ISSN 2282-6483

Alma Mater Studiorum - Università di Bologna
DEPARTMENT OF ECONOMICS

Benchmark-Guided Sample Construction from Non-Probability Firm Databases

Silvia Sarpietro
Tommaso Sonno

Quaderni - Working Paper DSE N° 1227

Benchmark-Guided Sample Construction from Non-Probability Firm Databases

Silvia Sarpietro* Tommaso Sonno†

July 2026

Abstract

Commercial firm-level databases are not probability samples, and their selective coverage can distort estimates of unconditional population objects such as aggregate productivity and firm-size distributions. At the same time, researchers may want to work with a smaller analytical sample because the full database is costly to process or uneven in data quality. This paper proposes a benchmark-guided stratified subset-selection algorithm for this setting. Given a target sample size and official cell-level aggregate statistics, the method selects firms so that the retained sample matches official sector-by-size cell counts and aligns within-cell benchmark moments. The selected sample is then adjusted using mild post-stratification weights. The algorithm can be viewed as a sparse, design-stage analogue of calibration. We evaluate the method using Italian manufacturing firms from Orbis in 2022, with OECD Structural Business Statistics as external benchmarks. In the raw Orbis pool, cell-weighted turnover per employee is overstated by about 14%. The proposed algorithm reduces this distortion to about 5% and produces near-uniform final weights, while post-stratification alone leaves the distortion essentially unchanged. Full-pool calibration reproduces the moments it constrains, but in this application it performs less well on non-targeted functionals, and bounded cell-level calibration becomes infeasible in many cells under finer stratifications. The results show that benchmark-guided subset selection can complement weighting-based corrections when researchers work with large non-probability firm databases and want a smaller sample aligned with official aggregate statistics.

JEL codes: C13, C81, C83.

Keywords: non-probability samples, firm databases, Orbis, stratification, calibration, effective sample size.

*Department of Economics (DSE), University of Bologna. Email: silvia.sarpietro@unibo.it.

†Department of Economics (DSE), University of Bologna, and Centre for Economic Performance (CEP), London School of Economics. Email: tommaso.sonno@unibo.it.

Non-technical summary

Commercial firm-level databases have become indispensable in empirical economics. Thousands of studies rely on them to investigate productivity, market concentration, employment dynamics, and firm growth across countries. Yet these databases are not statistical samples: they are assembled from administrative filings and commercial data providers, with no sampling design. As a result, their composition can differ markedly from the actual population of firms. Large companies tend to be over-represented, small firms are frequently missing, and coverage varies widely across countries, sectors, and years. When researchers compute aggregate statistics directly from these raw data, the numbers can be substantially misleading. In Italian manufacturing, for example, the average turnover per employee of the typical firm environment, calculated from raw Orbis records, overshoots the official figure by roughly fourteen percent.

The standard remedy is to reweight observations so that the weighted sample matches known population totals. Reweighting helps, but it has limits: when certain types of firms are entirely absent or extremely scarce in the database, no amount of weight adjustment can compensate for what is not there. This paper proposes a different approach. Instead of reweighting the full database, we use official aggregate statistics from the OECD Structural Business Statistics to *select* a subset of firms whose composition, by sector and size class, mirrors the known population structure. Only after this selection step do we apply mild post-stratification weights to fine-tune the match.

We formalise the algorithm, characterise the conditions under which subset selection outperforms reweighting alone, and test the method on Italian manufacturing data for the year 2022. The results are encouraging: the fourteen percent distortion drops to about five percent, and to two percent for the national aggregate; the selected subset exhibits a stable, well-behaved weight distribution; and the correction carries over to firm outcomes, such as profits and assets, that the algorithm never uses directly. The method is transparent, replicable, and can be applied to any country and sector for which official benchmark statistics are available.

1 Introduction

Commercial firm-level databases have become central to empirical economics. They provide standardised financial data on millions of firms worldwide and underpin a large and growing body of research on productivity, market power, ownership structure, and firm dynamics. For many empirical questions, these databases are well suited: studies of large firms, multinational networks, within-firm panel variation, or conditional relationships estimated on the intensive margin can tolerate, or even benefit from, the particular composition of a commercial database. Research on the rise of market power based on listed-firm data (De Loecker et al., 2020), for instance, can proceed with data that over-represent large enterprises, because the objects of interest are defined at the firm level or are conditional on observables.

Yet a distinct class of empirical objects requires unconditional population quantities, such as aggregate labour productivity, firm size distribution, employment share of small and medium enterprises, or average turnover per worker. These objects depend on the composition of the sample mirroring that of the population along dimensions where coverage is selective. When it does not, the resulting estimates can be substantially biased, and the direction of the bias is typically predictable: because large firms file more completely and are more likely to appear in commercial registers, raw aggregates tend to overstate average firm size, productivity, and capital intensity, while understating the prevalence of small enterprises. The question, then, is not whether a specific database is "good" or "bad" in general, but whether the specific empirical object at hand requires representativeness along dimensions where coverage is known to be non-random.

We ground the argument in the leading case: Bureau van Dijk's Orbis, the most widely used commercial firm-level database. The starting point for any serious use of Orbis is careful data construction. Kalemli-Özcan et al. (2024) demonstrate that a disciplined vintage-by-vintage downloading and cleaning protocol, validated against official statistics, substantially improves the representativeness of Orbis for cross-country research; indeed, they argue that when their guidelines are followed, no reweighting is needed, and that reweighting with ad hoc weights can introduce further errors. Their contribution has set a new standard for the field, establishing that much of the perceived "Orbis bias" reflects avoidable errors in data handling rather than inherent deficiencies in the source. Our work is complementary. We study what to do when the available extract of the database still displays economically consequential distortions relative to official cell-level benchmarks. The construction history of the extract matters for how large these residual distortions are, and we document it explicitly for our application; the question we ask is what can be done about the distortions that remain, whatever their origin. Bajgar et al. (2020) document such distortions systematically, showing that Orbis coverage varies widely across countries, sectors, and size classes, with small firms particularly under-represented, and,

importantly for our framework, that the tilt towards larger and more productive firms persists even within size classes. The patterns they identify are precisely those that generate first-order bias in unconditional aggregates. The practical implication is that even after implementing the best available cleaning protocol, a researcher computing population-level statistics from Orbis may face residual compositional distortions that are large enough to change substantive conclusions.

The natural response to compositional distortions is reweighting. Calibration estimators (Deville and Särndal, 1992), entropy balancing (Hainmueller, 2012), and their bounded and trimmed variants (Deville et al., 1993; Théberge, 1999; Kott, 2006) are well understood and widely available. We do not claim that weighting is a new idea, nor do we argue that it is ineffective. Our question is narrower and, we believe, practically important: under what conditions is reweighting the full pool insufficient, and when does a design-stage intervention, in the form of benchmark-guided subset selection, yield lower bias on the objects of interest? The distinction matters because calibration weighting solves an optimisation problem over the entire available pool. When some cells of the cross-classification (sector by size class) are nearly empty, when within-cell distributions exhibit heavy tails, or when the selection process that determines which firms appear in the database is correlated with the target variable conditional on the stratifying variables, calibration weights can become extreme, unstable, or unable to correct the bias at all. In these settings, reducing the pool to a subset whose cell composition matches the population structure before applying mild post-stratification weights can outperform full-pool calibration, because the design stage absorbs compositional distortions that the estimation stage cannot handle gracefully.

A second, complementary motivation concerns sample size. Researchers frequently want an analytical sample far smaller than the full database: processing and validating hundreds of thousands of balance sheets is costly, data quality is uneven across firms, and downstream tasks such as manual checks, record linkage, or repeated estimation scale poorly with sample size. Fixing a target sample size and asking which observations to retain turns representativeness into a design problem. The terms of this trade-off should be stated plainly: constructing the sample requires access to the full pool once, so the gains accrue downstream, where every subsequent analysis operates on a small, quality-screened, benchmark-aligned dataset. Discarding observations is justified in this setting not because information is free to lose, but because part of the pool is of doubtful quality for the purpose at hand and because the deliverable itself is a manageable sample.

Hellerstein and Imbens (1999) and Chen et al. (2023) develop related ideas using aggregate auxiliary information, as we do; the difference is that their corrections operate entirely through reweighting a fixed sample, whereas we study when intervening at the design stage, by choosing which observations enter the sample, adds value over weighting alone.

This connects to a broader methodological literature on inference from non-probability samples. The design-versus-estimation question itself is as old as survey sampling: Neyman (1934) contrasted purposive selection, choosing units so that the sample matches known population characteristics, with stratified random sampling, and Smith (1983) analyses the conditions under which inference from such non-random selections is valid. The modern incarnation with a large online pool is sample matching (Rivers, 2007), and Meng (2018) formalises the sense in which a small, well-composed sample can dominate a much larger but selectively distorted one. On the estimation side, Wu (2022), Elliott and Valliant (2017), and Valliant (2020) provide comprehensive treatments of the available corrections, including propensity-based approaches, mass imputation (Kim et al., 2021), and doubly robust estimators (Chen et al., 2020). A common thread in this literature is the availability of a reference probability sample or, at minimum, a micro-level sampling frame against which the non-probability sample can be calibrated (Särndal et al., 1992). In our setting, neither is available. The researcher observes a non-random pool of firms (Orbis) and a set of aggregate cell-level benchmarks from official statistics (OECD Structural Business Statistics), but has no micro-level frame and no probability sample from the same population. Balanced sampling (Deville and Tillé, 2004) offers a design-based approach to selecting samples that satisfy balancing constraints, but it presupposes a complete enumeration of the population from which units can be drawn with known inclusion probabilities. Relative to the purposive-selection lineage, then, the distinctive features of our setting are that the selection is guided by cell-level aggregate moments alone, and that we evaluate the algorithm explicitly against feasible weighting alternatives on both bias and stability. The classical vulnerability of purposive and quota methods, selection on unobservables within cells, is precisely what the structural-link condition of our theoretical framework (Section 3) addresses.

The contribution of this paper is to study when and why benchmark-guided stratified subset selection from a non-probability firm database improves the estimation of unconditional population objects, relative to feasible weighting-only corrections applied to the full pool. We characterise the conditions under which subset selection is expected to add value: sparse cells in the cross-classification, heavy-tailed within-cell distributions, and within-cell selection distortions that correlate the database inclusion process with the target variable. We formalise the algorithm, which can be viewed as a sparse, design-stage analogue of calibration: excluded firms receive a weight of exactly zero, and the retained sample carries near-uniform weights. We characterise the conditions under which this design-stage correction is expected to outperform full-pool calibration. The algorithm is intended as a complement to, not a replacement for, the careful data construction advocated by Kalemli-Özcan et al. (2024). It applies downstream, after the database has been cleaned and standardised, and addresses residual compositional distortions that cleaning alone cannot eliminate. The key insight is that the choice of which observations to retain

(design) and the choice of how to weight retained observations (estimation) are distinct levers, and that their relative performance depends on the estimand: in the application, the weighting lever is more accurate on the functional built from the very moments both methods target, while the design lever is more accurate on the aggregate functional that neither targets directly.

We apply the algorithm to Italian manufacturing firms drawn from Orbis for the year 2022, using OECD Structural Business Statistics as the benchmark, and we are careful to define the estimands first: the aggregate ratio of totals (an employment-weighted object) and the firm-count-weighted average of cell-level ratios (which reflects the typical firm environment) are different population quantities, and each requires its own sample analogue. The raw Orbis pool overstates cell-weighted turnover per employee by 14% and the aggregate ratio by 6.8%. Post-stratification leaves the first distortion untouched, by construction: the problem is within-cell composition, not cell shares. The proposed algorithm reduces the two distortions to 4.9% and 2.0% (means across 20 independent runs of the stochastic search), matches over 90% of cells within $\pm 10\%$, and produces near-uniform weights (effective sample size 2,913 out of 2,920 firms; Kish, 1965). The comparison with full-pool calibration reveals specialisation rather than dominance: calibration is nearly exact on the functional built from the very moments it constrains (1.8% deviation), but less accurate than the proposed algorithm on the aggregate functional that neither method targets (5.0% versus 2.0% on turnover per employee), and both estimators prove stable under pool perturbations once they are properly re-estimated rather than mechanically truncated. Finally, the correction propagates to outcomes the algorithm never targets: estimates of EBITDA, assets, and profits per employee move by 5 to 9% when the composition is corrected, confirming that the choice of correction method matters for the variables researchers ultimately study.

The remainder of the paper is organised as follows. Section 2 describes the benchmark-guided stratification algorithm. Section 3 presents the theoretical framework. Section 4 reports the empirical application to Italian manufacturing. Section 5 concludes.

2 The Algorithm

This section describes the benchmark-guided subset-selection algorithm in operational terms. We first present an overview of the three stages (Section 2.1), then discuss how the algorithm handles cells that cannot be tightly matched (Section 2.2), provide a formal interpretation of the algorithm as an approximate optimisation problem (Section 2.3), and conclude with a discussion of scope and limitations (Section 2.4).

2.1 Overview

The algorithm operates in three stages: stratification, benchmark-guided subset selection, and post-stratification weight correction. Each stage is described in turn.

Stage 1: Stratification. Partition firms into cells $c \in \mathcal{C}$, where $\mathcal{C}$ denotes the set of cells, defined by the Cartesian product of country, sector (NACE Rev. 2, 2-digit), and employee size class. We adopt four size classes throughout: 10–19, 20–49, 50–249, and 250+ employees. This classification follows the standard breakdown used in the OECD Structural Business Statistics (SBS), which serves as the source of official benchmarks.

For each cell c in which the official statistics report data, we observe the number of enterprises N_c and the cell-level moments m_c^x of a vector of benchmark variables x: average employment per enterprise, value added per employee, and turnover per employee. The cell shares $\pi_c = N_c / \sum_{c'} N_{c'}$, where the sum runs over all cells, define the target composition of the representative subset.

Stage 2: Benchmark-guided subset selection. For each cell c, we select a subset S_c of size n_c from the pool of available firms in the non-probability database (Orbis). The target cell sizes n_c are set proportional to the population shares π_c, scaled to the desired total sample size n, with oversampling to ensure a minimum of 2 firms per non-empty cell:

$$n_c = \max\{2,\ \lceil \pi_c \cdot n \rceil\}\,, \tag{1}$$

where $\lceil \cdot \rceil$ denotes rounding up to the nearest integer.

The selection within each cell proceeds through a stochastic search with the following steps.

(i) **Score firms on data completeness.** Each firm i receives an observation quality score $q_i \in \{0, 1, \ldots, 5\}$ equal to the number of non-missing values among five key variables: age, revenue, intangible assets, tangible assets, and number of employees. This score does not enter the acceptance criterion directly; it serves as a soft priority mechanism to favour firms with richer data. Because reporting quality may correlate with performance, the score is itself a potential selection channel: the acceptance criterion disciplines its effect on the benchmark means, and the empirical application verifies directly that neutralising the score does not drive the corrections (Section 4).

(ii) **Draw a candidate subsample.** Firms are sorted by descending quality score q_i, with ties broken by $u_i \sim \text{Uniform}(0, 1)$, drawn independently for each firm at each iteration. The top n_c firms with non-missing values of the relevant benchmark variables are retained as the candidate subsample. The combination of a deterministic quality score and a random perturbation ensures that different draws explore different subsets while maintaining a preference for data-rich observations.

(iii) **Check the acceptance criterion.** For each benchmark variable x_k (the k-th component of the vector x) available in the official statistics for cell c, compute the sample mean $\bar{x}_k(S_c)$ over the candidate subsample and check whether it falls within the acceptance interval:

$$\bar{x}_k(S_c) \in \left[\alpha \cdot m_c^{x_k},\ (2-\alpha) \cdot m_c^{x_k}\right] \tag{2}$$

where $\alpha \in (0, 1]$ is the lower multiplier and $(2 - \alpha)$ is the upper multiplier, so that the interval is symmetric in proportional terms around the official value $m_c^{x_k}$. A candidate subsample is accepted if and only if it passes the check for every available benchmark variable.

(iv) **Progressively widen the acceptance interval.** The acceptance interval starts tight ($\alpha = 0.9$, corresponding to $\pm 10\%$) and widens in stages as iterations accumulate without a successful draw. The schedule is reported in Table 1. The first candidate that passes all checks is accepted. If no candidate is accepted after 15,000 iterations, the cell is recorded as unmatched.

Table 1: Acceptance-band schedule of the stochastic search

Iterations	α	Acceptance band
1–4,000	0.9	$\pm 10\%$
4,001–5,000	0.8	$\pm 20\%$
5,001–6,000	0.7	$\pm 30\%$
6,001–8,000	0.5	$\pm 50\%$
8,001–11,000	0.3	$\pm 70\%$
11,001–15,000	0.1	$\pm 90\%$

Notes: The stochastic search draws candidate subsamples within each cell and accepts the first draw whose sample means of all available benchmark variables fall jointly within the acceptance band around the official cell moments. The multiplier α defines the band $[\alpha \cdot m_c^{x_k}, (2-\alpha) \cdot m_c^{x_k}]$ of equation (2): the band starts at $\pm 10\%$ and widens stepwise as iterations accumulate without an accepted draw. Cells with no accepted draw after 15,000 iterations are recorded as unmatched.

Not all benchmark variables are available for every cell. The algorithm adapts: only benchmarks for which both the official moment $m_c^{x_k}$ and the corresponding firm-level variable are non-missing and non-zero in the pool are used in the acceptance check. Cells where no benchmark is available are skipped.

Stage 3: Post-stratification weight correction. After selection, the composition of the selected sample may not perfectly match the population cell shares, because of minimum-size oversampling, data-availability constraints, or cells that could not be

matched. Post-stratification weights correct for this discrepancy:

$$w_c \;=\; \frac{\pi_c}{\hat{\pi}_c^S} \;=\; \frac{N_c \,/\, \sum_{c'} N_{c'}}{n_c^{\text{sel}} \,/\, \sum_{c'} n_{c'}^{\text{sel}}} \tag{3}$$

where $\hat{\pi}_c^S = n_c^{\text{sel}} \big/ \sum_{c'} n_{c'}^{\text{sel}}$ is the share of cell c in the selected sample and n_c^{sel} is the number of firms actually retained in cell c. These weights are cell-level (all firms in the same cell receive the same weight) and adjust for composition only, not for within-cell selection. In a well-matched sample, the weights are close to one and exhibit low dispersion.

2.2 Handling of Difficult Cells

Not all cells can be matched tightly. A cell may be difficult for several reasons: the Orbis pool for that cell may contain too few firms relative to the target n_c; the within-cell distribution may be so distorted that no subsample of the required size can reproduce the official moment; or the benchmark variable may be available in the official statistics but systematically missing in the firm-level data.

The progressive widening of the acceptance interval (from ±10% to ±90%) is designed to handle these cases gracefully. The iteration count at which a cell is matched provides a natural diagnostic: cells matched within the first 4,000 iterations are well-approximated; cells requiring wider intervals are flagged as difficult.

Transparent reporting of cell-matching quality is essential. Three diagnostics matter: the share of cells matched within each tolerance band (±10%, ±20%, ±30%, and wider); the identity and characteristics (sector, size class, pool size) of cells that could not be matched or entered at all; and the sensitivity of aggregate results to restricting the sample to well-matched cells only, for instance retaining only cells matched within ±20% and recomputing weights accordingly. The application in Section 4 reports all three diagnostics. The sensitivity diagnostic is particularly important: if aggregate estimates are insensitive to dropping poorly matched cells, the results are robust; if they are sensitive, the poorly matched cells may be driving the findings through artefacts rather than genuine correction.

2.3 Formal Interpretation

The algorithm can be interpreted as an approximate solution to the following constrained optimisation problem. Let $S = \{S_c\}_{c \in \mathcal{C}}$ denote a partition of the selected sample into cells, and let $w = \{w_c\}_{c \in \mathcal{C}}$ denote the associated post-stratification weights. The objective is:

$$\min_{S,\,w} \sum_{c \in \mathcal{C}} \| \hat{m}_c^x(S, w) - m_c^x \|^2 + \mathcal{P}(S, w) \tag{4}$$

subject to cell-level representation (n_c proportional to N_c), minimum cell size ($n_c \geq 2$ for all non-empty cells), weight feasibility ($w_c > 0$ for all c), and moment acceptance ($\bar{x}_k(S_c) \in [\alpha \cdot m_c^{x_k},\ (2-\alpha) \cdot m_c^{x_k}]$ for each cell and available benchmark variable), where $\hat{m}_c^x(S, w)$ is the weighted sample mean of the benchmark variables in cell c, $\|\cdot\|$ is the Euclidean norm, and $\mathcal{P}(S, w)$ is a penalty term that captures weight dispersion or sample instability (e.g., the variance of w_c across cells).

The operational algorithm solves this problem approximately through cell-by-cell stochastic search. Within each cell, random candidate subsamples are drawn and evaluated against the acceptance criterion, with the acceptance region widening progressively if no satisfactory draw is found within a fixed number of iterations. This cell-by-cell structure has the practical advantage of being parallelisable across cells and transparent in its diagnostics, at the cost of not jointly optimising across cells.

A natural alternative would be to select, in each cell, the subset that minimises the distance between sample moments and benchmark moments, in the spirit of the method of moments. We use an acceptance rule with widening bands instead, for three reasons. First, exact minimisation over subsets of a fixed size is a combinatorial problem whose cost grows rapidly with the cell pool, whereas the acceptance rule requires only repeated draws and mean comparisons. Second, the official moments are themselves estimates, and the minimising subset is a single deterministic object that overfits their specific values; drawing at random from the set of subsets whose moments are acceptably close both avoids this overfitting and preserves the randomness needed to assess the variability of the algorithm across seeds, as we do in the application. Randomising among near-minimisers of the distance would achieve the same two goals; the acceptance rule is a computationally cheap implementation of exactly that idea, with the tolerance set by the band schedule rather than by the unknown minimum. Third, the band at which each cell is accepted is recorded, so every accepted cell carries an explicit, cell-specific bound on its moment deviation. The bound is adaptive rather than uniform – the schedule allows it to widen to $\pm 90\%$ in the hardest cells – but it is always observed and reported, which converts matching quality from an assumption into a diagnostic (Section 2.2). There is no contradiction with the optimisation view of (4): the acceptance rule targets the *feasible set* of that problem rather than its minimiser, and the band width bounds the value of the objective attained by any accepted draw.

Remark 1. The operational algorithm is a heuristic approximation to the optimisation problem in (4). The theoretical analysis in this paper focuses on the *principle* of stratify-then-select versus reweight-only, not on proving optimality of the specific search procedure. The distinction between the methodological principle and the computational implementation is important: a different search algorithm that solves (4) more efficiently or more precisely would serve the same methodological purpose, and the theoretical results apply to the class of procedures that satisfy the constraints, not to the specific heuristic.

Remark 2 (Selection as sparse weighting). The algorithm can be viewed as a sparse, design-stage analogue of calibration. A selected sample with post-stratification weights is equivalent to a weighting of the full pool in which every excluded firm (including all firms in skipped cells, see Section 2.4) receives a weight of exactly zero and the retained firms within a cell share a common positive weight; the equivalence is exact for ratio (Hájek) estimators such as (10), whose weights are defined up to a common scale. Standard calibration keeps all observations and, in its bounded form, imposes a strictly positive lower bound on every weight; the algorithm instead sets most weights to zero, performing subset selection on observations in the same way that LASSO-type methods perform selection on variables. The implicit penalty on sample size that this entails is justified when part of the pool is of doubtful quality for the target objects, so that removing observations is preferable to down-weighting them, or when the deliverable itself is a small analytical sample; absent such motivations, discarding data has no intrinsic virtue.

2.4 Scope and Limitations

Several features of the algorithm deserve explicit discussion of their scope and limitations.

The algorithm is designed for estimating unconditional population objects: aggregate means, distributional statistics, population shares. If the object of interest is conditional on the variables used for stratification (within-cell regressions, or firm-level panel models with sector-size fixed effects), the correction is irrelevant by construction, since conditioning on the cell absorbs exactly the margins being corrected. The method matters when the research question requires aggregating across cells or when cell composition affects the estimate.

A second boundary concerns support. If entire categories of firms are absent from the non-probability database, no selection algorithm can represent them: the algorithm improves representation *within* the observed support. When a cell contains zero firms in the database, or lacks the benchmark moments needed for the acceptance check, the cell is skipped. It receives no selected firms, and the estimator implicitly renormalises over the covered cells. Skipped cells are therefore unrepresented in the final sample, a limitation that must be disclosed cell by cell in any application. The method is limited to settings where the database and the benchmarks jointly cover the relevant cells.

The gains from subset selection over full-pool reweighting are largest in precisely the settings where calibration weights become unstable: sparse cells, with few firms over which to redistribute weight; heavy-tailed within-cell distributions, where outliers receive extreme weights under calibration; and non-random within-cell selection, where the firms present in the database are systematically different from the population even within narrowly defined cells. In settings where the database pool is large, well-distributed, and representative within cells, full-pool calibration may perform equally well, and the addi-

tional complexity of subset selection is not warranted.

The algorithm also takes official benchmarks as ground truth, so its validity depends on the quality, timeliness, and comparability of the official statistics used as targets. Benchmark variables must be defined consistently across the official source and the firm-level database. In practice, definitional differences (in the treatment of part-time employees, for instance, or in the valuation of value added) can introduce systematic discrepancies that are not due to sample selection. Any application should document known definitional differences and assess their potential impact.

Finally, the logic of the algorithm is not specific to firm-level databases. It applies, in principle, to any setting where one observes a non-probability pool of units, official aggregate statistics provide cell-level moments for key variables, and the research question requires unconditional population objects: surveys with selective non-response, administrative data with incomplete coverage, or web-scraped datasets. The feasibility of the extension depends on the availability of suitable benchmarks, the comparability of variable definitions, and the degree of support overlap between the pool and the population.

3 Theoretical Framework

This section develops the theoretical foundations of the proposed algorithm. We begin by formalising the distinction between benchmark variables and target outcomes, then define the data-generating environment, introduce the estimators, and state the core result. The result takes the form of a conjecture: it identifies the conditions under which the design-stage correction is expected to improve on full-pool weighting, and it is supported by the theoretical intuition developed below and by the empirical evidence of Section 4.

3.1 Benchmark Variables and Target Outcomes

Official aggregate statistics provide information on a vector of *benchmark variables* x. In the Orbis/OECD setting, the components of x are employment per enterprise, value added per employee, and turnover per employee, observed at the level of sector $\times$ size-class cells. The *target outcome* y that the researcher ultimately wants to estimate at the population level (labour productivity, the firm size distribution, SME employment shares, or other unconditional objects) may or may not coincide with one or more components of x.

The theory therefore works with two objects throughout: the vector $x \in \mathbb{R}^p$ of benchmark-linked variables for which official cell-level moments are known, and the target outcome of interest $y \in \mathbb{R}$. We emphasise that the paper does *not* claim that matching the moments of x uniformly improves estimates of an arbitrary y. Such a claim would be false in general: if y is unrelated to x and to the selection process that shapes the pool, correcting the distribution of x has no bearing on y. The theoretical contribution is to

characterise conditions under which reducing within-cell distortion in x *also* reduces bias for y.

3.2 Population and Pool

Consider a finite population of N firms, partitioned into C cells $c \in \mathcal{C} = \{1, \ldots, C\}$. A cell is defined by the intersection of a sector classification and a size class. Within each cell c, the population joint distribution of (x, y) is F_c, with marginal means $\mu_c^x = E_{F_c}[x]$ and $\mu_c^y = E_{F_c}[y]$. The population cell share is $\pi_c = N_c/N$, where N_c is the number of firms in cell c, so that the unconditional population mean of y is

$$\mu^y \;=\; \sum_{c=1}^{C} \pi_c \, \mu_c^y \,. \tag{5}$$

We do not observe the full population. Instead, we have access to a *non-probability pool* $\mathcal{D}$ that arises from a commercial data-collection process with no controlled sampling design. Within each cell c, the pool contains $|\mathcal{D}_c|$ firms drawn from a distribution G_c that in general differs from F_c:

$$G_c \;\neq\; F_c \qquad \text{for at least some } c \in \mathcal{C} \,. \tag{6}$$

The inequality $G_c \neq F_c$ captures selective coverage: the pool over-represents certain firm types even *within* a given sector $\times$ size-class cell. We model this selection through a latent index η_i associated with each firm i in the population. The index η jointly affects the probability of inclusion in the pool and the realisations of both x and y. For instance, larger, more profitable, and better-reporting firms have higher η, which increases their probability of appearing in the database and simultaneously shifts both their employment-based characteristics (x) and their productivity-based outcomes (y). Formally, conditional on cell c,

$$P(i \in \mathcal{D} \mid c, \eta_i) \;=\; s(\eta_i) \,, \qquad \text{with } s(\cdot) \text{ increasing}, \tag{7}$$

and η_i is correlated with both x_i and y_i within cell c. As a consequence, $E_{G_c}[x] \neq E_{F_c}[x]$ and $E_{G_c}[y] \neq E_{F_c}[y]$ in cells where selection is non-trivial.

Beyond within-cell distortions, the pool also exhibits incorrect cell proportions: $\tilde{\pi}_c \neq \pi_c$, where $\tilde{\pi}_c = |\mathcal{D}_c|/|\mathcal{D}|$.

Available benchmarks. We assume that the researcher has access to two types of aggregate information from official statistical sources: the population cell shares π_c for all $c \in \mathcal{C}$, and the population cell means of the benchmark variables, $m_c^x = \mu_c^x$, for at least some cells. This "benchmarks-only" information structure is the empirically relevant case. The researcher observes cell-level aggregates (from, e.g., OECD Structural Business

Statistics) but does *not* have access to a reference probability sample at the firm level.

3.3 Three Estimators

We compare three approaches to estimating μ^y. A fourth, intermediate estimator (post-stratification only, denoted PS in the empirical section) reweights the full pool to match population cell shares π_c without any within-cell moment matching; it is a special case of Estimator W with the x-moment constraints removed, and we use it as an additional reference point in the application.

Estimator R (Raw pool). The naïve estimator uses the unweighted sample mean of y over the entire pool:

$$\hat{\mu}^R \;=\; \frac{1}{|\mathcal{D}|}\sum_{i\in\mathcal{D}} y_i\,. \tag{8}$$

This estimator corrects for nothing. Its bias combines both the between-cell distortion (wrong cell proportions) and the within-cell distortion (wrong within-cell composition).

Estimator W (Full-pool calibration). Find weights $\{\omega_i\}_{i\in\mathcal{D}}$ on the full pool such that the weighted cell shares match the population cell shares π_c and the weighted cell means of x match the official benchmarks m_c^x. The weights are chosen to minimise a distance criterion (e.g., chi-squared distance from uniform weights, or Kullback–Leibler divergence as in entropy balancing), subject to feasibility and stability constraints such as lower and upper bounds $0 < L \le \omega_i \le U$ for all i. The strictly positive lower bound is the defining feature of weighting relative to selection: every observation in the pool retains a positive weight. Allowing $\omega_i = 0$ for some observations turns calibration into subset selection, which is the sparse case this paper studies (Remark 2). The estimator is

$$\hat{\mu}^W \;=\; \sum_{i\in\mathcal{D}} \omega_i\, y_i\,. \tag{9}$$

This estimator addresses between-cell distortions by construction. It also addresses within-cell distortions *to the extent that* the x-moment constraints, combined with feasible weights, can shift the effective within-cell distribution of y towards its population counterpart. When within-cell distortions are severe, cells are sparse, or weight bounds are binding, this correction may be incomplete.

Estimator S (Stratify-then-select-then-weight). For each cell c, select a subset $S_c \subset \mathcal{D}_c$ of size n_c such that the sample mean of x on the selected subset approximates the official benchmark: $\bar{x}_{S_c} \approx m_c^x$. Then estimate μ^y by the weighted sample mean over

the selected firms, with each firm in cell c receiving the cell-level weight w_c:

$$\hat{\mu}^S \;=\; \frac{\sum_{c=1}^{C}\sum_{i\in S_c} w_c\, y_i}{\sum_{c=1}^{C}\sum_{i\in S_c} w_c} \;=\; \frac{\sum_{c=1}^{C} w_c\, n_c^{\mathrm{sel}}\, \bar{y}_c^S}{\sum_{c=1}^{C} w_c\, n_c^{\mathrm{sel}}}\,, \tag{10}$$

where $\bar{y}_c^S = |S_c|^{-1}\sum_{i\in S_c} y_i$ and w_c are post-stratification weights correcting any residual discrepancy between the cell composition of the selected sample and the population:

$$w_c \;=\; \frac{\pi_c}{n_c^{\mathrm{sel}} \big/ \sum_{c'} n_{c'}^{\mathrm{sel}}}\,, \tag{11}$$

with $n_c^{\mathrm{sel}} = |S_c|$ the number of firms actually retained in cell c.

The key distinction from Estimator W is that the correction of within-cell composition happens at the *design stage*: by selecting firms whose benchmark-variable means match official targets, the algorithm removes within-cell distortion in x before any weighting takes place. The post-stratification weights that follow are therefore mild, serving only to adjust residual between-cell imbalances rather than to compensate for large within-cell distortions.

3.4 Core Proposition

The central theoretical result characterises conditions under which the design-stage correction embodied in Estimator S yields more reliable estimates than the full-pool calibration of Estimator W. We state it as a conjecture.

Proposition 1 (Conjecture)**.** *Under the following conditions:*

(i) ***Within-cell selection distortion.*** *The pool $\mathcal{D}$ is selectively distorted within cells: $E_{G_c}[x] \neq E_{F_c}[x]$ for at least some $c \in \mathcal{C}$.*

(ii) ***Benchmark availability.*** *Official benchmarks $m_c^x = \mu_c^x$ and cell shares π_c are available for those cells.*

(iii) ***Structural link between x and y.*** *Selection within cells acts through the latent index η that jointly shifts x and y. The conditional expectation $E[y \mid x, c]$ is sufficiently regular (for instance, monotone non-decreasing in x or Lipschitz continuous in x with a bounded Lipschitz constant), so that reducing the distortion in the marginal distribution of x within cell c also reduces the distortion in the conditional mean of y.*

(iv) ***Calibration instability or insufficiency.*** *At least one of the following holds:*

(a) cells are sparse: $|\mathcal{D}_c|$ is small relative to the number of moment conditions, so that the calibration problem is ill-conditioned;

(b) *the within-cell distribution G_c has heavier tails than F_c (e.g., outlier firms that inflate cell-level averages), so that unconstrained calibration weights become extreme;*

(c) *feasible weight bounds $[L, U]$ are binding, preventing the calibration weights from fully correcting within-cell distortions in x;*

(d) *a substantial fraction of firms in the pool have missing values on the benchmark variables, reducing the effective pool for calibration.*

Then the absolute bias of $\hat{\mu}^S$ for μ^y is smaller than the absolute bias of $\hat{\mu}^W$ under the same weight constraints:

$$\left| \, Bias(\hat{\mu}^S) \, \right| \; < \; \left| \, Bias(\hat{\mu}^W) \, \right|. \tag{12}$$

Remark 3. Condition (iii) is essential. If y is independent of x conditional on cell membership (that is, if the benchmark variables carry no information about the target outcome beyond what is already captured by the cell partition), then matching the moments of x within cells has no effect on the distribution of y, and neither estimator dominates the other. The condition is not directly testable, because it involves the joint population distribution of (x, y), which is exactly what the non-probability pool fails to represent. Two sufficient structures make it concrete. First, a single-index structure: suppose that, within each cell, both x and y are increasing functions of the latent index η up to idiosyncratic noise independent of selection, and that selection tilts the pool towards high-η firms as in (7). Any correction that operates by re-tilting the within-cell distribution of η towards its population counterpart in the first-order stochastic dominance sense – as a selection of firms does when it removes part of the high-η over-representation – then moves the distributions, and in particular the means, of both x and y towards their population values. Matching the cell mean of x does not mechanically equalise the mean of y unless the two index loadings are proportional, but under this structure the direction of the correction is the same for both variables, which is what condition (iii) requires. Second, a regression structure with ignorable selection given (x, c): suppose inclusion in the pool is independent of y conditional on (x, c), so that $E[y \mid x, c]$ is the same function in the pool and in the population. If $E[y \mid x, c]$ is linear in x, the bias in the mean of y equals β_c' times the bias in the mean of x, so that matching the cell means of x within the acceptance band bounds the residual bias in y by $\|\beta_c\|$ times the band width. If $E[y \mid x, c]$ is only Lipschitz in x, the bias in the mean of y is bounded by the Lipschitz constant times the Wasserstein distance between the within-cell distributions of x in the pool and in the population; correcting the mean of x alone then reduces, but need not bound, the bias in y, and the guarantee is strongest when the conditional mean is close to linear over the relevant range. Both structures are plausible in firm-level settings: when the selection process captured by η favours larger, more profitable, and more export-oriented firms, it simultaneously shifts employment-based benchmarks (components of x) and productivity-based

outcomes (typical components of y) in the same direction. In the Italian manufacturing application of this paper, the correlation between the benchmark variables (value added per employee, turnover per employee) and observable proxies of the target outcomes is strong and positive, providing indirect support for the condition.

Remark 4. Condition (iv) identifies the environments in which the two estimators genuinely differ. If the pool is large relative to the number of cells, within-cell distributions are well behaved, and calibration weights are unconstrained and stable, then Estimator W can in principle correct both between-cell and within-cell distortions through the moment constraints on x. In that setting, there may be no advantage to the design-stage selection of Estimator S, and the additional variance introduced by working with a subset rather than the full pool could even make S inferior. The case for subset selection arises specifically when full-pool calibration is constrained or unstable, and the mechanism deserves to be spelled out. Within a cell, non-negative weights that sum to one can reproduce only benchmark vectors that lie in the convex hull of the observed firm-level values of x in that cell. When $|\mathcal{D}_c|$ is small, the hull has few vertices and covers a small region: the official moment vector m_c^x is then more likely to fall outside the hull, in which case no feasible weights exist and the constraint system is infeasible, or near its boundary, in which case the solution concentrates weight on the few extreme observations and the bounds $[L, U]$ bind.[1] When $|\mathcal{D}_c|$ falls to $p + 1$, where p is the number of moment conditions, the system of p moment constraints plus the normalisation is exactly identified whenever the x_i are affinely independent: the weights are then the barycentric coordinates of m_c^x in the simplex spanned by the observed points, are feasible only if m_c^x lies inside that simplex, and respond one-for-one to any perturbation of an individual observation. With fewer than $p + 1$ firms the reproducible vectors form a lower-dimensional set and the system is generically infeasible altogether. This is the precise sense in which sparsity makes the calibration problem ill-conditioned, and it is the mechanism documented directly in the 3-digit application, where the constraint system fails in 73% of cells. The empirical application at 2-digit granularity, where the pool is large and rich benchmarks are available, is not the setting where condition (iv) binds most strongly. Consistent with this, calibration performs particularly well there on the functionals closest to its constraints, while the proposed algorithm is more accurate on aggregate functionals that neither method targets directly. The conjecture's operative content concerns accuracy on target outcomes y that are not mechanically tied to the constrained moments; establishing it requires settings where the truth for y is known, which is the role of controlled simulation rather than of field data.

[1] With bounds $0 < L \leq \omega_i \leq U$ the set of reproducible benchmark vectors is a polytope strictly interior to the convex hull, since no observation can carry all the weight. Bounded calibration therefore becomes infeasible strictly before m_c^x reaches the hull boundary; the sparsity mechanism described here is, if anything, stronger under the bounded weights used in practice.

3.5 Intuition for the Bias Reduction

The bias of any estimator of μ^y can be decomposed into two components: a *between-cell* term, arising from incorrect cell proportions, and a *within-cell* term, arising from distorted composition within cells. Formally,

$$\text{Bias}(\hat{\mu}) \;=\; \underbrace{\sum_c (\hat{\pi}_c - \pi_c)\, \mu_c^y}_{\text{between-cell}} \;+\; \underbrace{\sum_c \hat{\pi}_c\, (\hat{\mu}_c^y - \mu_c^y)}_{\text{within-cell}}, \tag{13}$$

where $\hat{\pi}_c$ and $\hat{\mu}_c^y$ denote the effective cell shares and cell means implied by the estimator.

Both Estimators W and S address the between-cell component by matching official cell shares π_c. The estimators differ in how they handle the within-cell component.

Estimator W (full-pool calibration) attempts to correct within-cell distortions through the weight vector $\{\omega_i\}$. The calibration constraints require that the weighted cell means of x match the official benchmarks. However, this approach faces three interrelated difficulties. First, when the pool within a cell is contaminated by outlier firms or exhibits heavier tails than the population distribution, the weights needed to shift the weighted mean of x towards the benchmark can be extreme: a few firms must receive very large or very small weights to offset the influence of outliers. Second, when cells are sparse (few firms in the pool for a given sector–size-class combination), the calibration problem is under-determined or ill-conditioned, and the resulting weights are unstable across replications. Third, when weight bounds are imposed to ensure stability (as they must be in practice), the constrained weights may be unable to fully correct within-cell distortions, leaving a residual within-cell bias in y that persists even after calibration.

Estimator S (subset selection) addresses the within-cell component at the design stage. By selecting, within each cell, the subset of firms whose benchmark-variable means approximate the official targets, it removes much of the within-cell composition bias *before* any weighting takes place. When condition (iii) holds and the selection distortion that shifts x also shifts y, the subset whose x-means are corrected is also the subset whose y-means are closer to the population values. The post-stratification weights applied after selection serve only to fine-tune residual between-cell imbalances and are correspondingly mild. In the Italian application, for example, 90% of post-stratification weights fall between 0.93 and 0.97, indicating that the design-stage selection has absorbed nearly all of the correction.

Remark 5. Two related diagnostics measure how evenly an estimator spreads its information across observations. The first is the effective sample size (ESS) in the sense of Kish (1965):

$$\text{ESS} \;=\; \frac{\left(\sum_i w_i\right)^2}{\sum_i w_i^2}, \tag{14}$$

computed on the firm-level weights implied by the estimator; when weights are uniform, ESS $= n$. The second, and in our setting the more informative, is the tail of the weight distribution: the maximum weight relative to the mean, and the share of weight carried by the most heavily-weighted observations. The ESS summarises average dispersion and can remain high even when a small number of observations carry extreme weights; the tail diagnostics capture the fragility channel described in condition (iv), namely that a handful of firms with extreme weights can shift the aggregate, so that any data-quality issue in those firms propagates directly to the final estimate. Estimator S, by construction, has near-uniform weights: its maximum relative weight is close to one and its ESS is close to the selected sample size. For Estimator W, the acceptance of extreme weights is the price of exact moment matching. In the empirical application (Section 4) we report both diagnostics and complement them with a direct perturbation exercise that measures how much each estimate moves when the most heavily-weighted observations are removed. These are variance and robustness diagnostics, not measures of bias; they speak to the stability of the estimate, which is the dimension along which the design-stage correction has its comparative advantage.

3.6 Granularity Trade-off

The choice of cell granularity (how finely to partition the population by sector and size class) introduces a trade-off that we state here as a conjecture to be explored through simulation.

On one hand, finer cells (larger C) reduce within-cell heterogeneity: if cells are narrowly defined, the firms within each cell are more homogeneous in both x and y, and the scope for within-cell selection distortion is smaller. In the limit, if each cell contained a single firm type, matching cell shares alone would suffice.

On the other hand, finer cells increase sparsity. As C grows, the number of pool firms per cell $|\mathcal{D}_c|$ shrinks, and in some cells the pool may contain too few firms for reliable subset selection. The variance of $\bar{y}_c^S$ rises, the stochastic search is more likely to fail to find a satisfactory subset, and the acceptance tolerance on $\bar{x}_{S_c} \approx m_c^x$ must be widened. This tension implies that, in finite samples, there exists an interior optimum: a level of granularity that balances the bias reduction from narrower cells against the variance increase from sparser pools. Sparsity also bites the two correction strategies asymmetrically: the calibration problem within a cell requires enough observations to satisfy several moment constraints with well-behaved weights, whereas the selection algorithm only requires enough observations to fill a small target with an acceptable draw. The empirical application provides direct evidence on both margins by repeating the analysis at 3-digit granularity (Section 4).

4 Empirical Application

We apply the benchmark-guided stratification algorithm to Italian manufacturing firms for the year 2022. This section describes the data sources (Section 4.1), documents the compositional distortions present in the raw Orbis pool (Section 4.2), evaluates the algorithm's ability to correct aggregate bias (Section 4.3), and examines cell-level matching quality, weight stability, and sectoral composition (Sections 4.4–4.6). Section 4.7 compares the algorithm with full-pool calibration, Section 4.8 evaluates outcomes not targeted by the algorithm, Section 4.9 repeats the comparison in the sparse-cell regime created by 3-digit stratification, and Section 4.10 concludes with a synthesis.

4.1 Data and Setting

Firm-level data. Firm-level data were obtained from Bureau van Dijk's Orbis database. The raw extract covers approximately 4.8 million firm-year observations for Italian firms over the period 2015–2024, with 21 variables including BvD identifiers, 4-digit NACE Rev. 2 industry codes, number of employees, operating revenue (turnover), value added, EBITDA, total assets, and profit/loss after tax. The extract is a single download rather than a vintage-by-vintage assembly in the sense of Kalemli-Özcan et al. (2024); part of the raw-pool discrepancies documented below may therefore reflect construction choices that their protocol is designed to avoid. This caveat qualifies the level of the distortions we document, but not their cross-cell pattern, nor the comparison between correction methods, which all operate on the same pool. We restrict the sample to manufacturing firms (NACE sections C10–C33) with at least 10 employees in the year 2022, yielding a pool of 52,650 firms. The 10-employee threshold matches the size-class detail of the benchmark extract used here and avoids the segment where Orbis coverage is known to be thinnest. No sampling design governs which firms appear in Orbis: the database aggregates mandatory filings, voluntary disclosures, and commercial data sources, with well-documented variation in coverage by country, sector, and firm size (Bajgar et al., 2020; Kalemli-Özcan et al., 2024).

Benchmark statistics. Population-level benchmarks come from the OECD Structural Business Statistics (SBS), April 2025 vintage; the 2022 figures may be subject to subsequent revision by the source. The SBS provides counts of enterprises, total employment, number of employees, aggregate turnover, and aggregate value added at factor cost, tabulated by 2-digit NACE sector and four size classes based on number of employees (10–19, 20–49, 50–249, 250 and above). For Italy in 2022, the SBS reports 67,691 manufacturing firms with 10 or more employees, distributed across 92 non-empty cells out of a potential $24 \times 4 = 96$ (four small sectors have no firms in the largest size class). These data constitute the "population structure" against which we calibrate the subset selection.

Coverage. The Orbis pool covers 77.8% of the OECD population count (52,650/67,691). Coverage varies non-monotonically by size class: roughly 70% for firms with 10–19 employees, 86% for 20–49, 92% for 50–249, and 64% for firms with 250 or more employees. The high coverage of mid-sized firms is consistent with the stylised facts documented by Bajgar et al. (2020) and reflects the filing obligations and commercial visibility of established firms. The low coverage at the top is a consequence of our cleaning protocol, which retains unconsolidated accounts: a number of very large firms file consolidated statements only and drop out of the pool. The coverage gaps mean that even if the Orbis pool were perfectly representative *within* each cell, using the raw pool to estimate unconditional aggregates would still produce bias whenever the missing firms differ systematically from those present.

Algorithm parameters. We apply the algorithm described in Section 2 with a target sample of $n = 3{,}000$ firms (approximately 4.4% of the OECD population). Target counts per cell are computed by allocating 3,000 firms proportionally to the OECD distribution, rounding up to the next integer, with a minimum of 2 firms per non-empty cell. Within each cell, candidate draws prioritise firms with more complete reporting (the quality score of Section 2), breaking ties at random; the stochastic search uses a pseudo-random seed of 20260405. The acceptance criterion requires the sample means of all benchmark variables available for the cell (value added per employee, turnover per employee, and employees per enterprise) to fall jointly within the current acceptance interval of the OECD benchmarks. The interval starts at ±10% and widens in steps (to ±20%, ±30%, ±50%, ±70%, and ±90%) over a maximum of 15,000 iterations per cell. In this application, every cell that entered the search was matched: no cell exhausted the iteration limit. Six of the 92 non-empty OECD cells, however, could not enter the search because the SBS does not report the cell-level moments needed for the acceptance check (cells in sectors C14, C18, C19, and C21, mostly in the smaller size classes). These cells are excluded from the sample, and post-stratification weights do not reallocate mass to them; we return to the implications in Section 4.6. The final selected sample contains 2,920 firms across 86 cells.

4.2 Size Class Distribution

Table 2 compares the size-class distribution across three datasets: the OECD population, the raw Orbis pool, and the stratified sample produced by the algorithm. The raw Orbis pool over-represents medium-sized firms (20–49 and 50–249 employees) at the expense of the smallest size class (10–19 employees), which accounts for 55.3% of the OECD population but only 50.0% of the Orbis pool. This 5.3 percentage-point gap reflects the well-documented under-coverage of small firms in commercial databases. At the upper end of the distribution, firms with 250 or more employees are slightly under-represented

in Orbis (1.8% versus 2.2%), likely because a small number of very large firms file in consolidated rather than unconsolidated accounts and are excluded by our cleaning protocol.

The stratified sample restores the population structure. The share of small firms (10–19 employees) rises to 55.6%, closely matching the OECD benchmark of 55.3%. The remaining size classes are similarly well aligned. These corrections are achieved entirely through the selection mechanism; no reweighting has been applied at this stage.

Table 2: Size class distribution: population, raw pool, and stratified sample

Size class (employees)	OECD population	Orbis raw	Stratified sample
10–19	55.3%	50.0%	55.6%
20–49	28.8%	32.0%	27.7%
50–249	13.7%	16.2%	14.0%
250+	2.2%	1.8%	2.7%
Total N	67,691	52,650	2,920

Notes: Percentages are shares of total firms in each dataset; columns sum to 100 up to rounding. OECD population figures are from the OECD Structural Business Statistics: Italy, manufacturing (NACE C10–C33), firms with 10 or more employees, year 2022. Orbis raw is the pool of firms extracted from Bureau van Dijk's Orbis database for the same perimeter and year. The stratified sample is selected by the benchmark-guided algorithm of Section 2 with target $n = 3{,}000$; the realised size of 2,920 firms reflects the six cells excluded for lack of benchmark moments and the cells whose pool contains fewer firms than the target.

4.3 Estimands and Aggregate Estimates

Before comparing estimates, the estimands must be defined precisely, because the SBS aggregates and simple firm-level means are different population objects. We work with two estimands throughout. **Estimand A** is the national ratio of totals: aggregate turnover (or value added) divided by aggregate employment. This is an employment-weighted object, dominated by large firms, and it is directly published by the SBS. **Estimand B** is the firm-count-weighted average of cell-level ratios: within each cell, the ratio of totals; across cells, the average weighted by the number of firms. This object weights each cell by how many firms it contains, so it reflects the productivity of the typical firm environment rather than of the typical euro of activity. Both are legitimate targets; they differ substantially (for TOPE, the SBS values are 368,294 for A and 258,044 for B), and conflating them, or comparing either with an unweighted mean of firm-level ratios, mixes selection bias with a functional mismatch.[2]

Table 3 reports percentage deviations from the SBS benchmarks for both estimands, computed with the correct sample analogue in every cell. For the proposed algorithm

[2]The unweighted firm-level mean of TOPE in the pool is 297,435, which deviates from the estimand-B benchmark by +15.3%; the correct estimand-B sample analogue (cell-level ratios of totals, aggregated with population cell shares) deviates by +14.0%. The functional mismatch thus accounts for 1.2 percentage points, and the remainder is composition.

we report the mean and standard deviation across 20 independent runs of the stochastic search, so that the figures reflect the estimator rather than a single draw; the reference realisation (seed 20260405) used for the cell-level diagnostics below lies within one standard deviation of the multi-seed means on every estimand.

Table 3: Deviations from SBS benchmarks, by estimand and estimator

	Estimand B (cell-weighted)		Estimand A (aggregate ratio)	
Estimator	VAPE	TOPE	VAPE	TOPE
R (Raw pool)	+2.1%	+14.0%	+1.2%	+6.8%
PS (Post-stratification)	+2.1%	+14.0%	+1.1%	+6.1%
W (Calibration)	+1.3%	+1.8%	+3.1%	+5.0%
S (Proposed, mean of 20 seeds)	−0.3%	+4.9%	+1.0%	+2.0%
(s.d. across seeds)	(0.8)	(0.9)	(2.4)	(2.4)

Notes: Entries are percentage deviations from the OECD SBS benchmarks (Italy, manufacturing, 10+ employees, 2022). Benchmark levels: estimand B, VAPE 69,851 and TOPE 258,044; estimand A, VAPE 90,728 and TOPE 368,294 (EUR). VAPE = value added per employee; TOPE = turnover per employee. All sample analogues are computed as ratios of (weighted) totals, within cells for estimand B and nationally for estimand A. R and PS coincide on estimand B because post-stratification changes cell weights but not within-cell composition, and estimand B fixes cell weights at their population values by construction. S rows are the mean and standard deviation of the deviations across 20 independent runs of the stochastic search.

Four findings emerge. First, the raw pool's distortion is economically material on both estimands: +14.0% on cell-weighted TOPE and +6.8% on the aggregate ratio. The distortion is larger for estimand B because the object gives most weight to the small-firm cells where Orbis coverage is most selective, and larger for TOPE than for VAPE because turnover scales more steeply with the characteristics that drive inclusion. Second, post-stratification alone accomplishes essentially nothing here: it leaves estimand B unchanged by construction (it corrects cell shares, which estimand B already fixes) and moves estimand A by less than one point. The problem is within-cell composition, exactly the margin that between-cell reweighting cannot touch. Third, the proposed algorithm corrects most of the distortion on both estimands: to +4.9% on cell-weighted TOPE and +2.0% on the aggregate ratio, with VAPE essentially unbiased. Fourth, full-pool calibration nearly eliminates the deviation on estimand B (+1.8%), which is unsurprising because estimand B is built from the same cell-level moments that calibration constrains; on estimand A, a functional neither method targets directly, calibration performs worse than the proposed algorithm (+5.0% versus +2.0% on TOPE, +3.1% versus +1.0% on VAPE). We return to this pattern in Section 4.7.

4.4 Cell-Level Matching Quality

The aggregate improvements reported in Table 3 could in principle mask substantial heterogeneity at the cell level. If matching quality is concentrated in a few large cells while small cells are poorly approximated, the aggregate bias reduction might not survive disaggregation. Table 4 addresses this concern by reporting the fraction of cells matched within progressively wider tolerance bands, for each of the three benchmark variables.

The evaluation covers the 86 cells in which firms were selected. At the ±10% tolerance, 94% of cells are matched on VAPE, 94% on TOPE, and 91% on employees per enterprise. At ±20%, full coverage is reached on VAPE and the matching rate is 98% on TOPE. All three variables are targeted jointly by the acceptance criterion, so these figures measure how tightly the accepted draws sit around the benchmarks after the progressive widening of the tolerance.

The iteration count at which each cell was accepted provides the complementary design-side diagnostic: 77 of the 86 cells were accepted at the initial ±10% band, 6 at ±20%, 1 at ±30%, and 2 at ±50%. The two hardest cells are tobacco (C12, size class 20–49), where the pool contains a single eligible firm against a target of two, and the smallest size class of wearing apparel (C14), where the within-cell distribution of the pool is most distorted; no cell required the ±70% or ±90% bands. The acceptance band should not be read as a measure of achieved quality in large cells: the two largest cells (fabricated metals and food, smallest size class) were accepted at iterations 4,001 and 4,050, immediately after the first widening, because the means of large draws barely vary across iterations, so acceptance at ±10% requires the pool itself to sit within ±10%; their achieved cell means nonetheless lie within 5 to 20% of the benchmarks. A sensitivity check based on achieved quality is more informative: restricting the cell-weighted estimand to the 83 cells whose achieved cell means all lie within ±20% of the benchmarks (dropping 4.3% of population mass, including the wearing-apparel cell) moves the deviations from −0.6% to +1.3% on VAPE and from +5.0% to +6.9% on TOPE; movements of this size are comparable to two seed-level standard deviations and are driven mechanically by the reallocation of the dropped cells' population shares.

Table 5 provides additional detail on the distribution of cell-level ratios (sample mean divided by OECD benchmark) across the 86 evaluated cells. For VAPE, the median ratio is 0.965, with the 10th and 90th percentiles at 0.908 and 1.070, respectively: the distribution is tight and centred just below unity. TOPE ratios are similarly concentrated, with a median of 1.007 and 10th-to-90th percentile range of $[0.911, 1.088]$. The employees-per-enterprise ratio has a median of 1.032 and a slightly wider spread, reflecting the fact that firm counts are constrained to integers and small cells are sensitive to the inclusion or exclusion of individual firms.

Table 4: Cell-level matching quality: fraction of cells within tolerance

Tolerance band	VAPE		TOPE		Employees/enterprise	
±10%	81/86	(94%)	81/86	(94%)	78/86	(91%)
±20%	86/86	(100%)	84/86	(98%)	84/86	(98%)
±30%	86/86	(100%)	85/86	(99%)	85/86	(99%)

Notes: Each entry reports the number (and percentage) of cells of the reference realisation for which the ratio of the selected-sample cell mean to the OECD benchmark falls within the indicated tolerance band. Evaluation covers the 86 cells with selected firms; the remaining 6 non-empty OECD cells lack the cell-level moments required by the acceptance check and receive no firms (see Section 4.1). VAPE = value added per employee; TOPE = turnover per employee.

Table 5: Distribution of cell-level ratios (sample mean / OECD benchmark)

Variable	Mean	Median	P10	P90	Min	Max
VAPE	0.981	0.965	0.908	1.070	0.841	1.178
TOPE	1.006	1.007	0.911	1.088	0.760	1.490
Employees/ent.	1.025	1.032	0.920	1.094	0.744	1.485

Notes: Distribution, across the 86 cells of the reference realisation, of the ratio of the selected-sample cell mean to the OECD benchmark; a ratio of 1.000 indicates exact replication of the benchmark. P10 and P90 denote the 10th and 90th percentiles. VAPE = value added per employee; TOPE = turnover per employee; Employees/ent. = employees per enterprise.

4.5 Weight Distribution

A well-functioning stratification algorithm should produce a sample whose composition is close enough to the population structure that post-stratification weights are mild. Extreme weights are a warning sign: they indicate that the design-stage selection failed to achieve compositional balance, leaving the weighting step to compensate for large discrepancies. In calibration approaches applied to the full pool, extreme weights are common and motivate the bounded and trimmed variants studied by Deville et al. (1993) and Kott (2006).

In the stratified sample, post-stratification weights have a mean of 0.96 and a median of 0.97, with a standard deviation of 0.046 (equivalently 0.048 when weights are normalised by their mean, the convention used in the comparison table below). The distribution is concentrated near unity: the 5th percentile is 0.93, the 25th percentile is 0.96, and the 75th through 99th percentiles are all 0.97. The left tail is short but present: the 1st percentile is 0.83 and the overall minimum is 0.13, corresponding to a handful of firms in very small cells where the minimum-2-firms rule forces oversampling relative to the cell's population share. Skewness is -11.76 and kurtosis is 172.27, driven by these few cells. The key substantive point is that the weights are close to 1, confirming that the design-stage selection, rather than aggressive post-hoc reweighting, is responsible for the bias corrections documented in Table 3. This is a direct consequence of the benchmark-

guided construction: by selecting firms in proportions that approximate the population cell structure, the algorithm ensures that the remaining weight adjustment is a fine-tuning exercise rather than a structural correction.

4.6 Sector-Level Patterns

The algorithm also corrects sectoral composition, although it targets size-class-by-sector cells rather than sector margins directly. In the raw Orbis pool, certain sectors diverge substantially from their OECD population shares. Wearing apparel (C14) accounts for 6.25% of the population but only 3.96% of the Orbis pool, an under-representation of 2.3 percentage points that reflects the predominance of very small firms in this sector. Fabricated metal products (C25), the largest manufacturing sector, is over-represented by 1.8 percentage points (23.58% versus 21.81%), consistent with the sector's relatively high share of medium-sized firms that are well covered by Orbis. Food products (C10) shows a similar under-representation pattern (8.96% versus 10.57%), again driven by the small-firm segment.

The stratified sample partially corrects these distortions. The share of wearing apparel rises from 3.96% to 4.42%, reducing the gap from 2.3 to 1.8 percentage points. Fabricated metals falls from 23.58% to 22.50%, closer to the population share of 21.81%. Food products increases from 8.96% to 10.92%, nearly matching the population benchmark. One sector deteriorates rather than improves: printing and reproduction of recorded media (C18) accounts for 2.22% of the population and 2.10% of the Orbis pool, but only 0.27% of the stratified sample (8 of the 2,920 selected firms). This is a direct consequence of the excluded cells documented above: the two smallest size classes of C18, which contain most of the sector's firms, lack the SBS cell-level moments required by the acceptance check and therefore receive no selected firms. This is a genuine limitation of the algorithm in its current form: when official moments are missing for a cell, the cell drops out of the sample entirely, and the post-stratification weights do not reallocate mass to it. A natural refinement, left to future work, is a fallback rule for such cells (for instance, random selection within the cell with cell-share weighting only). The remaining residual discrepancies reflect the fact that the algorithm targets cells rather than sector margins.

4.7 Comparison with Full-Pool Calibration

To evaluate the proposed algorithm against the natural alternative, we implement bounded calibration on the full Orbis pool, in the spirit of Deville and Särndal (1992). For each cell, starting from uniform weights, we iteratively apply linearised calibration updates so that the weighted cell means match the cell-level OECD benchmarks for turnover per employee, value added per employee, and employees per enterprise simultaneously; at each step, multiplicative adjustments are damped and weights are kept within bounds

of $[0.05/n_c,\ 20/n_c]$, where n_c is the cell pool size, to prevent degeneracy. The calibrated cell-level means are then aggregated using OECD population cell shares. This is a cell-by-cell implementation, chosen to mirror the cell-by-cell structure of the proposed algorithm; sample-level calibration with cross-cell constraints, as well as entropy balancing (Hainmueller, 2012), are natural additional comparators beyond the scope of this paper.

Table 6 reports the deviations of the three estimators defined in Section 3, together with the post-stratification-only reference (PS), on both estimands of Section 4.3, alongside the weight properties of each estimator. To make weight statistics comparable across estimators operating on different samples, all weights are normalised by their mean.

Table 6: Comparison of estimators: deviations by estimand and weight properties

	Estimand B		Estimand A		Weight properties		
Method	VAPE	TOPE	VAPE	TOPE	Std	Range	ESS
R (Raw pool)	+2.1%	+14.0%	+1.2%	+6.8%		(unweighted)	
PS (Post-strat.)	+2.1%	+14.0%	+1.1%	+6.1%	0.21	[0.50, 3.3]	50,382
W (Calibration)	+1.3%	+1.8%	+3.1%	+5.0%	0.51	[0.03, 11.9]	41,693
S (Proposed, mean of 20 seeds)	−0.3%	+4.9%	+1.0%	+2.0%	0.048	[0.13, 1.01]	2,913

Notes: Entries in the first four columns are percentage deviations from the SBS benchmarks of Table 3, computed with the correct sample analogue of each estimand. Post-stratification weights correct for cell shares only. Calibration implements bounded iterative calibration on the full pool (52,650 firms), matching cell-level means of the three benchmark variables simultaneously. Proposed = benchmark-guided subset selection (2,920 firms); its deviations are means across 20 seeds of the stochastic search, and its weight statistics refer to the reference realisation. Weight statistics (Std, Range) refer to firm-level weights normalised by their mean, so that a value of 1 is the average weight. ESS = effective sample size (Kish, 1965), computed as $(\sum w_i)^2 / \sum w_i^2$ on the firm-level weights implied by each estimator; values close to the sample size indicate near-uniform weighting.

The deviation columns of Table 6 show that neither weighting-based nor selection-based correction dominates across estimands. Calibration nearly eliminates the deviation on estimand B, which is expected rather than remarkable: estimand B is constructed from the same cell-level moments that the calibration constraints target, so accuracy there is close to mechanical. On estimand A, a functional that neither method constrains, the ranking reverses: the proposed algorithm deviates by +2.0% on TOPE and +1.0% on VAPE against calibration's +5.0% and +3.1%. Accuracy on the constrained moments, in other words, does not automatically carry over to other functionals of the same data, which is the empirical counterpart of the distinction between benchmark variables and target objects drawn in Section 3. The weight columns add a second, descriptive contrast: under calibration, individual firms carry up to 11.9 times the mean weight, and 5% of firms carry more than 1.9 times the mean; under the proposed algorithm, the maximum relative weight is 1.01, so the contribution of every firm to the estimate is transparent and essentially uniform.

Stability under perturbation and across seeds. A natural concern is that either estimator might depend on a small set of observations. We assess this at the estimator level: we perturb the pool, re-estimate each procedure from scratch (re-solving the calibration, re-running the stochastic search), and measure the movement of the estimates. Three exercises are informative. First, removing the top 0.5% of firms by relative calibration weight (264 firms) and re-calibrating moves the calibrated estimates by at most half a percentage point; without re-calibration the same removal moves the calibrated VAPE deviation by 1.9 points, but that exercise breaks the calibration constraints mechanically and measures exposure, not estimator sensitivity. Second, removing the top 0.5% of firms by turnover per employee, an extreme-value perturbation that is symmetric across methods, moves the re-calibrated estimates by at most 0.3 points and the mean re-selected estimates of the proposed algorithm by at most 0.5 points. Third, across 20 independent seeds of the stochastic search on the unperturbed pool, the proposed algorithm's deviations have standard deviations of 0.8 to 0.9 points on estimand B, so that any single run lies within roughly ± 2 points of the estimator's central tendency: the acceptance bands cap how far any accepted draw can wander from the benchmarks. Both estimators, properly re-estimated, are therefore stable; the naive leave-out exercise that does not re-estimate overstates the fragility of calibration, and a single-seed comparison understates the sampling variability of the proposed algorithm. A final exercise addresses the quality score of the algorithm's first step: neutralising the score, so that candidate draws are purely random within cells, changes the multi-seed mean deviations by about one percentage point, and not in a common direction (cell-weighted TOPE improves to $+3.6\%$ across five seeds against $+4.9\%$ with the score, while VAPE moves to -1.2% against -0.3%). The corrections relative to the raw pool are therefore not generated by the score-based prioritisation, which if anything trades a point of accuracy on one benchmark variable against the other; the corrections are disciplined by the acceptance criterion, and the score only tilts the selection towards firms with more complete balance sheets. The definitive comparison of accuracy on outcomes not tied to the benchmarks requires a setting where the truth is known, which is the role of the planned Monte Carlo study.

4.8 Outcomes Not Targeted by the Algorithm

The results so far evaluate the estimators on the same variables that enter the acceptance criterion. A more demanding test, and the empirically relevant one for condition (iii) of Proposition 1, is whether the correction propagates to outcomes that the algorithm never sees. The Orbis data include three balance-sheet variables that play no role in the selection: EBITDA, total assets, and profit or loss after tax. We express each in per-employee terms and winsorise at the 1st and 99th percentiles of the pool distribution to

limit the influence of extreme accounting values.

Two facts emerge. First, the firm-level correlations between the benchmark variables and these outcomes are strong: in the pool, the correlation of value added per employee with EBITDA per employee is 0.80, with profit per employee 0.85, and with assets per employee 0.66. This provides direct empirical support for the structural link postulated in condition (iii): the selection forces that shift the benchmark variables also shift these outcomes.

Second, correcting the sample composition materially changes the estimates of the untargeted outcomes, and the correction operates in the direction the theory predicts. No official benchmark exists for these variables, so bias cannot be computed; but since the raw pool overstates the targeted variables (by 14% on cell-weighted turnover per employee), and the untargeted outcomes are strongly positively correlated with them, the raw pool should overstate the untargeted outcomes as well, and composition corrections should pull the estimates down. They do: relative to the raw pool, the proposed algorithm lowers estimated EBITDA per employee by 7.0%, assets per employee by 4.6%, and profit per employee by 9.1%. Post-stratification alone moves the same estimates by 2.4 to 4.1%, and full-pool calibration by 3.5 to 7.6%. The choice of correction method therefore matters materially for downstream variables that researchers routinely study, not only for the variables used in the correction itself.

4.9 The Sparse-Cell Regime: Three-Digit Granularity

The theoretical case for design-stage selection (condition (iv) of Proposition 1) concerns settings where cells are sparse and the calibration problem is ill-conditioned. Italian manufacturing at 2-digit granularity is not such a setting: the pool is large, coverage is 78%, and most cells contain hundreds or thousands of firms. The same data, however, allow us to create the sparse regime directly, by refining the stratification from 2-digit to 3-digit NACE sectors. The SBS download reports complete benchmark moments for 229 3-digit cells, covering 64,475 firms (95% of the 2-digit population); the Orbis pool populates 227 of them, with a median of 61 firms per cell, a 25th percentile of 16, and 37% of cells below 30 firms. This is the environment the theory is about.

Table 7 reports the same comparison as Table 6 at 3-digit granularity, with estimands defined analogously on the 3-digit cell structure. Three results stand out.

First, the raw distortion grows with granularity, from 14.0% to 16.7% on cell-weighted TOPE: finer benchmarks reveal more of the within-cell selection that coarse cells average away. Post-stratification, as before, changes nothing.

Second, the mechanics of the two corrections diverge sharply, in the direction the theory predicts. The calibration constraint system becomes infeasible or fails to converge in 165 of the 227 cells (73%), and the weights that do converge push against the imposed

Table 7: The sparse-cell regime: deviations at 3-digit granularity

	Estimand B (3-digit cells)		Estimand A (aggregate)	
Estimator	VAPE	TOPE	VAPE	TOPE
R (Raw pool)	+3.5%	+16.7%	−3.1%	+0.9%
PS (Post-stratification)	+3.5%	+16.7%	−1.8%	+0.3%
W (Calibration)	+2.5%	+2.6%	−3.0%	−5.2%
S (Proposed, mean of 10 seeds)	−0.3%	+5.8%	−8.7%	−9.1%
(s.d. across seeds)	(0.7)	(0.7)	(0.8)	(0.8)

Notes: Entries are percentage deviations from the 3-digit SBS benchmarks, defined as in Table 3 but on the 3-digit cell structure. Benchmark levels: estimand B, VAPE 68,426 and TOPE 249,632; estimand A, VAPE 92,690 and TOPE 389,364 (EUR). The SBS reports complete benchmark moments for 229 3-digit cells; the Orbis pool populates 227 of them. Calibration diagnostics: the constraint system fails to converge, or the cell is too small to calibrate, in 165 of 227 cells (73%); the maximum weight relative to uniform is 20.8, at the imposed bound. Selection diagnostics: 1 cell unmatched per run; mean selected sample of 3,089 firms; deviations are means and standard deviations across 10 seeds.

bounds. The selection algorithm, by contrast, fails to match only one cell per run, and its deviations remain tightly bounded across seeds (standard deviations of 0.7 to 0.8 points). The point estimates tell a more nuanced story: calibration's aggregate deviation on estimand B remains low (+2.6%) despite the widespread constraint failures, because even the non-converged weights move cell means towards the benchmarks; the selection algorithm delivers +5.8%, essentially unchanged from the 2-digit regime. Operationally, however, only one of the two procedures still works as designed: a researcher relying on the calibration output at this granularity would be using weights that violate the stated constraints in three cells out of four, a failure that is invisible unless convergence is checked cell by cell.

Third, both methods deteriorate on the aggregate estimand A, and the selection algorithm deteriorates more (−9.1% against −5.2% on TOPE). The reason is structural and well understood in business-survey design: estimand A is dominated by the largest firms, and a sample of roughly 3,000 firms allocated proportionally across 227 cells selects only a handful of firms in each open-ended 250+ class, missing the extreme upper tail that drives employment-weighted totals. The standard remedy is a take-all stratum, including with certainty all firms above a size threshold, which is a natural refinement of the algorithm and part of the planned extensions. The scope implication is worth stating plainly: fixed-size samples without a certainty stratum, however well composed, are not suited to tail-dominated aggregates.

Taken together, the two granularity levels provide the empirical counterpart of the granularity trade-off conjectured in Section 3: finer cells expose more selection bias and make the benchmarks more informative, but they degrade the feasibility of weighting-based corrections much faster than they degrade design-based selection, while tail-dominated

functionals eventually require explicit tail coverage under either approach.

4.10 Discussion

The empirical application supports five conclusions. First, compositional distortions in the raw Orbis pool are real and economically material once the estimand is defined precisely: 14% on cell-weighted turnover per employee at 2-digit granularity, rising to 16.7% at 3-digit, and the distortion propagates, through the strong correlation between benchmark variables and balance-sheet outcomes, to variables well beyond those used in the correction. Second, between-cell reweighting alone does not address the problem: post-stratification leaves the cell-weighted estimand unchanged by construction, confirming that the operative distortion is within-cell composition. Third, the benchmark-guided algorithm corrects most of the distortion on both estimands (to 4.9% and 2.0% on turnover per employee at 2-digit granularity, with value added essentially unbiased), matches over 90% of cells within $\pm 10\%$, and produces near-uniform weights, so that the contribution of every firm to the estimates is transparent. Fourth, at coarse granularity the comparison with full-pool calibration shows specialisation rather than dominance: calibration is nearly exact on the functional built from the moments it constrains, the proposed algorithm is more accurate on the aggregate functional that neither method targets, and both estimators are stable under perturbation once properly re-estimated. Fifth, in the sparse-cell regime the operational asymmetry predicted by the theory emerges: the calibration constraint system fails in 73% of 3-digit cells while the selection algorithm continues to deliver bounded, stable estimates, although tail-dominated aggregates then require a take-all stratum under either approach. Which correction to prefer depends on the estimand the researcher needs and on the granularity of the benchmarks, which reinforces the paper's central message that the estimand must be fixed before the correction is chosen.

The limitations of the analysis, and the extensions they call for, are collected in the concluding section.

5 Conclusion

This paper studies when benchmark-guided stratified subset selection from a non-probability firm database improves the estimation of unconditional population objects relative to feasible weighting-only corrections. We describe the algorithm, which selects a subset of firms from a commercial database so that the resulting sample matches the cell structure of official statistics across a sector-by-size-class cross-classification. We outline the theoretical conditions under which this design-stage intervention improves upon full-pool calibration: within-cell distributional distortions, sparse or near-empty cells in the cross-classification, heavy-tailed firm-level variables, and a latent selection index that links

database inclusion to the target outcomes conditional on the benchmark stratifying variables. We apply the algorithm to Italian manufacturing firms drawn from Orbis for the year 2022, using OECD Structural Business Statistics as the external benchmark, taking care to define the estimands precisely before comparing estimators. The raw Orbis pool overstates cell-weighted turnover per employee by 14% and the aggregate ratio of totals by 6.8%; post-stratification leaves the first distortion unchanged by construction; the proposed algorithm reduces the two to 4.9% and 2.0% (means across 20 seeds of the stochastic search), with over 90% of cells matched within ±10% and near-uniform weights. The comparison with full-pool calibration reveals specialisation rather than dominance at coarse granularity: calibration is nearly exact on the functional built from the moments it constrains, while the proposed algorithm is more accurate on the aggregate functional that neither method targets, and both estimators are stable under pool perturbations once properly re-estimated. Repeating the analysis at 3-digit granularity, where the median cell contains 61 pool firms, reveals the operational asymmetry the theory predicts: the calibration constraint system fails in 73% of cells, while the selection algorithm continues to deliver stable, bounded estimates, failing in only one cell per run. The correction also propagates to outcomes not used by the algorithm, moving estimates of EBITDA, assets, and profits per employee by 5 to 9%.

Several limitations should be acknowledged. First, the empirical application covers only Italian manufacturing firms with ten or more employees, in a single year; extensions to other countries, years, and service sectors are needed to assess external validity. Second, the calibration comparison uses a cell-by-cell bounded implementation chosen to mirror the structure of the proposed algorithm; sample-level calibration and entropy balancing (Hainmueller, 2012), together with trimmed calibration baselines (Deville et al., 1993), would further clarify the boundary conditions. Third, the core theoretical result is stated as a conjecture rather than a theorem. Fourth, the algorithm cannot recover firms that are entirely absent from the commercial database, and cells for which official benchmark moments are unavailable currently drop out of the sample entirely, as the application documents for the printing sector; a fallback rule for such cells is a needed refinement. Fifth, we do not yet provide Monte Carlo evidence on the boundary conditions under which the proposed algorithm dominates strong calibration competitors, nor on the sensitivity of relative performance to the degree of cell sparsity, tail heaviness, and within-cell selection intensity.

Future research will address these gaps on several fronts. First, a Monte Carlo study will compare the estimation strategies (raw pool, post-stratification, standard and bounded calibration, entropy balancing, and the proposed algorithm) on target outcomes whose true values are known by construction, across controlled variation in cell sparsity, tail heaviness, within-cell selection intensity, and the correlation between benchmark variables and target outcomes; this is the setting the field data cannot provide, and the

one where the accuracy comparison can be settled. Second, the core proposition will be formalised with a complete proof, including explicit sufficient conditions. Third, the empirical application will be extended to Germany, France, and Spain. Fourth, two refinements suggested directly by the empirical results will be incorporated into the algorithm: a take-all stratum for the largest firms, to serve tail-dominated aggregates, and a fallback rule for cells without usable benchmark moments.

Despite these limitations, the evidence suggests that benchmark-guided subset selection is a promising complement to existing weighting approaches for researchers working with non-probability firm databases, particularly when unconditional population objects are of primary interest and when stability of the estimates across database vintages and data-quality perturbations matters.

References

Bajgar, M., Berlingieri, G., Calligaris, S., Criscuolo, C., and Timmis, J. (2020). Coverage and representativeness of Orbis data. Science, Technology and Industry Working Paper 2020/06, OECD.

Chen, R., Chen, G., and Yu, M. (2023). Entropy balancing for causal generalization with target sample summary information. *Biometrics*, 79(4):3179–3190.

Chen, Y., Li, P., and Wu, C. (2020). Doubly robust inference with nonprobability survey samples. *Journal of the American Statistical Association*, 115(532):2011–2021.

De Loecker, J., Eeckhout, J., and Unger, G. (2020). The rise of market power and the macroeconomic implications. *Quarterly Journal of Economics*, 135(2):561–644.

Deville, J.-C. and Särndal, C.-E. (1992). Calibration estimators in survey sampling. *Journal of the American Statistical Association*, 87(418):376–382.

Deville, J.-C., Särndal, C.-E., and Sautory, O. (1993). Generalized raking procedures in survey sampling. *Journal of the American Statistical Association*, 88(423):1013–1020.

Deville, J.-C. and Tillé, Y. (2004). Efficient balanced sampling: The cube method. *Biometrika*, 91(4):893–912.

Elliott, M. R. and Valliant, R. (2017). Inference for nonprobability samples. *Statistical Science*, 32(2):249–264.

Hainmueller, J. (2012). Entropy balancing for causal effects: A multivariate reweighting method to produce balanced samples in observational studies. *Political Analysis*, 20(1):25–46.

Hellerstein, J. K. and Imbens, G. W. (1999). Imposing moment restrictions from auxiliary data by weighting. *Review of Economics and Statistics*, 81(1):1–14.

Kalemli-Özcan, Ş., Sørensen, B. E., Villegas-Sanchez, C., Volosovych, V., and Yeşiltaş, S. (2024). How to construct nationally representative firm-level data from the Orbis global database: New facts on SMEs and aggregate implications for industry concentration. *American Economic Journal: Macroeconomics*, 16(2):353–374.

Kim, J. K., Park, S., Chen, Y., and Wu, C. (2021). Combining non-probability and probability survey samples through mass imputation. *Journal of the Royal Statistical Society: Series A*, 184(3):941–963.

Kish, L. (1965). *Survey Sampling*. John Wiley & Sons, New York.

Kott, P. S. (2006). Using calibration weighting to adjust for nonresponse and coverage errors. *Survey Methodology*, 32(2):133–142.

Meng, X.-L. (2018). Statistical paradises and paradoxes in big data (i): Law of large populations, big data paradox, and the 2016 US presidential election. *Annals of Applied Statistics*, 12(2):685–726.

Neyman, J. (1934). On the two different aspects of the representative method: The method of stratified sampling and the method of purposive selection. *Journal of the Royal Statistical Society*, 97(4):558–625.

Rivers, D. (2007). Sampling for web surveys. In *Proceedings of the Survey Research Methods Section, Joint Statistical Meetings*. American Statistical Association.

Särndal, C.-E., Swensson, B., and Wretman, J. (1992). *Model Assisted Survey Sampling*. Springer.

Smith, T. M. F. (1983). On the validity of inferences from non-random samples. *Journal of the Royal Statistical Society: Series A*, 146(4):394–403.

Théberge, A. (1999). Extensions of calibration estimators in survey sampling. *Journal of the American Statistical Association*, 94(446):635–644.

Valliant, R. (2020). Comparing alternatives for estimation from nonprobability samples. *Journal of Survey Statistics and Methodology*, 8(2):231–263.

Wu, C. (2022). Statistical inference with non-probability survey samples. *Survey Methodology*, 48(2):283–311.

Alma Mater Studiorum - Università di Bologna

DEPARTMENT OF ECONOMICS

Strada Maggiore 45
40125 Bologna - Italy
Tel. +39 051 2092604
Fax +39 051 2092664
http://www.dse.unibo.it