threshold, meaning it captures the taxable population. In SPI Table 3.14 (2022–23), the average taxable employment income in Greater Essex is reported as £41,900. Multiplying this by the number of people in Essex with taxable employment income (578,000) gives a total of around £24.2 billion. It's worth noting that the SPI average is conditional; it only reflects those with taxable income from employment, excluding anyone with zero income or whose main income comes from pensions or self-employment.

An alternative employment income estimate was produced using ASHE data (2024), where the average gross annual pay for employee jobs in Greater Essex was £41,915. ASHE - the Annual Survey of Hours and Earnings. This is a business survey based on employer payroll records and provides earnings data for employees in the UK. This figure includes individuals earning below the personal allowance and those in part-time or low-wage employment. It was applied to the total count of employee jobs in the region (637,000). However, ASHE measures jobs rather than individuals, meaning those with multiple jobs may be double counted. The data also exclude the self-employed and individuals not paid during the reference period.

Both sources are limited by definitional constraints and population coverage differences. SPI omits non-taxpayers and has stronger coverage of high earners; ASHE offers more inclusive employment coverage but excludes non-employees. The UKMOD estimate sits slightly above the external range, suggesting higher average income or

coverage effects after reweighting. This is expected given the use of grossed-up Family Resources Survey data scaled to reflect regional population and income distribution targets from Experian and administrative benchmarks.

Distributional Comparison (Percentiles and Cumulative Earnings)

Table 2 compares monthly employment income percentiles across UKMOD and ASHE for Greater Essex. The two distributions are well aligned at the median: UKMOD gives a P50 of £2,392, while ASHE reports £2,535. That match is encouraging, given how different the two sources are. UKMOD income is self-reported in the FRS and includes everyone with non-zero earnings, while ASHE is job-based and reflects PAYE records for employees in April.

At the lower end (P10 to P30), UKMOD is consistently a bit higher. For example, P10 is £867 in UKMOD versus £839 in ASHE. This probably reflects two things. First, UKMOD includes more part-time and low-hour workers, people who may not be captured in ASHE if they weren't paid that month or are on irregular contracts. Second, there's smoothing in UKMOD because some incomes are imputed based on characteristics, which means extremely low reported values, including zeroes, are less common.

At the upper end from P70 to P90, UKMOD continues to track ASHE well. Some values are higher, but not by much. For instance, P70 is £3,324 in UKMOD compared to £3,313 in ASHE. At P90, however, UKMOD shows a noticeably higher value (£5,759) than ASHE (£4,667), suggesting more top-coding or tail adjustment in the ASHE sample. Beyond P80, the ASHE data are more limited, especially for male and female breakdowns, so comparisons become harder.

Figure 12 shows this in cumulative terms. The UKMOD curve is slightly to the right of ASHE across the middle of the distribution, meaning that at any given earnings level, slightly fewer people fall below that amount in UKMOD. This fits with what we see in the percentiles: UKMOD is shifted slightly higher but follows the same shape.

Overall, the comparison suggests that the reweighted UKMOD sample gives a credible distribution of earnings for Greater Essex. It slightly overestimates income at the bottom, but it tracks well at the median and upper end.

Distributional Comparison (Monthly Employment Income)

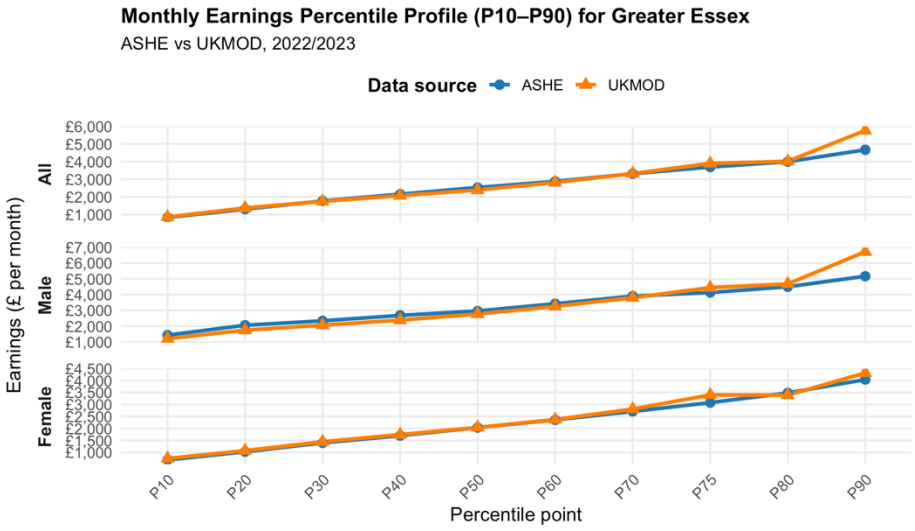
The table below compares monthly employment income percentiles between UKMOD and ASHE for Greater Essex. Percentiles are based on gross income and are grouped by sex.

Table 2: Monthly Employment Income Comparison

Percentile	ASHE (All)	UKMOD (All)	ASHE (Male)	UKMOD (Male)	ASHE (Female)	UKMOD (Female)
P10	£839	£867	£1,431	£1,205	£689	£750
P20	£1,309	£1,378	£2,059	£1,746	£1,028	£1,075
P30	£1,777	£1,755	£2,337	£2,063	£1,399	£1,447
P40	£2,155	£2,076	£2,683	£2,383	£1,702	£1,746
P50 (Median)	£2,535	£2,392	£2,964	£2,773	£2,038	£2,045
P60	£2,880	£2,817	£3,422	£3,254	£2,359	£2,375
P70	£3,313	£3,324	£3,905	£3,805	£2,712	£2,812
P75	£3,692	£3,900	£4,135	£4,450	£3,079	£3,400
P80	£3,996	£4,017	£4,500	£4,680	£3,484	£3,393
P90	£4,667	£5,759	£5,167	£6,725	£4,042	£4,320

Note: ASHE figures are gross monthly pay estimates derived from annual values. UKMOD estimates are gross monthly earnings from simulated microdata, reweighted for Greater Essex.

Figure 12: Monthly Earnings Percentile Profile



Self-Employment Income

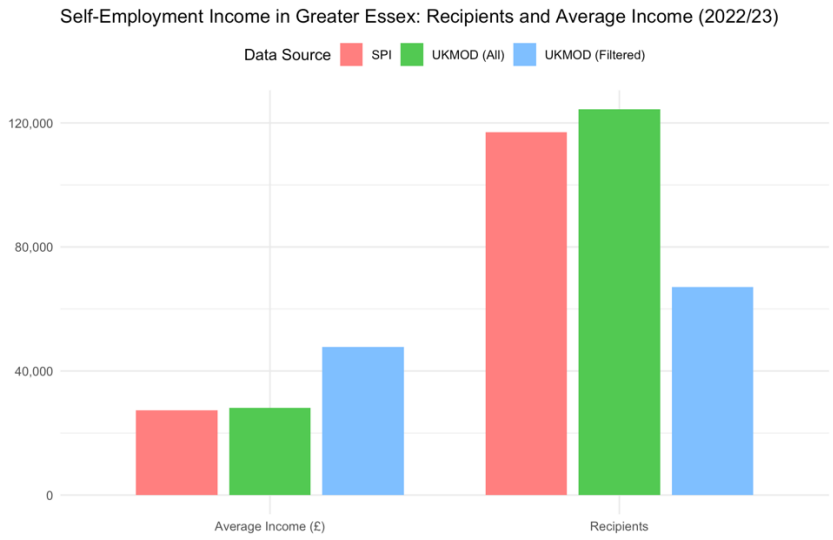
The external benchmark for self-employment income is derived from SPI Table 3.13 and SPI Table 3.14, based on a total of 117,000 individuals in Greater Essex with taxable self-employment income. The average income for this group is £27,365, giving a total of approximately £3.2 billion. These values are calculated directly from the SPI headcount (in thousands) and the corresponding mean income across the 14 Greater Essex districts.

As noted in SPI documentation, these estimates include only individuals whose self-employment income was taxable in 2022–23. That means the figure excludes anyone whose income was fully offset by losses or capital allowances or whose profit was below the reporting threshold. The reported income reflects net taxable profits after deductions. The SPI average is calculated conditionally: it includes only those with non-zero taxable self-employment income.

The unfiltered UKMOD estimate, based on all individuals with non-zero self-employment income, yields a total of £3.50 billion across 124,325 recipients, with an average income of £28,147 as shown in figure 12. This reflects gross reported income before deductions and includes individuals whose income may fall below the tax threshold or would not be considered taxable under SPI definitions.

To align more closely with SPI, a filtered UKMOD estimate was produced by restricting the population to individuals with annual self-employment income above £12,570, matching the personal allowance threshold. Under this condition, UKMOD identifies 66,990 recipients with total income of £3.20 billion and an average income of £47,790. While the number of recipients is lower than the SPI total, the total income matches exactly, lending strong support to the calibration and reweighting process. The higher average income in the filtered sample reflects the use of gross pre-deduction amounts and the exclusion of lower earners. Taken together, the unfiltered and filtered UKMOD estimates provide a credible range for self-employment income and affirm the validity of the reweighted dataset for this component.

Figure 13: Self-Employment Income



9. Discussion

In this paper we have presented a new approach to reweighting input data for microsimulation models, with an application to a regional variant for greater Essex of the UK-wide UKMOD tax-benefit model. This regional variant differs from the UK counterpart only with respect to the input data, but the tax-benefit model remains the same. Both the model and the data are freely available for download.[7] An online version also exists that allows users to design bespoke tax-benefit scenarios, run them online and compare results with baseline simulations, without the need to download the software, the model, and the input data.[8]

Our approach to creating new input data for Essex, based on a Gradient Boosted Machine to estimate propensity scores, leverages on the availability of large household-level data for the region, which provide a target joint distribution for some relevant variables. This allowed us to move beyond targeting marginal distributions coming from official statistics or other sources, as in most of the current literature on spatial microsimulation, although we also performed a final raking stage to address some quality issues of the target micro data. Our validation of the outcomes of the resulting regional microsimulation model shows a good fit with external statistics.

But is our proposed approach, with its positive results, of more general interest or should it be considered a display of technical prowess in a lucky case where regional information was already available? Three things should be considered in this respect. First, the (commercially) available micro-data for Greater Essex were limited to a handful of variables and were by no means sufficient in themselves to be used as inputs for a tax-benefit model. Second, while we had an almost one-to-one sample of the Essex population, information was in some cases departing from official sources, most likely because of imputation issues in the micro-data. These two things together imply that our case is favourable, but not exceptionally so. Our third point is that our setting is also not likely to be exceptionally rare. We live in an era of data, and new sources – from network companies such as utilities or social media, from monitoring agencies, etc. – are

[7] See https://www.microsimulation.ac.uk/ukmod/access/ for more information.
[8] See https://www.microsimulation.ac.uk/ukmod/ukmod-explore/.

becoming increasingly available. We have therefore the hope that our work will be useful for other applications.

Limitations and Technical Considerations

Several limitations of the proposed approach warrant careful reflection. First, while the use of gradient boosted machines (GBM) improves covariate balance over traditional reweighting methods, it comes at a significant computational cost. Training and tuning GBM models particularly with cross-validation and interaction terms involves substantial time and processing power, which may constrain uptake in settings without dedicated computational resources.

Second, despite being less reliant on strict parametric assumptions, machine learning models introduce other forms of complexity. Model selection, prevention of overfitting, and interpretability all pose challenges, particularly in high-dimensional or imbalanced datasets. Our use of residualised income, interaction terms, and post-hoc diagnostics helped mitigate these risks, but the approach still requires considerable technical oversight.

Critically, the performance of the method hinges on the quality and representativeness of the target data. The success of our matching and weighting pipeline reflects the relatively complete coverage and internal consistency of the Experian dataset used for calibration. In scenarios where only sparse, biased, or inconsistently defined commercial data are available, similar results may be difficult to achieve. Moreover, the need for post-weighting raking to align with official marginal distributions indicates that even sophisticated models struggle to account for all dimensions of population heterogeneity.

Finally, our findings highlight the importance of sample design. Naively matching the full Experian sample (740,000 households) against a much smaller control pool (25,000 UKMOD households) led to poor overlap and unstable weights. This underscores a critical insight for applied users: increasing sample size alone does not guarantee better results especially when the distributions are structurally mismatched. In such cases, stratified or balanced subsampling can materially improve performance and should be considered.

Broader Applicability and Future Directions

Although this project was grounded in the Greater Essex context, the methodology is broadly applicable to regional microsimulation development across the UK and beyond. The growing availability of commercial and administrative microdata opens new possibilities for fine-grained policy analysis, provided that appropriate matching and reweighting techniques are applied.

To support broader use, future work should test this approach across diverse local contexts, including areas with more limited data infrastructure. Exploring alternative machine learning methods such as neural networks, extreme gradient boosting, or hybrid ensemble models could further enhance robustness, particularly for extreme class imbalance or nonlinear relationships.

Policy and Practical Implications

From a policy perspective, locally calibrated microsimulation tools like UKMOD-Essex provide much-needed granularity for understanding the effects of tax and benefit policies at the regional level. By grounding analysis in real, locally reflective data, policymakers can assess how proposed reforms would play out in their specific jurisdiction capturing differences in employment patterns, household structure, housing costs, and more.

Equally important is the model's accessibility. The open-source nature of UKMOD-Essex, the availability of the micro-data, and the integration of the model into an online platform allow both analysists with computational skills and non-specialists from local government staff to civil society groups to run policy scenarios of interest. This usability is key to promoting wider engagement with evidence-based policy tools, encouraging more democratic and inclusive decision-making.

Ultimately, this work demonstrates the feasibility and value of blending traditional microsimulation techniques with modern machine learning and commercial data to meet the evolving needs of policymaking. Continued methodological innovation,

combined with clear communication and stakeholder collaboration, will be essential to maximise their contribution to evidence-based policymaking.

10. References

Austin, P. C. (2011) ‘An introduction to propensity score methods for reducing the effects of confounding in observational studies’, Multivariate Behavioral Research, 46(3), pp. 399–424.

Austin, P. C. and Stuart, E. A. (2015) ‘Moving towards best practice when using inverse probability of treatment weighting (IPTW) using the propensity score to estimate causal treatment effects in observational studies’, Statistics in Medicine, 34(28), pp. 3661–3679.

Boscolo, S., Figari, F., Fiorio, C., Matranga, M., Matsaganis, M. (2025, forthcoming) ‘EUROMODspatial Italy: a tax-benefit spatial microsimulation model for the analysis of public policies at local level’ *mimeo*.

Breiman, L. (2001) ‘Random Forests’, Machine Learning, 45(1), pp. 5–32.

Cortes, C., Mohri, M. and Storcheus, D. (2019) ‘Regularised Gradient Boosting’, Advances in Neural Information Processing Systems, 32.

Crump, R. K., Hotz, V. J., Imbens, G. W. and Mitnik, O. A. (2009) ‘Dealing with limited overlap in estimation of average treatment effects’, Biometrika, 96(1), pp. 187–199.

DiNardo, J., Fortin, N. M. and Lemieux, T. (1996) ‘Labor market institutions and the distribution of wages, 1973–1992’, Econometrica, 64(5), pp. 1001–1044.

Fortin, N., Lemieux, T. and Firpo, S. (2010) ‘Decomposition methods in economics’, NBER Working Paper No. 16045.

Friedman, J. H. (2001) ‘Greedy function approximation: A gradient boosting machine’, Annals of Statistics, 29(5), pp. 1189–1232.

Imbens, G. W. (2000) ‘The role of the propensity score in estimating dose-response functions’, Biometrika, 87(3), pp. 706–710.

King, G. and Zeng, L. (2001) ‘Logistic regression in rare events data’, Political Analysis, 9(2), pp. 137–163.

Lee, B. K., Lessler, J. and Stuart, E. A. (2010) ‘Improving propensity score weighting using machine learning’, Statistics in Medicine, 29(3), pp. 337–346.

Leite, W., et al. (2024) ‘Machine learning for propensity score estimation: a systematic review and reporting guidelines’, Preprint on OSF.

Lunceford, J. K. and Davidian, M. (2004) ‘Stratification and weighting via the propensity score in estimation of causal treatment effects: A comparative study’, Statistics in Medicine, 23(19), pp. 2937–2960.

Panori, A., Ballas, D., Psycharis, Y. (2017) ‘SimAthens: A spatial microsimulation approach to the estimation and analysis of small area income distributions and poverty rates in the city of Athens, Greece’ Computers, Environment and Urban Systems, 63, pp. 15-25.

Richiardi, M., Collado, D. and Popova, D. (2021) ‘UKMOD – A new tax-benefit model for the four nations of the UK’, International Journal of Microsimulation, 14(1), pp. 92–101.

Robins, J. M., Hernán, M. A. and Brumback, B. (2000) ‘Marginal structural models and causal inference in epidemiology’, Epidemiology, 11(5), pp. 550–560.

Sologon, D. M., Doorley, K. and O'Donoghue, C. (2023) ‘Drivers of income inequality: What can we learn using microsimulation?’, in Zimmermann, K. F. (ed.) Handbook of Labor, Human Resources and Population Economics. Springer.

Stuart, E. A. (2010) ‘Matching methods for causal inference: A review and a look forward’, Statistical Science, 25(1), pp. 1–21.

Zhao, P., Su, X., Ge, T., Fan, J. (2016) ‘Propensity Score and Proximity Matching Using Random Forest’ Contemp Clin Trials, (47), pp. 85–92.

Appendix 1- Full Sample

Figures A1 – A5

Figure A1: Propensity Score Distribution After Common Support Trimming

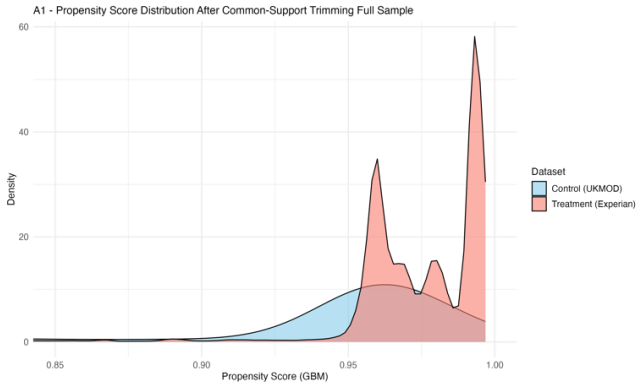


Figure A1 shows that common support trimming with the full sample produces poor distributional overlap between the datasets. Despite removing extreme propensity scores, the fundamental imbalance persists with UKMOD households concentrated in lower ranges and Experian households dominating higher ranges, revealing inadequate common support for reliable matching procedures.

Figure A2: Propensity Score Distribution After Matching

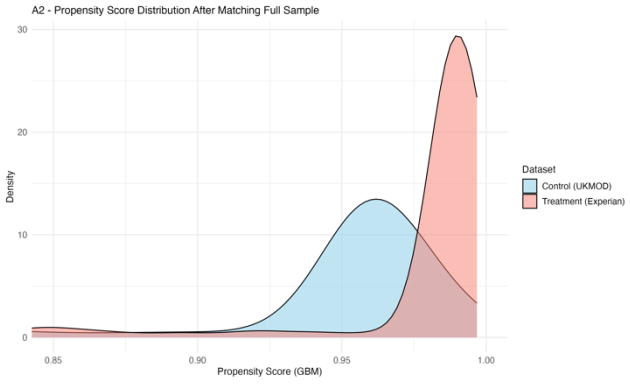


Figure A2 shows that matching with the full sample achieves minimal improvement in distributional overlap. The matching procedure struggles to find adequate comparable units due to the extreme sample size disparity, leaving the distributions largely unchanged with persistent separation between UKMOD and Experian households.

Figure A3: Propensity Score Distribution After IPW Weighting

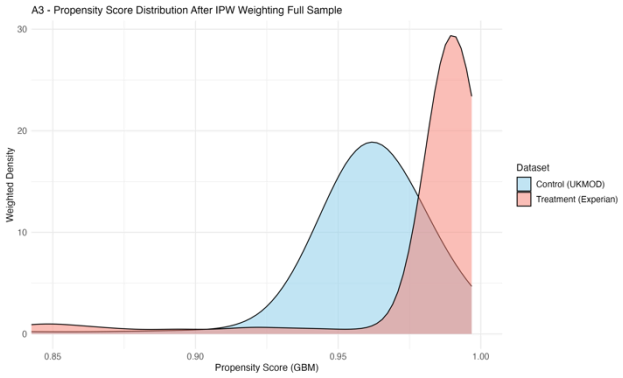


Figure A3 shows that IPW weighting following matching provides marginal distributional improvement but fails to achieve adequate balance. While the weighting procedure attempts to adjust for remaining imbalances, the fundamental separation persists, with both distributions maintaining distinct peaks and limited overlap throughout the propensity score range.

Figure A4: Propensity Score Distribution After Raking

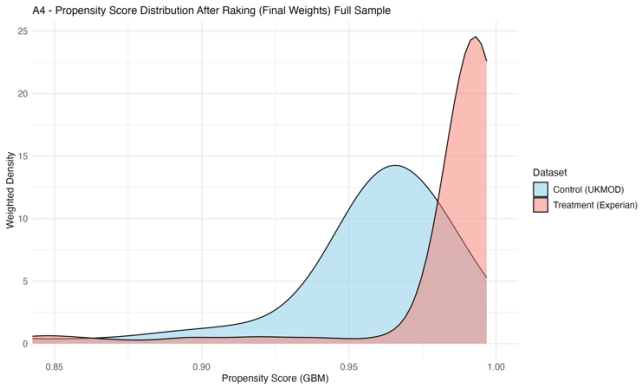


Figure A4 shows that final raking adjustments with the full sample provide only modest distributional improvements. While raking to population margins creates slightly better overlap compared to IPW alone, the underlying structural imbalance remains evident, with the distributions still exhibiting distinct peaks.

Figure A5: Covariate Balance

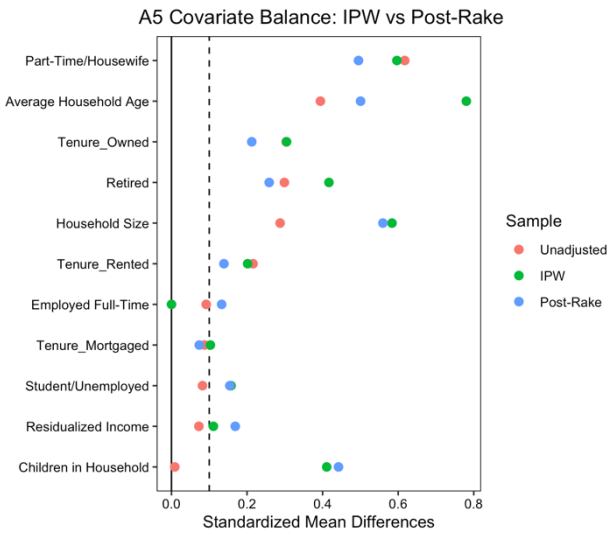


Figure A5 above shows that covariate balance with the full sample remains poor despite sequential adjustments. Multiple variables exceed the 0.1 threshold for standardized mean differences, with several covariates showing substantial imbalances even after post-raking corrections.