Autostrata

Improved Automatic Stratification for Coarsened Exact Matching

Jo Inge Arnes[1], Alexander Hapfelmeier[2], and Alexander Horsch[3]
[1]UiT The Arctic University of Norway, Department of Computer Science, Tromsø, Norway, jo.i.arnes@uit.no
[2]Technical University of Munich, Institute of AI and Informatics in Medicine, München, Germany
[3]UiT The Arctic University of Norway, Department of Computer Science, Tromsø, Norway

Abstract
We commonly adjust for confounding factors in analytical observational epidemiology to reduce biases that distort the results. Stratification and matching are standard methods for reducing confounder bias. Coarsened exact matching (CEM) is a recent method using stratification to coarsen variables into categorical variables to enable exact matching of exposed and nonexposed subjects. CEM's standard approach to stratifying variables is histogram binning. However, histogram binning creates strata of uniform widths and does not distinguish between exposed and nonexposed. We present Autostrata, a novel algorithmic approach to stratification producing improved results in CEM and providing more control to the researcher.

Keywords
Analytic epidemiology, confounder bias, stratification, coarsened exact matching, algorithms

1 INTRODUCTION

Epidemiologists conduct analytical observational studies [1] to investigate associations between exposures and outcomes. Instead of assigning a treatment or exposure to the participants of a randomized experiment [2], we rely on observations of the subjects in their usual environment with minimal interference. There are many established ways of designing observational studies, from cross-sectional, cohort, and case-control studies to more complex prospective cohorts with several nested case-control and cross-sectional designs [3, 4, 5]. A common theme for these is awareness of biases. Confounding factors [6, 7, 8] are a common source of bias that can, if measured, be adjusted for in the analysis [9, p. 1020]. Stratification [10], for example, can control for confounding by dividing study subjects into groups based on observed confounders. Iacus et al. [11] present the *coarsened exact matching* (CEM) method that adjusts for bias by turning confounder covariates into categorical variables through stratification, which we can then use to match comparable subjects exactly. Blackwell et al. [12] introduce a Stata (https://www.stata.com) implementation of CEM, and Iacus et al. [13] provide an implementation for R (https://www.r-project.org). In addition, a web page with an overview of implementations for other programming languages and platforms is available (https://gking.harvard.edu/cem). The same webpage also informs that CEM is officially qualified for scientific use by the U.S. Food and Drug Administration. The CEM implementations let users create strata manually or use automatic stratification. The built-in automatic stratification creates uniform width bins by applying general rules of thumb for constructing histograms. The three binning algorithms included in both Stata and R are Sturges' rule [14], Scott's rule [15], and Freedman-Diaconis' rule [16]. Additionally, Stata includes an implementation of Shimazaki-Shinomoto's rule [17].

Blackwell et al. [12, p. 534] demonstrate that manually defining strata based on domain knowledge can sometimes give better results than the current automatic approach. In their example, the manually defined strata are less imbalanced while giving a higher number of matched units. However, according to King et al. [18, p. 439], researcher biases are highly likely to affect qualitative choices even when researchers attempt to avoid them. 'The literature makes clear that the way to avoid these biases is to remove researcher discretion as much as possible,' following King. On the other hand, the general histogram binning rules do not support the specific challenges of stratifying confounders:

- The histogram binning algorithms do not distinguish between different groups of units and include no concept of matching.
- They do not take into account multivariate imbalance between groups.
- The strata have uniform widths, i.e., all strata for a covariate have the same width.
- The researcher cannot in advance give parameters to influence the stratification process.

Against this background, we researched and developed a novel algorithmic approach to the stratification problem that addresses the shortcomings above. We implemented the algorithm and experimentally compared it to CEM's built-in histogram binning with good results.

We conclude the introduction with a brief example of Autostrata's applicability to health-related studies. For instance, say we want to study if coffee consumption is associated with a beneficial effect on the risk of liver cancer. In the respective observational study, we must be cautious of possible systematic differences between the compared groups, such as smoking habits. Failing to adjust for these differences can challenge the validity of the results. Autostrata improves such adjustments when using

The 18th Scandinavian Conference on Health Informatics, Tromsø, Norway, August 22-24, 2022. Organized by UiT The Arctic University of Norway. Conference Proceedings published by Linköping University Electronic Press at https://doi.org/10.3384/ecp187. © The Author(s). This work is licensed under the Creative Commons Attribution-NonCommercial 4.0 International License. To view a copy of this license, visit http://creativecommons.org/licenses/by-nc/4.0/

CEM. The method creates more precise results and keeps more study participants included in the analysis.

After the introduction, the structure of the paper is as follows: First, we provide essential background for understanding the problem. Next, we describe our approach and algorithm. We then present the experiments and results, followed by a discussion. Last, we briefly touch upon related work before concluding.

1.1 A note on terminology

The paper primarily uses the terms *treated and controls* instead of *exposed and nonexposed* due to their use in CEM and the general causal inference literature. In addition, although case-control studies are different from cohort studies that focus on exposed and unexposed, they are interchangeable in this paper because we concentrate on the stratification of confounder covariates in isolation from these differences. Further, we use the general statistical term units instead of subjects, individuals, or study participants often seen in epidemiology.

2 BACKGROUND

Before presenting our approach, we provide the background necessary to understand the challenges of the stratification problem.

2.1 Confounding

We often assess whether the risk of a health event (outcome) is increased or decreased among an exposed or treated group compared to a control group. To quantify the relationship between an exposure or treatment and the outcome, we calculate risk ratios, odds ratios, or other measures. However, other factors not directly under investigation can skew the results or even lead us to the opposite conclusion of what is correct. Figure 1 illustrates how confounding factors influence both the exposure and the outcome. Note that the confounder is not in the direct causal pathway between the two. Also, a relevant property of confounders is that the compared groups have differently distributed values for the confounder covariate. If the confounders are measured and included in the dataset, we can adjust for confounders during analysis, which is the purpose of the stratification discussed in this paper. It is worth noting that according to Wacholder et al. [9, p. 1020], the use of stratification or matching can, in effect, adjust for unknown or unmeasured confounders through reduced variability because this variability is measured conditionally on the levels of other studied variables.

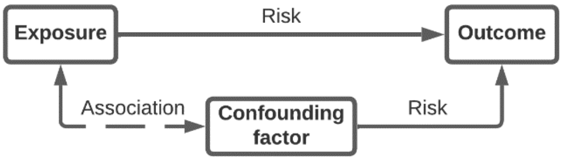

Figure 1 shows an exposure that is associated with a risk of an outcome. The confounding factor is associated with both the exposure and the outcome without being in the direct causal pathway of the two.

2.2 Counterfactuals and imbalance

The Neyman-Rubin causal model (RCM) [19] is one of the notable influences on the understanding of causal inference in observational studies. According to the model, to estimate the effect of a treatment on an outcome, we should ideally compare the treated subjects with the same subjects without treatment. Except for the treatment, all other conditions must be the same, including the time. The latter is a counterfactual and is impossible to observe. We instead compare to relatively similar, untreated controls. However, the treated and controls in our sample are often systematically different or imbalanced for the confounding factors, which leads to bias. Lowering this imbalance between treated and controls to make them more similar is thus a strategy to reduce the bias.

2.3 Coarsened exact matching

As earlier explained, the confounder covariates are distributed differently for the compared groups. Thus, we can view the bias as stemming from an imbalance in the data. Coarsened exact matching (CEM) [11] is a method for adjusting confounder bias as a preprocessing step before analysis. It belongs to a class of monotonic imbalance bounding (MIB) methods, enabling the researcher to set a maximum imbalance between treated and controls for the confounder covariates or reduce the maximum imbalance for a covariate independently of others. The theoretical foundation of CEM is outside the scope of this paper, but its use is relatively straightforward.

We partition the confounder covariates into subintervals. Each subinterval then represents a single value of a categorical variable. For example, a covariate for years of education can be partitioned into subintervals representing the highest level of education instead. In CEM, this is called coarsening and opens for simple, exact matching of similar treated and control units. It additionally helps balance the sample by pruning treated and control units without suitable matches. The coarsening is temporary and not passed to subsequent analysis steps.

The described coarsening corresponds to stratification. We stratify each covariate, and each treated and control unit will then belong to a multi-dimensional stratum. Although the current CEM software packages use uniform width histogram binning for automatic stratification, CEM as a method is not restricted to strata of uniform widths. For example, manual stratification and non-uniform widths are supported. Autostrata is an alternative approach to automatically stratifying covariates, which constructs strata of non-uniform widths.

2.4 Imbalance and unmatched trade-off

The most commonly described imbalance measure for CEM involves the relative difference between the number of treated and control units per stratum. However, the software packages use an imbalance measure based on a per stratum difference in means between the covariate values for the two groups as default. This is similar to what Appendix B of [11, p. 34] describes. We thus base our approach on the latter.

As shown in Figure 2, two strata with the same number of treated and control units can have a different internal imbalance because the covariate means are different for the groups. Nevertheless, the maximum imbalance is bounded by the stratum widths because the differences cannot be greater than the widths. Therefore, the narrower the stratum is, the lower its maximum imbalance. The lowest maximum imbalance is when each stratum only has a single unit or equal-valued units. A stratum with only one type of unit contributes zero to the imbalance, while multiple equal-valued units have an imbalance of zero. The challenge is that there is a trade-off.

CEM prunes unmatched units from the sample. If all units in a stratum are from the same group, these units are unmatched and discarded. Recall that the confounder covariates for treated and control units have different distributions. Hence, various degrees of overlap and densities will be found along the covariate axes, restricting how narrow a stratum containing both types of units can be. As we decrease the maximum imbalance, the number of unmatched units generally increases, and vice versa. Autostrata aims to lower this trade-off.

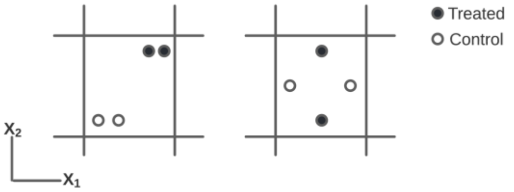

Figure 2 illustrates two strata for covariates x_1 and x_2. Both strata have two treated and two controls, but the left stratum has a higher mean difference. Also, the maximum difference is bounded by the width between the stratum edges.

2.5 Stratification problem properties

Before concluding the background section, we describe a few properties of the stratification problem relevant to solving it algorithmically.

First, the number of relevant stratum edges is finite. The reason is that a stratum edge for a covariate can be placed anywhere between two adjacent observations without changing stratum memberships. If an observation coincides with an edge, it belongs to the higher stratum. The exact position of an edge does not matter, only that it separates two adjacent observations for the given covariate. Neither do multiple stratum edges between two neighboring observations change any memberships. Further, if two or more observations have equal values for a covariate, they cannot be separated by adding stratum edges for the given covariate. Conclusively, the maximum number of relevant stratum edges equals the number of distinct values per covariate.

Second, the number of possible combinations of the stratum edges, from including no edge to including all edges, grows exponentially with the number of distinct covariate values, i.e., the problem space is non-polynomial.

Figure 3 shows all possible combinations of stratum edges for four distinct values, organized as a tree of nodes. The number of new stratifications that can be made by adding one stratum edge to a given stratification is illustrated in Figure 4.

We can deduce the number of different stratifications possible for a covariate. Let S be the set of possible stratifications for a covariate with n distinct values. Then the cardinality, $|S|$, is:

$$|S| = 1 + (n-1) + \sum_{i=2}^{n-1} 2^{i-2}(n-i) = 2^{n-1}$$

Given m covariates, the total number of combinations, $|S_{tot}|$, becomes:

$$|S_{tot}| = \prod_{i=1}^{m} |S_i| = \prod_{i=1}^{m} 2^{n_i-1}$$

For cases where all $n_i = n$ are equal:

$$|S_{tot}| = |S|^m = 2^{m(n-1)}$$

Thus, the *state space* of the problem grows exponentially with increasing numbers of distinct values and covariates. Furthermore, considering that each stratification can contain relatively many multi-dimensional strata and that we must compute imbalance measures and the number of unmatched units for each stratification, it quickly becomes computationally infeasible to perform a brute-force search through all combinations to find an optimal solution with the resources typically available to researchers.

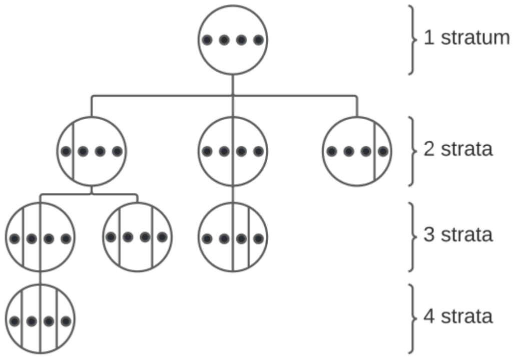

Figure 3. All possible stratifications of a covariate with four distinct observed values. The four values are illustrated as black dots within the tree nodes, and the stratum edges as vertical lines between the dots.

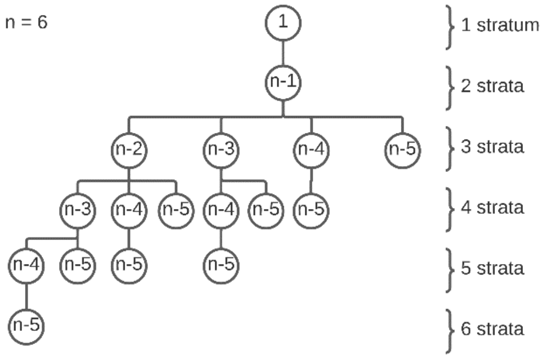

Figure 4. This tree illustrates a pattern in the number of different stratifications that can be made as we move from a given parent to a child node by adding a new stratum edge, as in Figure 3. In this case, the number of distinct values is n=6.

3 AUTOSTRATA

We now present Autostrata, a novel algorithmic approach for improved stratification of confounder covariates for CEM. Improving CEM's standard stratification method–histogram binning–is not trivial. However, analysis results need to be as free of bias as possible to avoid them from being invalid. Often, the imbalance is higher than we wanted, the number of unmatched units is high, or both. Autostrata aims to lower the trade-off between the imbalance and the number of unmatched units. Figure 5 shows a comparison of histogram binning and Autostrata.

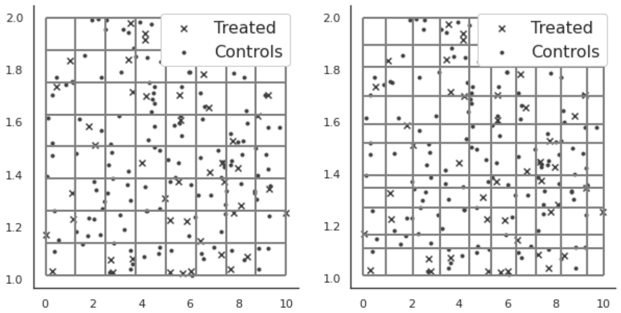

Figure 5 shows two stratifications for the same two-covariate dataset. The left plot is from histogram binning, and the right is from Autostrata. Each grid cell is a two-dimensional stratum. On the left, the strata have uniform widths. On the right, the strata widths are non-uniform.

3.1 Overall approach

This section gives an overall description of the Autostrata approach and explains its reasoning.

The generic histogram binning rules used in CEM work surprisingly well for stratification in our context. Therefore, understanding the underlying reasons is invaluable to improving the results: Any stratum containing both treated and controls is valid. Also, the sample's total maximum imbalance will be lower if the strata are narrower. To construct strata spanning over a mixture of treated and control units, regions of common support must be present for the sample, i.e., there must be some overlap in the distributions for treated and controls. The treated and controls in regions with sparse or no overlap are further apart and more dissimilar than units in denser and more overlapping regions. Because we usually have a reasonable common support level, the uniform width strata will readily contain both treated and control units. Further, units in the sparser and less overlapping regions are more likely to be pruned, as they should. These factors contribute to why histogram binning works well. Conclusively, knowing these factors makes it reasonable to assume that much of the potential for improvement is in the regions where the distributions for treated and controls overlap most.

Autostrata's strategy is to construct narrow strata while keeping the number of unmatched units low. The strata can be of varying widths. Having narrower strata on average is equivalent to more strata. We thus start with an initial stratification state where all possible stratum edges for all covariates are included (see section 2.5). This state represents the narrowest stratification that is relevant. All units will be in a stratum containing only a single unit or same-valued units. From there, we iteratively remove one edge at a time. This edge can belong to any of the covariate dimensions.

In its simplest form, the algorithm does not consider widths but removes edges one by one until the number of unmatched units is as low as requested by an input parameter. The main selection criterion for removing an edge, per iteration step, is the edge that gives the most significant reduction in unmatched units when removed. Removing a stratum edge for one dimension (covariate) merges one or more strata divided by stratum edges for other dimensions. Merging strata for a given covariate results in strata that are wider, so the increase in the average width of the strata for a covariate is strictly monotonic.

The crux of the algorithm is: For each stratum edge that we remove from the initial state, the average maximum imbalance increases. If the algorithm reaches the requested maximum number of unmatched in fewer steps, i.e., by removing fewer edges, the *average maximum imbalance* will be lower than if more steps are spent. Thus, to reduce the number of iterations needed to reach the goal number of unmatched, for each iteration, we remove the edge that gives the greatest reduction in the number of unmatched, after assessing all currently remaining edges in any dimension. If several equally good options are found, the one giving the narrowest width is chosen. In Section 3.2, we describe how the widths for different covariates are scaled to be comparable.

Autostrata also provides the researcher with input parameters for more control over the resulting stratification:

- The maximum wanted numbers of unmatched treated and controls
- The maximum allowed widths between stratum edges per covariate

The researcher can specify maximum numbers of unmatched treated and controls as two separate input parameters. The stratification process will continue until reaching both numbers or until the point when there is no closer solution. For example, suppose the stratification algorithm reaches one of the requested maximum numbers of unmatched for either treated or controls. It will then continue until reaching the requested number of unmatched for the other group. It continues iterating, and the numbers can continue to improve for both treated and controls. Section 3.2 describes how Autostrata incorporates weights to account for the difference in the requested maximum numbers of unmatched treated and controls while iterating.

Further, Autostrata has a parameter for the maximum allowed stratum width per covariate, and it will not create strata wider than the given widths. If widths are not of importance, a large or infinite value can be given as input instead. The background for the maximum width parameter is that researchers may want to set a maximum difference, *caliper*, between treated and controls for the covariates—for example, max five years age difference or five points difference for a given performance score. In addition, setting a maximum width restricts the maximum imbalance. Another reason to set widths, which concerns the algorithm, is to prevent a single or a few strata from expanding too much while leaving others unchanged. Broader strata have a higher potential imbalance. It is possible to imagine that, on average, a large stratum combined with many narrow ones may somewhat cancel each other out imbalance-wise, but it is probably not what we want. A large stratum will still have a greater risk of being imbalanced. Lastly, we can use the widths produced by CEM's histogram binning as input to Autostrata. Histogram binning only supports uniform width strata, but Autostrata can use these widths as the maximum allowed when defining strata of non-uniform widths.

3.2 Heuristics

In section 3.1, we gave an introduction to the overall approach. Autostrata is an algorithmic approach to stratifying covariates that starts with an initial state where all stratum edges are present and iteratively removes one edge at a time until the end criterion is met or no further improvements are found. Here, we describe the heuristics in more detail.

When we remove a stratum edge along the direction of one dimension (covariate), two and two strata become merged to form new, wider strata. Removal of an edge usually results in more than two strata being merged because there are also edges along the other dimensions separating the covariate values into distinct strata. If two neighboring strata contain only treated and only controls, respectively, merging the two strata results in a stratum with a mix of both types. These units are no longer unmatched and, thus, not pruned from the sample.

Autostrata has two criteria for choosing which stratum edge to remove for each iteration. The first criterion has the highest priority, and the second criterion applies only to alternatives with equally good values for the first. The two criteria are:

1. Choose the greatest relative increase in matched treated and controls if the stratum edge is removed
2. Choose the stratum with the narrowest width

Instead of using the increase in matched units directly, Autostrata uses a weighted measure for increase. Let Δ_t and Δ_c be the increase in the number of matched treated and control units, respectively, when we remove a given stratum edge. The relative increase, Δ_{rel}, is then:

$$\Delta_{rel} = w_t \Delta_t + w_c \Delta_c$$

, where w_t and w_c are weights. The weight for the treated group, $w_t = w(t)$, and control group, $w_c = w(c)$, is found as follows:

$$w(g) = \begin{cases} \dfrac{m_g^{cur} - m_g^{max}}{n_g - m_g^{max}}, & m_g^{cur} - m_g^{max} \geq 0 \\ 0, & m_g^{cur} - m_g^{max} < 0 \end{cases}$$

, where g is the group, m_g^{cur} is the number of currently unmatched units for the group, m_g^{max} is the requested maximum number of unmatched for the group, and n_g is the total number of units from the group in the sample. Here, we also assume that $n_g > m_g^{max}$.

The purpose of the weights is threefold:

1. If one group is represented less than the other, each new matched unit from the group should weigh more.
2. The researcher can set parameters for how many unmatched (pruned) treated and controls are acceptable. The difference $n_g - m_g^{max}$ takes into account that the gap between available and discardable units can differ between groups.
3. If Autostrata has reached the goal for the number of unmatched units for one group, an increase in the other groups should weigh more when choosing an edge to remove. As one group comes closer to the goal, reducing the number of unmatched for the other group is prioritized higher. The difference $m_g^{cur} - m_g^{max}$ is the remaining units to match for the given group.

Width is the second selection criterion for edge removal. The widths must be scaled because Autostrata compares stratum edges from all covariates per iteration. We compute a scale factor by removing outliers and taking the min-max difference. Observations having a standard score, $|z| \geq 3$, are outliers. The data can be scaled once as an initial step. In that case, the maximum widths must be scaled as well. Also, we must restore the resulting stratum edges to the original scale. For clarity, the pseudocode in Listing 1 does not scale the data until needed.

3.3 Algorithm

Here we present the algorithm in pseudocode form. The pseudocode is at an abstraction level sufficient to implement the algorithm. However, we omit implementation details and performance enhancements that do not contribute to the understanding. Listing 1 presents the algorithm in pseudocode form, and Table 1 describes the variables used in the listing.

Variable	Meaning
tr and ct	The covariate values for the treated and the control units
Δ_{best}	The best relative increase in matched units for the current iteration
Δ_{cur}	The relative increase in matched units for currently assessed edge
Δ_t and Δ_c	The increase in the number of matched treated and controls for assessed edge
m_t^{cur} and m_c^{cur}	The current number of unmatched treated and controls
m_t^{max} and m_c^{max}	The requested maximum number of unmatched treated and controls
covariates	The covariates (dimensions)
cov	The current covariate
edges	The current set of edges, including the outer left- and rightmost edge per covariate
n_{edg}	The number of edges in the current set of edges
$\text{edges}_{\text{cov}}$	The current set of edges for the current covariate, *excluding* the outer left and right edges
e_{cur}	The currently assessed edge
e_{sel}	The currently best edge for the iteration and candidate for selection
e_{l} and e_{h}	e_{cur}'s lower and higher adjacent edges
$\text{width}_{\text{cur}}$	The scaled widths of merged strata if we remove the currently assessed edge
$\text{width}_{\text{sel}}$	The scaled widths of strata if removing the iteration's current candidate for best edge
$\text{widths}_{\text{max}}$ and $\text{width}_{\text{max}}$	The set of maximum allowed stratum widths, and the maximum width for the current covariate

Table 1. The pseudocode variables and their meaning

```
Autostrata Algorithm
Input: tr, ct, $\text{widths}_{\text{max}}$, $m_t^{max}$, $m_c^{max}$
Output: edges
Initialization of variables:
$\text{edges}_{\text{cov}}$ ← one edge per distinct covariate value
$m_t^{cur}, m_c^{cur}$ ← calculate the initial number of unmatched treated and controls
Stratification:
```

```
while ($m_t^{cur} > m_t^{max}$ or $m_c^{cur} > m_c^{max}$)
and ($n_{edg} > 0$) do
    $\Delta_{best} \leftarrow -1$
    $width_{sel} \leftarrow \infty$
    $e_{sel} \leftarrow$ nil
    for cov in covariates do
        for $e_{cur}$ in $edges_{cov}$ do
            get $e_l$ and $e_h$
            $width_{unscaled} \leftarrow | e_h - e_l |$
            if $width_{unscaled} > width_{max}$ then
                continue // stratum too wide
            $width_{cur} \leftarrow$ scaled_width($e_l$, $e_h$)
            $\Delta_t$, $\Delta_c \leftarrow$ the difference in numbers of
            unmatched (for the multi-dimensional
            strata) between $e_l$ and $e_h$ before and
            after removing $e_{cur}$
            $\Delta_{cur} \leftarrow$ relative_increase($\Delta_t$, $\Delta_c$)
            if ($\Delta_{cur} > \Delta_{best}$) or ($\Delta_{cur} == \Delta_{best}$
            and $width_{cur} < width_{sel}$) then
                $\Delta_{best} \leftarrow \Delta_{cur}$
                $width_{sel} \leftarrow width_{cur}$
                $e_{sel} \leftarrow e_{cur}$
            end // if
        end // for $e_{cur}$
    end // for cov
    if $e_{sel}$ == nil then
        break // no more improvements found
    else
        remove $e_{sel}$ from edges
        update $m_t^{cur}$ and $m_c^{cur}$
    end
    if $m_t^{cur} \le m_t^{max}$ and $m_c^{cur} \le m_c^{max}$ then
        break // goal reached
end // while
return edges
```

Listing 1. Pseudocode for the algorithm

3.4 Implementation

A version of the algorithm corresponding to Listing 1 was implemented in Python 3.9 (https://python.org), with some added performance enhancements. For example, we utilize Numba (https://numba.pydata.org) for counting unmatched units in strata, yielding a speedup [20, p. 125] of 2.25 for the algorithm as a whole when stratifying for Dataset 3 in Table 2 on an Intel i7-8850H CPU with 12 logical cores. A far more significant performance enhancement is achieved by caching already computed results for each stratum. The same strata are visited repeatedly during the iterations, and the algorithm finishes 17.67 times faster for Dataset 1 in Table 2 when reusing already computed results. Further, strata not affected by removing a given edge are not visited unnecessarily. Lastly, only relevant units are included in computations regarding subsets of strata.

Still, there is plenty of room to enhance performance. Many of the algorithm's computational tasks can be performed independently, e.g., the difference in unmatched units if a given edge is removed. Such independent computations that are well suited for parallelization are often termed embarrassingly parallel [21, p. 79-98]. A systematic approach to parallelizing algorithms is found in Foster's methodology [22]. In addition to parallelization, we can enhance the performance by designing data structures for efficient access to frequently used data and extensively reusing previously computed results in the algorithm's iterations. For clarity, we concentrate on the basic algorithm in this paper, leaving the suggested performance enhancements to future work.

The accompanying source code for the paper is available on GitHub (https://github.com/jo-inge-arnes/autostrata).

4 EXPERIMENTS AND RESULTS

4.1 Datasets

A generator for synthetic data was implemented that lets us draw random samples from a composition of distributions for treated and controls. Figure 6–Figure 8 show the datasets as violin and swarm plots, and Table 2 shows the number of units and the mixed distributions for the datasets.

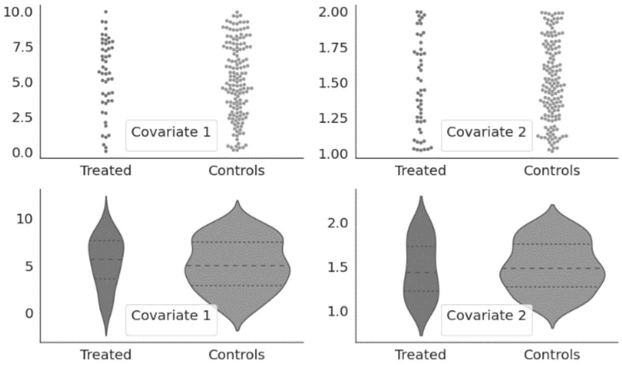

Figure 6 shows the swarm and violin plots of Dataset 1 with two uniformly distributed covariates.

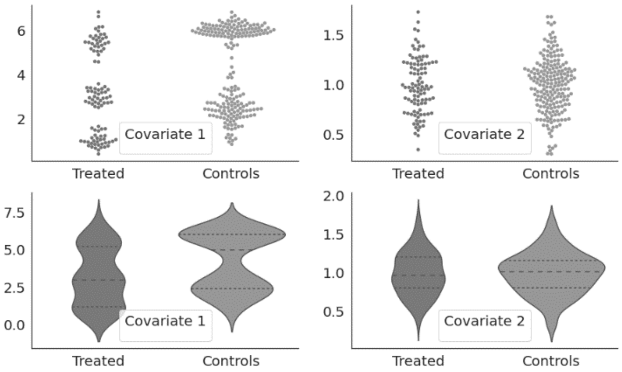

Figure 7 shows the swarm and violin plots of Dataset 2 with a mixture of Gaussians.

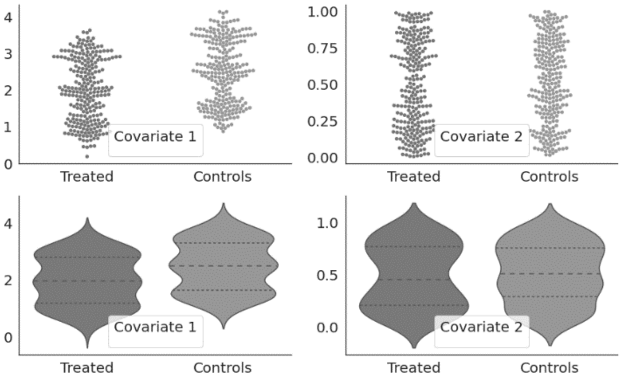

Figure 8. Dataset 3 has a mixture of Gaussians for Covariate 1 and uniform distribution for Covariate 2.

	Treated	Controls
Dataset 1	50 units	150 units
Cov. 1	U(0, 10)	U(0, 10)
Cov. 2	U(1, 2)	U(1, 2)
Dataset 2	100 units	200 units

Cov. 1	N(1, $\frac{1}{3.5}$)
	N(2, $\frac{1}{3.5}$)
	N(5.5, $\frac{1.5}{3.5}$)
		N(2.5, $\frac{2.5}{3.5}$)
		N(6, $\frac{1}{3.5}$)
		–

Cov. 2	N(0.95, $\frac{1}{3.5}$)	N(1.0, $\frac{1}{3.5}$)
Dataset 3	250 units	250 units

Cov. 1	N(1, $\frac{1}{3.5}$)
	N(2, $\frac{1}{3.5}$)
	N(3, $\frac{1}{3.5}$)
		N(1.5, $\frac{1}{3.5}$)
		N(2.5, $\frac{1}{3.5}$)
		N(3.5, $\frac{1}{3.5}$)

Cov. 2	U(0,1)	U(0, 1)

Table 2 shows the number of units and the mixed distributions for the datasets. U(min, max) stands for uniform and N(μ, σ) for normal distribution.

4.2 Experiments

To automate the experiments, we wrote Python and R scripts. The role of the R scripts is to call the CEM library. A reference manual for the CEM library is available online (https://CRAN.R-project.org/package=cem). In the code for the experiments, rpy2 (https://rpy2.github.io) is used to bridge between Python and R.

The experiments are as follows:

1. We call CEM to get pre-stratification scores and statistics for the given dataset.
2. Next, CEM is used to stratify the covariates by applying Scott's rule for histogram binning. It also computes the number of unmatched units, imbalance scores, and other statistics.
3. We then pass CEM's outputted number of unmatched units and stratum widths to Autostrata.
4. Autostrata stratifies the covariates.
5. Autostrata's outputted stratum edges are given as input to CEM, which uses them to stratify and compute statistics equivalent to step 2.

Two experiments are conducted per dataset. They differ only in how the results are passed to Autostrata in Step 3:

Input type	Input parameters
P1	The numbers of unmatched treated and controls from histogram binning are passed as m_t^{max} and m_c^{max} and the bin widths are passed as $\text{widths}_{\text{max}}$.
P2	The m_t^{max} and m_c^{max} values are as in P1, but $\text{widths}_{\text{max}}$ values are set to infinity.

Table 3. Input parameters. See Table 1 for variables

4.3 Results

Table 4 shows the experiment results.

DS1, DS2, and DS3 are headers for the results of the three datasets. The top column headers stand for 'results before stratification' (Before), 'stratification with histogram binning' (Hist.), and the input types P1 and P2 from Table 3. 'Res.' is an abbreviation for results, and 'Imp.' is the percent improvement compared to histogram binning. The row labels denote multivariate imbalance measure (MIM), total unmatched (UM$_{\text{TOT}}$), unmatched treated (UM$_{\text{TR}}$), and unmatched controls (UM$_{\text{CT}}$).

TOI is the percent improvement in the trade-off, which is the sum of the improvements for UM$_{\text{TOT}}$ and MIM.

	Before	Hist.	P1		P2	
	Res.	Res.	Res.	Imp.	Res.	Imp.
DS1						
MIM	0.240	0.199	0.167	16%	0.187	6%
UM$_{\text{TOT}}$	0	65	80	-23%	54	17%
UM$_{\text{TR}}$	0	4	5	-25%	3	25%
UM$_{\text{CT}}$	0	61	75	-23%	51	16%
TOI	–	–	–	-7%	–	**23%**
DS2						
MIM	0.465	0.273	0.266	3%	0.335	-23%
UM$_{\text{TOT}}$	0	88	84	5%	72	18%
UM$_{\text{TR}}$	0	16	24	-50%	16	0%
UM$_{\text{CT}}$	0	72	60	17%	56	22%
TOI	–	–	–	**7%**	–	-5%
DS3						
MIM	0.348	0.290	0.174	40%	0.280	3%
UM$_{\text{TOT}}$	0	53	92	-74%	39	26%
UM$_{\text{TR}}$	0	21	53	-152%	21	0%
UM$_{\text{CT}}$	0	32	39	-22%	18	44%
TOI	–	–	–	-34%	–	**30%**

Table 4. Results from experiments. Best TOI results per dataset are in bold and thicker cell borders.

5 DISCUSSION

Table 4 shows that both imbalance and the total number of unmatched units are lower for Autostrata for all three datasets. The input parameter type P2 gave the best results for DS1 and DS3, while P1 gave the best for DS2. The difference is that P2 sets the maximum allowed stratum widths to infinity, which effectively disables the parameter. By visually comparing the swarm plots in Figure 6–Figure 8, we see the difference between DS2 and the other two: DS2 has several regions with minimal overlap between treated and controls. As Section 3.1 explains, finding narrow strata with mixed types of units is easier in regions with high overlap. Therefore, restricting the widths is usually not necessary in such regions. Autostrata also works well for sparser overlap, but as illustrated by the experiment for DS2, setting maximum widths is more important.

Autostrata competed with CEM's best effort in the experiments, and we passed parameters not necessarily ideal for non-uniform widths. It is possible to adjust these parameters manually or programmatically, but for objectivity, we use the unchanged output from CEM as input to Autostrata.

Lastly, Autostrata can be used stand-alone. A researcher can decide the acceptable differences between treated and controls based on domain knowledge. The researcher can also request a maximum number of unmatched units. Autostrata thus provides researchers with more up-front control. After stratification, the researcher can input the stratum edges to the CEM software as manual cutpoints. A combination is even possible, where Autostrata stratifies a subset of the covariates given to CEM.

6 RELATED WORK

Aikens, R.C. et al. [23] present *Stratamatch*, a method for stratification of covariates for CEM. Only datasets from a minimum of 5 000 up to millions of observations are recommended. The method divides the dataset into training (pilot) and analysis sets, and the resulting strata are close to equal-sized. The size must be manually decided.

Jackson, B. et al. [24] present an algorithm for optimal data partitioning on an interval that Scargle, J.D. et al. [25] apply for astronomical time series. The algorithm supports custom fitness functions, and we tried defining a function. However, a common issue is the unwanted case of one subinterval per value; thus, the researcher must choose an expected number of subintervals. Also, while theoretically possible to extend for multivariate data, the algorithm is primarily univariate.

7 CONCLUSION

We have presented Autostrata, an algorithmic approach to stratifying confounder covariates. Autostrata shows improved results compared to the standard CEM stratification. In addition, it provides the researcher with parameters for controlling the stratification. Autostrata can be used stand-alone.

8 REFERENCE

[1] Ranganathan, P., Aggarwal, R. "Study Designs: Part 3 – Analytical Observational Studies" in *Perspectives in Clinical Research*, Vol. 10, Issue 2, pp. 91-94. 2019.

[2] Hariton, E., Locascio, J.J. "Randomised Controlled Trials – The Gold Standard for Effectiveness Research" in *BJOG: An International Journal of Obstetrics & Gynaecology*, Vol. 125, Issue 13, pp. 1716-1716. 2018.

[3] Arnes, J.I., Bongo, L.A. "The Beauty of Complex Designs" in *Advancing Systems Epidemiology in Cancer: Exploring Trajectories of Gene Expression*, pp. 23-47. Scandinavian University Press, 2020.

[4] Kim, S. "Case-Cohort Studies vs Nested Case-Control Studies" in *Datum Newsletter Division of Biostatistics*, Vol. 22, Issue 1, pp. 1-2. 2016.

[5] Ngo, L.H., et al. "Methodologic Considerations in the Design and Analysis of Nested Case-Control Studies: Association Between Cytokines and Postoperative Delirium" in *BMC Medical Research Methodology*, Vol. 17, Issue 1, pp. 88. 2017.

[6] Alexander, L.K., et al. "Confounding Bias, Part I" in *ERIC Notebook*, Vol. 11. 2015.

[7] Alexander, L.K., et al. "Confounding Bias, Part II and Effect Measure Modification" in *ERIC Notebook*, Vol. 12. 2015.

[8] Howards, P.P. "An Overview of Confounding. Part 1: The Concept and How to Address It" in *Acta Obstetricia et Gynecologica Scandinavica*, Vol. 97, Issue 4, pp. 394-399. 2018.

[9] Wacholder, S., et al. "Selection of Controls in Case-Control Studies: I. Principles" in *American Journal of Epidemiology*, Vol. 135, Issue 9, pp. 1019-1028. 1992.

[10] Tripepi, G., et al. "Stratification for Confounding – Part 1: The Mantel-Haenszel Formula" in *Nephron Clinical Practice*, Vol. 116, Issue 4, pp. 317-321. 2010.

[11] Iacus, S.M., King, G., Porro, G. "Matching for Causal Inference Without Balance Checking" in *SSRN Electronic Journal*. 2008.

[12] Blackwell, M., et al. "cem: Coarsened Exact Matching in Stata" in *The Stata Journal*, Vol. 9, Issue 4, pp. 524-546. 2009.

[13] Iacus, S.M., King, G., Porro, G. "cem: Software for Coarsened Exact Matching" in *Journal of Statistical Software*, Vol. 30, Issue 9, pp. 1-27. 2009.

[14] Sturges, H.A. "The Choice of a Class Interval" in *Journal of the American Statistical Association*, Vol. 21, Issue 153, pp. 65-66. 1926.

[15] Scott, D.W. "On Optimal and Data-Based Histograms" in *Biometrika*, Vol. 66, Issue 3, pp. 605-610. 1979.

[16] Freedman, D., Diaconis, P. "On the Histogram as a Density Estimator: L2 theory" in *Zeitschrift für Wahrscheinlichkeitstheorie und Verwandte Gebiete*, Vol. 57, Issue 4, pp. 453-476. 1981.

[17] Shimazaki, H., Shinomoto, S. "A Method for Selecting the Bin Size of a Time Histogram" in *Neural Computation*, Vol. 19, Issue 6, pp. 1503-1527. 2007.

[18] King, G., Nielsen, R. "Why Propensity Scores Should Not Be Used for Matching" in *Political Analysis*, Vol. 27, Issue 4, pp. 435-454. 2019.

[19] Sekhon, J.S. "The Neyman-Rubin Model of Causal Inference and Estimation via Matching Methods" in *The Oxford Handbook of Political Methodology*, Vol. 2, pp. 1-32. 2008.

[20] Pacheco, P.S. *An Introduction to Parallel Programming*. Morgan Kaufmann, Burlington, MA, 2011.

[21] Wilkinson, B., Allen, M. *Parallel Programming: Techniques and Applications Using Networked Workstations and Parallel Computers*. Pearson/Prentice Hall, Upper Saddle River, N.J, 2005.

[22] Foster, I. *Designing and Building Parallel Programs*. Addison-Wesley, Reading, MA, 1995. Also available from: https://www.mcs.anl.gov/~itf/dbpp/ Accessed 2022-06-26.

[23] Aikens, R.C., et al. "stratamatch: Prognostic Score Stratification Using a Pilot Design" in *arXiv preprint arXiv:2001.02775*. 2020.

[24] Jackson, B., et al. "An Algorithm for Optimal Partitioning of Data on an Interval" in *IEEE Signal Processing Letters*, Vol. 12, Issue 2, pp. 105-108. 2005.

[25] Scargle, J.D., et al. "Studies in Astronomical Time Series Analysis. VI. Bayesian Block Representations" in *Astrophysical Journal*, Vol. 764, Issue 2, pp. 167. 2013.

9 ACKNOWLEDGEMENT

We want to thank Marc Weitz for inspiring the use of swarm and violin plots.