The Impact of Fair Aggregation on Forecast Accuracy and Fairness: An Application to Enrollment Forecasting

Jade Zhang

Abstract

Fairness across racial/ethnic groups in college enrollment is an important social topic; however, fairness concerns have not been addressed in the forecasting scheme, although enrollment forecasting is a critical and fundamental practice for institutions. To investigate this issue, we developed a hierarchical structure for forecasting regarding race/ethnicity based on enrollment data from the University of California, Berkeley. Since forecast accuracy and fairness are sensitive to aggregation levels, we proposed a flexible fair aggregation method based on binary hierarchical K-medoids clustering with both similarity and fairness as criteria. Our results show that the hierarchical forecasting structure can increase the forecast accuracy, and the fair aggregation method can improve the forecasting performance and affect the fairness. Specifically, fair aggregation can provide flexibility for institutions to present fairness and valuable insights in terms of fairness and similarity across groups. In general, one interested in selecting aggregation levels to improve hierarchical forecasts can adopt this methodology.

Keywords: Enrollment Forecasting, DEI (Diversity, Equity, and Inclusion), F4SG (Forecasting for Social Good), Fairness, Time-series Aggregation

1. Introduction

Fairness, defined as "the absence of any prejudice or favoritism toward an individual or group based on their inherent or acquired characteristics", has recently gained significant attention in various real-world applications (Mehrabi et al., 2021). For example, Google's photo application once misidentified black people as "gorillas" (Zhang, 2022). Following that, ProPublica published a report claiming that the COMPAS (Correctional Offender Management Profiling for Alternative Sanctions) software used to predict recidivism has treated black defendants unfairly (Angwin et al., 2016). Furthermore, a study analyzing three gender classification algorithms found that the algorithms were more accurate when classifying light-skinned men compared with other groups (Buolamwini and Gebru, 2018). Besides the machine learning algorithms, matters of fairness should be addressed in forecasting applications because an unfair forecasting model may result in different levels of forecast accuracy for different groups (Tsai et al., 2022). In terms of decision-making, it is necessary to analyze fairness due to some issues associated with forecasting certain properties. Consideration of fairness can assist forecasters and decision-makers with more informed planning.

One prototypical example is enrollment forecasting, which is used to determine how many students that will enroll in the future. Enrollment forecasting is essential for institutions to prepare for all enrollment-related activities, such as tuition policy, budget planning, and faculty staffing (Brinkman and McIntyre, 1997; Weiler, 1987). From a fairness perspective, institutions may want to design fair algorithms for enrollment forecasting to ensure that different racial/ethnic groups are treated fairly. Furthermore, many institutions want to admit more students from minority groups, particularly because higher education is crucial for attaining positions of power and influence in society (Anderson and Svrluga, 2022). With accurate forecasts, institutions can design plans that more accurately reflect their goals.

Although most enrollment forecasting literature has focused on finding a suitable algorithm (including comparing existing ones) to forecast the total enrollment at the institutional level, several official reports have considered enrollment forecasting based on race/ethnicity (e.g., Hahn et al., 2015; Hussar and Bailey, 2020, 2019). One previous study adopted a bottom-up approach forecasting enrollment by race/ethnicity and then aggregating the results into a total to account for the differences across racial/ethnic groups (Grip, 2009). Because the study is specific to its data, it might be beneficial to evaluate a similar approach in other scenarios with the same consideration for race/ethnicity as one of the demographic factors influencing enrollment (Brinkman and McIntyre, 1997). Particularly, in the enrollment forecasting setting, the patterns of historical enrollments for different races/ethnicities can vary greatly.

To the best of our knowledge, first, the bottom-up approach with respect to race/ethnicity has never been used in college enrollment forecasting. Second, although race/ethnicity has been considered in enrollment forecasting, the issue has not been investigated with an eye toward fairness. Third, the aggregation rule has been inconsistent. Because different aggregation rules can lead to different forecast accuracy and fairness (e.g., Blyth, 1972; Zotteri et al., 2005), the aggregation rules for race/ethnicity can be analyzed on the basis of these aspects, namely, forecast accuracy and fairness. Furthermore, previous studies only forecast enrollment at the aggregate race/ethnicity level without considering disaggregate categories. Institutions may be interested in the forecast of more detailed categories; for example, the under-representation of Chinese students in US colleges has raised concerns for several institutions (Flannery, 2022). Accurate forecasting of detailed categories can assist institutions in developing advertisements and promoting scholarships.

Considering the existing problems, we propose a hierarchical framework for enrollment forecasting based on race/ethnicity. Using the hierarchical structure, the enrollment forecast is first obtained at the aggregate race/ethnicity level. The bottom-up approach is then applied to forecast at the total level, whereas the top-down approach is applied to forecast at the disaggregate race/ethnicity level. The bottom-up approach considers differences across racial/ethnicity groups, and the forecast of total enrollment may be more accurate, whereas the top-down approach overcomes the problems of low-quality and randomness in the data of disaggregate racial/ethnic groups. In contrast with previous studies that only consider forecast accuracy, we propose a fair aggregation method based on a bisecting K-medoids algorithm that considers both fairness and similarity (to improve the forecast accuracy), in which the statistical disparity is adopted as a fairness metric and short time series distance (STS) is used to measure dissimilarity (Möller-Levet et al., 2003).

Forecasting for social good (FSG) is defined as "a forecasting process that aims to inform decisions that prioritize the thriving of humanity over the thriving of economies by enhancing the social foundation and ecological ceilings that impact the public as a whole on both local and global level" (Rostami-Tabar et al., 2022). Inspired by this concept, our study is the first to analyze the impact of aggregation methods on forecast performance and fairness in enrollment forecasting using fairness metrics, addressing the concern of fairness to advance social good.

Section 2 is a detailed review of the existing literature on enrollment forecasting, hierarchical forecasting, the impacts of aggregation on forecasting, aggregation and clustering methods, and fairness metrics. In Section 3, the methodology and forecast models used herein are defined. In Section 4, the empirical results of this study is are presented. In Section 5, the theoretical and practical significance of the results are discussed and the conclusion is provided.

2. Literature review

2.1. Enrollment forecasting with a focus on race/ethnicity

It is important to consider race/ethnicity in enrollment forecasting from both the forecasting accuracy and fairness perspectives. Notably, race/ethnicity has a significant impact on enrollment (Brinkman and McIntyre, 1997), and different racial/ethnic groups might exhibit different historical enrollment patterns and trends. For example, Baker et al. (2018) discovered that the college enrollment gaps between Black-White and Hispanic-White groups shrank between 1986 and 2014, indicating that enrollment trends over time differ across these racial/ethnic groups. Figure 1 shows the percentage of recent high school graduates enrolled in college by four racial/ethnic groups from 2010 to 2020 according to the US Census Bureau, highlighting the different patterns and trends among race/ethnic groups. The differences may be discernible for institutions as well. Additionally, because we focus on the enrollment counts instead of percentage in enrollment forecasting and the magnitudes of counts can be large at the institutional level, such differences can have a significant impact in terms of forecasting. Thus, addressing the differences among racial/ethnic groups may be beneficial for enrollment forecasting in terms of forecast accuracy.

Society must maintain the fairness of higher education among different racial/ethnic groups because a bachelor's degree provides one of the best opportunities for an economically secure life (Torche, 2011). To enhance social fairness, institutions are looking for methods to increase diversity and present the fairest opportunities possible. In this case, enrollment forecasts that emphasize ethnicity may assist institutions in gaining some insight into this issue and making appropriate plans.

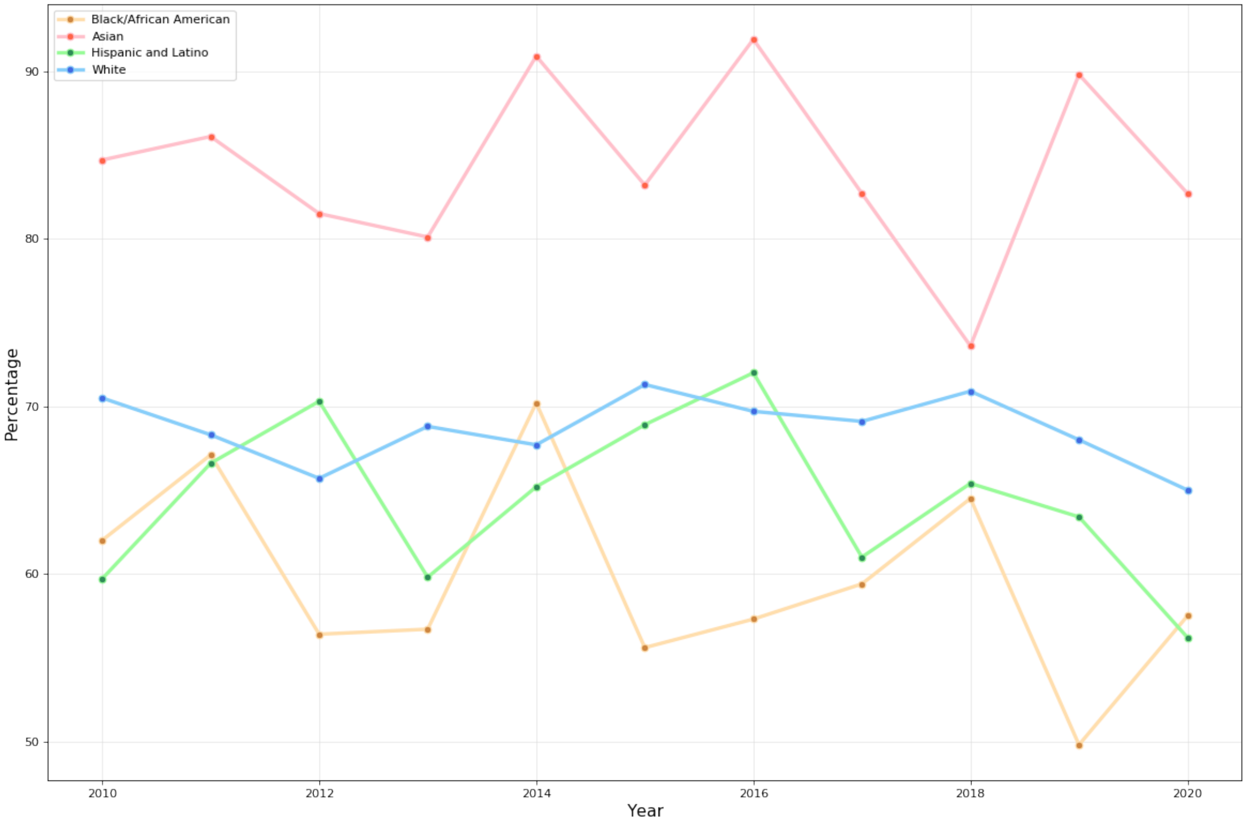

Figure 1: Percentage of recent high school graduates enrolled in college by race/ethnicity from 2010 to 2020

However, most enrollment forecasting studies have focused on either finding a suitable algorithm (including comparing the existing ones) (e.g., Song and Chissom, 1993; Stallings and Samanta, 2014; Yang et al., 2020) or choosing appropriate input variables by testing different models (e.g., Walczak and Sincich, 1999; Chen et al., 2019). Among all forecasting analyses focusing on algorithms and forecast techniques, we rarely see analysis regarding race/ethnicity, besides official reports, in which the different racial/ethnic groups have different trends and patterns in both historical enrollment counts and forecasts. (Hahn et al., 2015; Hussar and Bailey, 2019, 2020). Such differences cannot be addresses in the forecasts at the total level.

A previous study has addressed this issue in enrollment forecast applications for K-12 districts. Grip (2009) analyzed whether forecasting enrollments for K-12 districts by race/ethnicity produces more accurate results in New Jersey school districts. The study used a hierarchical structure and a bottom-up approach (forecasting enrollment by race/ethnicity and aggregating to a total) because it is expected that different racial/ethnic groups have different fertility rates and migration patterns. The results indicate that forecasting enrollment by race/ethnicity might be appropriate for larger districts with low majority percentages, and forecasts by race/ethnicity are greater in magnitude than those forecasts performed with all groups combined. Even though the analysis was limited to a specific dataset, it provides a hypothesis that forecasting enrollment by race/ethnicity can lead to more accurate results. Regarding the possible difference in the historical enrollment of undergraduate institutions, the assumption is likely to be true and worth testing.

In addition, the race/ethnicity categorization is usually ambiguous and inconsistent. For example, Hussar and Bailey (2020) followed the US Census Bureau, which uses "White, Black, Asian/Pacific Islander, Hispanic, American Indian/Alaska Native, Two or more races", whereas Hahn et al. (2015) used "White, African American, Hispanic, Other". Furthermore, the University of California (UC) defined more categories, such as "African American, American Indian/Alaska Native, Asian, Hispanic Latinx, Native Hawaiian Pacific Islander, Southwest Asian North African, and White" (University of California, 2021). Notably, in all of these examples, the race/ethnicity categories are aggregates. For instance, UC has 68 disaggregate races in total [1].

However, despite its importance, there has been little research on how to forecast the enrollment of disaggregate racial/ethnic groups. Although forecasting enrollment at the disaggregate ethnicity level can be challenging because of the low-quality data and high randomness, it may help the institutions and policymakers gain critical insights.

2.2. Hierarchical forecasting

We apply a hierarchical forecasting method to solve the problems that are currently associated with enrollment forecasting. As mentioned in the enrollment forecasting literature, the bottom-up approach can reflect individual components (individual racial/ethnic groups in our case) with different patterns of variation because individual forecasts for each segment are combined to produce a forecast of the aggregate level. With the assumption that each race/ethnicity has unique characteristics and patterns regarding enrollment counts, a bottom-up approach that aggregates the forecasts of each race/ethnicity for the forecasts of total enrollment counts may increase the forecast accuracy at the institutional level.

However, it is impractical to fit the forecast algorithms at the most disaggregate ethnicity level. For decades, the categorization rules have been changing because we need a more detailed categorization as time passes, resulting in the inconsistency and a lack of observations for some groups. Furthermore, the enrollment of some groups may have a small magnitude in most years, which can appear to be random. Such randomness and quality of data make it challenging to implement an algorithm that accurately forecasts for these groups.

The forecasts of the disaggregate race/ethnicity can then be obtained using the top-down approach, in which the forecasts of aggregate data are disaggregated to produce the forecasts for

[1] Complete ethnic groups in the UC categorization: Afghan, African, African American/Black, Algerian, American Indian/Alaska Native, Armenian, Asian Indian, Assyrian/Chaldean, Azerbaijani, Bahraini, Bangladeshi, Berber, Cambodian, Caribbean, Chinese, Circassian, Cuban, Djiboutian, East Indian, Egyptian, Emerati, Fijian, Filipino, Georgian, Guamanian/Chamoro, Hawaii Other Pacific Islander, Hawaiian, Hmong, Indonesian, Iranian, Iraqi, Israeli, Japanese, Jordanian, Korean, Kurdish, Kuwaiti, Laotian, Latin American/Latino, Lebanese, Libyan, Malaysian, Mauritanian, Mexican/Mexican American/Chicano, Moroccan, Omani, Other African American/Black, Other Asian, Other North African, Other Pacific Islander, Other Southwest Asian, Other Spanish American/Latino, Other White, Pakistani, Palestinian, Puerto Rican, Qatari, Samoan, Saudi Arabian, Somali, Sri Lankan, Sudanese, Syrian, Taiwanese, Thai, Tongan, Tunisian, Turkish, Vietnamese, White/Caucasian, White/Middle Eastern, White/North African, and Yemeni.

individual segments. Because the disaggregate forecast only needs a historical proportion of the corresponding disaggregate group, rather than fitting a model based on the disaggregate observations, the top-down approach can overcome the problems encountered when using the disaggregate race/ethnicity data. Additionally, aggregate data are less volatile than its individual components. Therefore, a forecast of the aggregate may be relatively more accurate than forecasts of its individual components (Lapide, 2006).

Thus, our objectives include obtaining enrollment forecasts at both the total (institutional) and disaggregate race/ethnicity level, with concerns about the differences among racial/ethnic groups and the low-quality data at the disaggregate level. One possible solution is to implement a hierarchical structure that includes enrollment counts at the total, aggregate race/ethnicity, and disaggregate race/ethnicity levels. Bottom-up and top-down approaches can also be combined to improve forecasting performance (Kahn, 1998), which involves forecasting at the aggregate race/ethnicity level and aggregating the forecasts to obtain the forecasts at the total level while assigning proportions to obtain the forecasts at the disaggregate race/ethnicity level. The aggregation level is a critical factor in such a process, and the aggregate groups must include disaggregate segments with similar patterns to ensure the forecast performance (e.g., Lapide, 2006; Zotteri et al., 2005; Goehry et al., 2019). In this case, it is necessary to investigate how to aggregate the disaggregate racial/ethnic groups. Moreover, rather than only using the common categorization rule, it is worth considering the other aggregation methods because the disaggregate groups that are classified in the same category based on common sense do not need to have a similar pattern.

2.3. Clustering methods for time series [2]

Various approaches for aggregation and clustering have been applied to address the similarities between disaggregate groups. The most fundamental approach is to cluster data based on characteristics and properties. For example, Zotteri et al. (2005) aggregated stores at the chain level and Mirowski et al. (2014) classified households based on their locations and sections. In this study, the disaggregate racial/ethnic groups can be grouped into broader categories (e.g., Chinese, Japanese, and Korean can be considered Asian). This approach is intuitive because we can sometimes assume that the time series in the same group share some similarities, but it might fail to address the hidden similarities between each time series.

The more advanced approaches can be categorized as model, feature, and raw data-based (Liao, 2005; Aghabozorgi et al., 2015). The first two approaches are more suitable for complex time series information because feature-based approaches typically involve dimensionality reductions or feature vector calculations (i.e., seasonality, trend, etc.), and the model-based approaches require model fitting to transform the raw series into model parameters. Because the enrollment series is one dimensional and usually very short (≤ 30 observations), we prefer the raw data-based approach.

[2] The difference between clustering and aggregating contexts is that clustering is usually used to categorize similar time series data so that the same model can be trained for each cluster, whereas aggregating involves summing up the individuals in one cluster. Herein, we discuss general clustering methods because the principles are similar.

The most conventional metrics for the raw data-based approach are Euclidean distance and its dynamic version, known as dynamic time warping (Aghabozorgi et al., 2015). However, these metrics can capture only the differences between magnitudes, but not the patterns and shapes of a time series. Because we aim to aggregate racial/ethnic groups with similar historical enrollment patterns, we use STS distance, proposed by Möller-Levet et al. (2003), to measure dissimilarity. STS considers short time series as piecewise linear functions and calculates the difference between the slopes of the functions. As a result, using STS distances, information about the relative change of different racial/ethnic groups can be included.

The clustering algorithm is also worth considering. Partitioning methods (e.g., K-means, C-means) are widely used for clustering purposes because they are computationally efficient and easy to implement. Because we want to use STS instead of the conventional Euclidean distance in a Euclidean space, we apply a K-medoids algorithm (Kaufman and Rousseeuw, 1990) that iterates over two phases as K-means ((1) finding a centroid for each cluster, (2) assigning data points to their closest centroid until some termination condition is met) but uses medoids, which are existing data points, as the centroids instead of means.

The major drawbacks of hard K-medoids are as follows: first, it is challenging to predefine the number of clusters and initial partitions due to a lack of efficient and universal methods (Fei et al., 2009); second, the algorithm might fail to recognize relatively small clusters (Kashef and Kamel, 2008). Because STS distance is sensitive to the magnitude of the time series and some racial/ethnic groups have much greater magnitudes than others (for example, see Figure 2), we want to avoid the case where the few racial/ethnic groups with large magnitudes are clustered together and the differences across groups with small magnitudes are ignored. One solution is to introduce the binary hierarchical K-medoids method as a variant of hard K-medoids. Our K-medoids algorithm first considers all the data in the same big cluster, and then separates one big cluster into two sub-clusters using K-medoids. The division is repeated at the sub-clusters until certain criteria are met. Then, we aggregate the disaggregate racial/ethnic groups in the same cluster, such that the similar time series according to our measurement can be aggregated together to improve the forecasting performance.

Furthermore, aggregation can affect not only forecast accuracy but also fairness, which is known as "Simpson's paradox" (Simpson, 1951). Given Events X, Y, and Z, even if we have $P(X|Y) < P(X|Y')$, it is possible that $P(X|YZ) \geq P(X|Y'Z)$. In a famous UC Berkeley case (Bickel et al., 1975), the aggregate data on graduate admission in the fall of 1973 show clear favoritism toward female applicants; however, the disaggregated data reveal that nearly as many departments favor male applicants as the ones that favor female applicants. Similarly, the College Factual website uses six race/ethnicity categories in general (Asian, White, Hispanic or Latino, Black or African American, and Other Races) (College Factual, 2021a). At the level of college majors, the categories are changed to "Asian, White, Hispanic or Latino, Black or African American, Other Races, and Nonresident Aliens" (College Factual, 2021b). For the major in mathematics and statistics, this new aggregation results in 178 "Nonresident Aliens". If the detailed information on "Nonresident

Aliens" is known, the aggregation result may be different. This category with the highest proportion can be confusing because of inconsistent and ambiguous categorization.

Therefore, we aim to examine how different race/ethnicity aggregation rules can affect fairness. Additionally, we can incorporate fairness into the aggregation method because modelders should not only analyze the fairness of the results but also try to improve fairness while modeling (Tsai et al., 2022). For both purposes, the proper metrics for fairness are required.

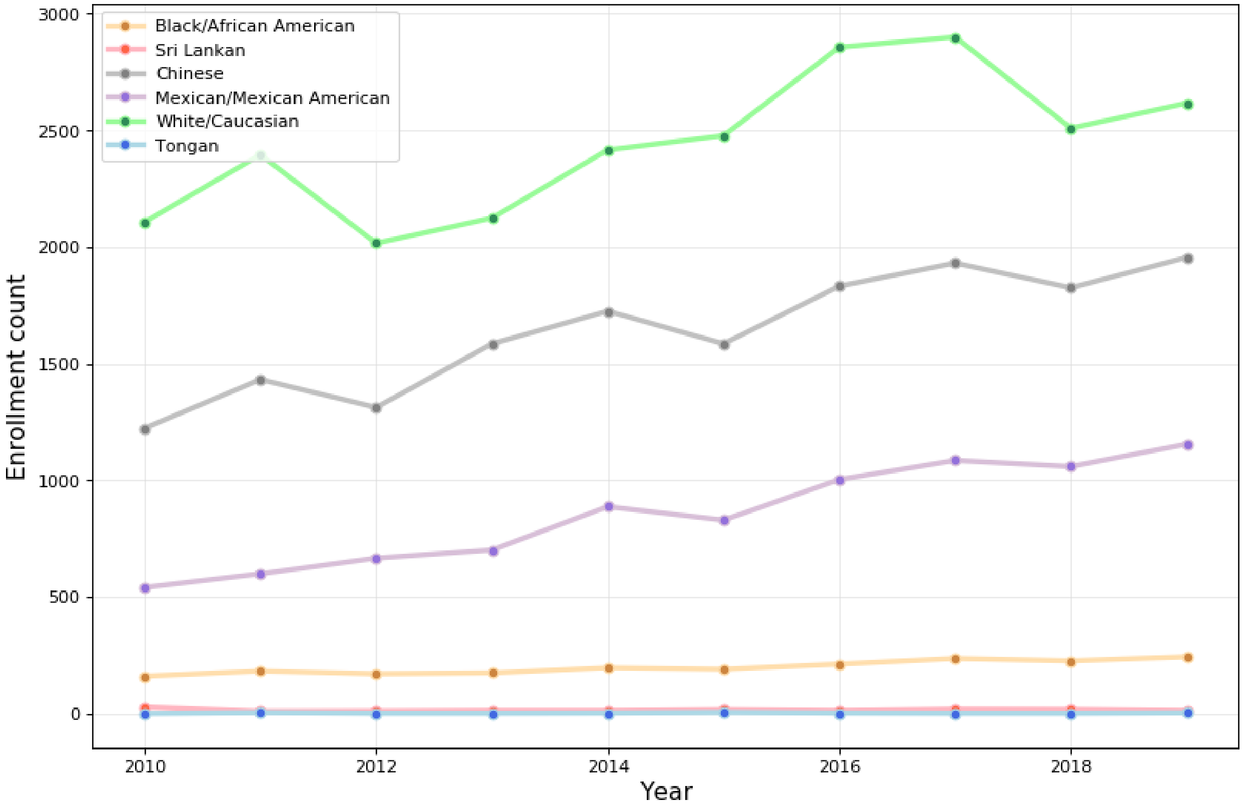

Figure 2: Undergraduate enrollment counts of six arbitrary races/ethnicities from 2010 to 2019 at UC Berkeley

2.4. Fairness metrics

We need appropriate fairness metrics for two reasons: enhance fairness during modeling and analyze the fairness of the results. For the first purpose, we can choose a quantitative metric to measure fairness and combine it with STS distances in the aggregation process. For the second purpose, it is also important to present a quantitative fairness metric because both a forecast of a phenomenon and its accuracy alongside the FSG metrics are important in FSG context (Rostami-Tabar et al., 2022). Furthermore, when analyzing the performance of a forecast model, different metrics may reveal different perspectives. For example, Duchemin and Matheus (2021) used multiple metrics for model comparisons, resulting in a different ranking of the model performance than using a single performance metric, such as accuracy. Forecasting results evaluated using fairness metrics combined with forecast accuracy can provide critical and more complete insights to institutions than using only forecast accuracy metrics.

Current fairness metrics can be categorized as either group or individual metrics. To compare fairness across racial/ethnic groups, we choose group fairness metrics. Commonly used fairness metrics for groups can be further divided into four major types: statistical parity, equal opportunity and equalized odds, balance for positive class and negative class, and predictive parity and calibration. For more information on these metrics, see Fu et al. (2020) and De-Arteaga et al. (2022).

These fairness metrics are designed for binary machine learning prediction problems only, and the use case across multiple groups instead of two groups remained ambiguous. To generalize the metrics to time series and forecasting problems, we use two fundamental concepts. The first is based on a generalized form of conditional disparity (Ritov et al., 2017), $P(\mathbf{x} \mid \mathbf{a} = a, \mathbf{z} = z) - P(\mathbf{x} \mid \mathbf{a} = a', \mathbf{z} = z)$, with different representations of the protected attributes $\mathcal{A}$ and the targeted outcome $\mathcal{Z}$. Because this metric can evaluate the fairness of any two racial/ethnic groups with respect to some other sensitive attributes (e.g., gender and residency), it can be adopted in both the modeling and output-analyzing stages. To evaluate the fairness of more than two groups, we can add the conditional disparity between all pairs of the groups and then calculate the average of the sum.

The second concept is to compare the model performance across groups. In the result analysis stage, the forecast accuracy across different racial/ethnic groups can be compared to determine whether the model treats different groups fairly.

2.5. Research gap

Unlike previous enrollment forecasting practices that mostly focused on the algorithms, we aim to investigate the impact of race/ethnicity in a forecasting scheme in terms of both forecast accuracy and fairness. Moreover, to the best of our knowledge, no previous aggregation and clustering approach has included fairness metrics. Our study proposes a new fair aggregation method based on STS, adopted fairness metrics for time series, and binary hierarchical K-medoids algorithm.

3. Empirical application

3.1. Data description and the hierarchical structure

The dataset was originally published by UC, Berkeley (University of California, 2021). It includes the yearly undergraduate student enrollment counts in the fall semester for 68 disaggregate racial/ethnic groups from 2010 to 2019. The disaggregate groups are categorized in aggregate categories as "African American, American Indian/Alaska Native, Asian, Hispanic Latinx, Native Hawaiian Pacific Islander, Southwest Asian North African, and White".

Because the original time series is too short, for convenience in the forecasting process, we generated synthetic time series with the same format and statistical properties with 25 observations using the synthetic data vault (SDV) (Patki et al., 2016). For records, 25 observations are reasonable, considering the number of historical data that institutions could obtain.

The data will be classified into total (institution), aggregate race/ethnicity, and disaggregate race/ethnicity levels, with the most disaggregate level consisting of 68 time series. We denote the disaggregate race/ethnicity, aggregate race/ethnicity, and total levels as l_2, l_1, and l_0, respectively. Figure 3 shows the three levels in a hierarchical structure based on the UC aggregation rule.

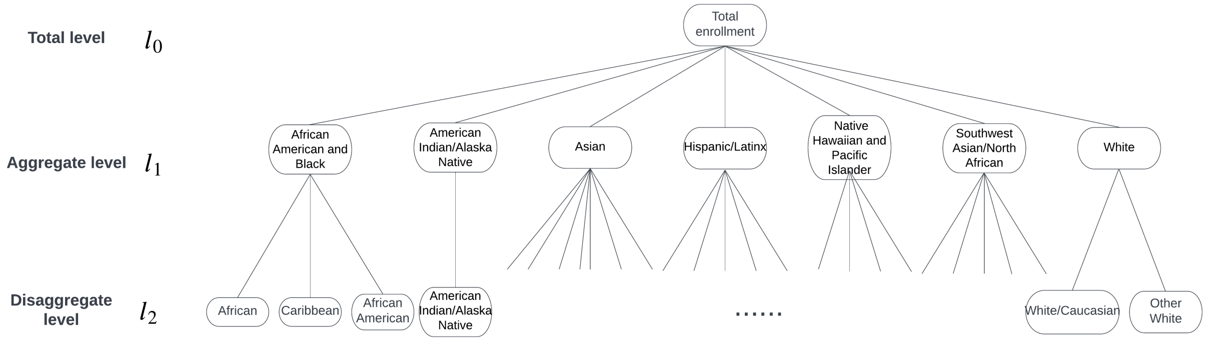

Figure 3: The hierarchical structure with UC aggregation rule

3.2. Forecasting models

Because we focus on the impact of aggregation levels rather than forecasting algorithms, given that detailed disaggregate categorization has recently become important and institutional enrollment records are usually short, we apply double exponential smoothing in our forecasting models (Brown and Meyer, 1961), which implies that recent observations have the most significant impact on the forecasts and considers the trend of historical data. Exponential smoothing is one of the most basic methods used in enrollment forecasting (for example, Hussar and Bailey, 2020; Bousnguar et al., 2022). The parameters α and β between 0 and 1 that generate the least mean absolute percentage error (MAPE) of the forecasting result are chosen as smoothing parameters for the forecasted series. In this study, we choose 5 years as the forecast horizon following previous literature. Thus, the training set has 20 observations and the test set has 5.

While such technique can be applied at different aggregation levels, we define three models: a disaggregate model, a standard aggregate model, and a fair aggregate model.

For consistency and convenience, we introduce the following notations, which will be used in subsequent sections:

Let vectors $Y_i^{l_a}$ represent the ith time series of enrollment at level a, where i is an integer group index and a is the level index, $i = 0, 1, ..., N_a$, $a = 0, 1, 2$, assuming that there are N_a groups at level a. Each vector has the entries $y_{i,t}^{l_a}$, where t is the year index. Correspondingly, $\hat{y}_{i,t+h}^{l_a}$ represents the h-step ahead forecast of the ith series at level a. When $a = 0$ (at the total level), we denote $y_{i,t}^{l_a}$ and $\hat{y}_{i,t+h}^{l_a}$ as $y_t^{l_0}$ and $\hat{y}_{t+h}^{l_0}$, respectively, because there is only one group at the total level.

Let $f_{t+h}(\alpha, \beta, y_{i,t}^{l_a})$ be the double exponential smoothing function, where h is the index of steps ahead and α and β are the smoothing parameters, $f_{t+h}(\alpha, \beta, y_{i,t}^{l_a}) : \mathbb{Z}^+ \to \mathbb{R}^+$. Specifically, $f_{t+h}(\alpha, \beta, y_{i,t}^{l_a}) = b_{i,t}^{l_a} + hc_{i,t}^{l_a}$, where $b_{i,t}^{l_a}$ represents the level equation and $c_{i,t}^{l_a}$ represents the trend equation, for the ith time series at level a. $b_{i,t}^{l_a} = \alpha y_{i,t}^{l_a} + (1-\alpha)(b_{i,t-1}^{l_a} + c_{i,t-1}^{l_a})$, $c_{i,t}^{l_a} = \beta(b_{i,t}^{l_a} - b_{i,t-1}^{l_a}) + (1-\beta)c_{i,t-1}^{l_a}$, where $0 \leq \alpha, \beta \leq 1$.

3.2.1. Disaggregation model

In the disaggregate model, we use only the data at the total and disaggregate levels. At the disaggregate level, the forecasting algorithm is applied separately for each racial/ethnic group; consequently, the smoothing parameters are chosen individually for each group.

As a result, the forecast at the total level does not cohere with the forecasts at the disaggregate level. This model serves as a benchmark for other models.

For the forecasts at the total level, $\hat{y}_{t+h}^{l_0} = f_{t+h}(\alpha, \beta, y_t^{l_0})$. For the forecast at the disaggregate level, $\hat{y}_{i,t+h}^{l_2} = f_{t+h}(\alpha, \beta, y_{i,t}^{l_2})$. There is no forecast available at the aggregate level because it is not defined in this model.

3.2.2. Standard aggregation model

In the standard aggregation model, we aggregate the disaggregate data into seven groups according to the standard categorization rule provided by UC. We apply double exponential smoothing on the aggregate level and then sum the forecasts to obtain the forecast at the total level. Meanwhile, the forecasts at the aggregate level are distributed for the forecasts at the disaggregate level based on the average historical proportions of the disaggregate series $Y_i^{l_2}$ in the corresponding aggregate group $Y_k^{l_1}$. In this case, we assume that the standard aggregation rule aggregates similar disaggregate groups together.

Denote the proportion of $Y_i^{l_2}$ in $Y_k^{l_1}$ as $p_{i,k}$, $p_{i,k} = \frac{1}{T}\sum_{t=1}^{T} y_{i,t}^{l_2}/y_{k,t}^{l_1}$

The forecast at the total level is $\hat{y}_{t+h}^{l_0} = \sum_k f_{t+h}(\alpha, \beta, y_{k,t}^{l_1})$, and the forecast of group i at the disaggregate level is $\hat{y}_{i,t+h}^{l_2} = p_{i,k} f_{t+h}(\alpha, \beta, y_{k,t}^{l_1})$, given that $Y_i^{l_2}$ belongs to $Y_k^{l_1}$.

3.2.3. Fair aggregation models

The forecasting process in fair aggregation model is similar to that in standard aggregation models, but the aggregate level is defined using different aggregation rules. The new criteria consist of both similarity and fairness, which are measured using STS and statistical disparity, respectively. In short, we perform binary hierarchical K-medoids clustering with the weighted linear combination of STS and statistical disparity as the pairwise distance between the disaggregate groups and then aggregate the disaggregate groups within each cluster to form an aggregate group.

To measure the dissimilarity between two time series in terms of shape, STS considers the short time series as piecewise linear functions and calculates the difference between the slopes of the functions in the same period. For the time series in the disaggregate level, we denote the STS between two series $Y_i^{l_2}, Y_j^{l_2}$ as $d_{i,j}$, where i, j are indexes of disaggregate groups, $i, j \in \mathbb{Z}^+, 1 \leq i, j \leq 68$. Let T be the number of available observations; then,

$$d_{i,j} = \sum_{t=0}^{T-1} |(y_{i,t+1}^{l_2} - y_{i,t}^{l_2}) - (y_{j,t+1}^{l_2} - y_{j,t}^{l_2})|, \tag{1}$$

Note that the time series are standardized for calculating STS in order to minimize the influence of scales.

Then, we define the statistical disparity between groups i and j. It makes little sense to consider only the difference between enrollment counts of different groups because such a difference can be perfectly valid because of differences in application numbers. For example, if group 1 has much more applicants than group 2, it would be normal that group 1 has more people enrolled than group 2, and we have insufficient information to address fairness. Thus, we shall consider some other attributes regarding specific concerns (e.g., gender, residency, and scholarship status) to see how differently the two groups are treated. For example, we can consider the difference between the probability of group 1 obtaining the scholarship and that of group 2.

In our study, we consider gender as the additional attribute for fairness. Specifically, we measure the difference in the proportion of female students enrolled between the two disaggregate racial/ethnic groups. Moreover, because the most recent data may have a greater influence on enrollment, we use the most recent observation to calculate fairness.

We denote being female as an attribute A, and define the statistical disparity between two series $Y_i^{l_2}, Y_j^{l_2}$ as $g_{i,j}$ as

$$g_{i,j} = |P(A|Y_i^{l_2}) - P(A|Y_j^{l_2})| = |a_{i,T}/y_{i,T}^{l_2} - a_{j,T}/y_{j,T}^{l_2}|, \tag{2}$$

where $a_{i,T}$ and $a_{j,T}$ are the numbers of females in groups i and j enrolled at time T, respectively.

We then combine $d_{i,j}$ (1) and $g_{i,j}$ (2) with a weight μ. The objective is to aggregate the disaggregate race/ethnicity groups that are similar in terms of distance and fairness, which correspond to small $d_{i,j}$ and small $g_{i,j}$. Let W be the dissimilarity matrix for the series whose (i,j)th entry $w_{i,j}$ is the dissimilarity (the clustering criteria) between i, j, which is then a linear combination of $d_{i,j}$ and $g_{i,j}$.

$$w_{i,j} = d_{i,j} + \mu g_{i,j}, \tag{3}$$

where $\mu \in \mathbb{R}_{\geq 0}$. The choice of μ depends on the distribution and scale of $d_{i,j}$ and $g_{i,j}$, as well as the purpose of modeling, i.e., greater μ indicates that the clustering is more based on fairness, the statistical disparity. The details of this matter will be discussed in Section 4.

As mentioned in Section 2, we utilize binary hierarchical K-medoids for clustering, which combines the hard K-medoids algorithm and hirarhical clustering process. Algorithm 1 illustrates the general hard K-medoids algorithm.

Algorithm 1: K-medoids algorithm

```
Input: a dataset $Y$ contains time series, dissimilarity matrix $W$, and the number of
       clusters $k$
Output: a set of $k$ clusters, $S = \{S_1, S_2, ..., S_k\}$ that minimizes the sum of dissimilarities
        between all series and their nearest medoids.
begin
    Arbitrarily choose $k$ series in $Y$ as the initial medoids;
    do
        Assign remaining series to the cluster with their nearest medoids;
        Compute the distance of all the series to their medoids (total distance) $D_{S_i}$ based
         on input $W$ for every cluster $S_i$, $i = 1, 2, ..., k$ ;
        for every medoid $m_j$, $j = 1, ..., k$ do
            for every other series $x_l$ in the same cluster, $i = 1, ..., |S_j|$ do
                Compute the total distance, $D_{S_j}^{Swap}$, assuming $m_j$ and $x_l$ are swapped;
            Select an $x$ that minimize the $D_{S_j}$;
            if $D_{S_j}^{Swap} \leq D_{S_j}$ then
                Swap the medoid $m_j$ with $x$;
                Update the total distance $D_{S_j}$;
    while Medoids do not change;
    return the cluster set $S$;
```

The binary hierarchical K-medoids performs K-medoids recursively. Initially, we consider all of the series in a single cluster. During the binary hierarchical clustering process, we divide the cluster into two sub-clusters and repeat the process until some stopping criteria is met, in which the relatively small clusters can be recognized. In this study, we choose the number of the series in the cluster as the stopping criteria because we would like to ensure that each cluster has a reasonable number of time series for the convenience of the forecasting process. For example, we may not need a cluster with too many time series because it may affect the forecast accuracy at the disaggregate level using the top-down approach. Algorithm 2 shows the details of the binary hierarchical K-medoids.

Algorithm 2: Binary hierarchical k-medoids algorithm

```
Input: a dataset $Y$ contains time series, dissimilarity matrix $W$, the greatest number of
       series $c$ every cluster can have
Output: a set of $k$ clusters, $S = \{S_1, S_2, ...\}$
begin
    Function binary_hierarchical_k_medoids($W$, $Y$, $c$):
        if $|Y| \leq c$ then
            return $[Y]$ ;
        else
            Define $W_{sub}$ as the submatrix of $W$ representing the dissimilarity matrix for the
              current cluster $Y$;
            Perform regular k-medoids (Algorithm 1) for $Y$ with $k = 2$ and dissimilarity
              matrix $W_{sub}$ to find two sub-clusters, denoted as $U_1$ and $U_2$;
            return binary_hierarchical_k_medoids($W$, $U_1$, $c$) append
              binary_hierarchical_k_medoids($W$, $U_2$, $c$) ;
    $S$ = binary_hierarchical_k_medoids($W$, $Y$, $c$)
```

Note that the binary hierarchical K-medoids cannot guarantee that the intracluster distance (the distance among members of a cluster) is minimized; however, it serves our purpose as we seek to find clusters with similar sizes without ignoring the relatively small differences in the criteria between disaggregate groups.

3.3. Evaluation Criteria

3.3.1. Forecast accuracy

Because the forecasts are obtained at different levels in all models, we evaluate the models (i.e., disaggregation, standard aggregation, and fair aggregation models) at each level. To evaluate forecast accuracy, we adopt MAPE for the forecasts at the total and aggregate levels (if available). The MAPE of series i at level a is defined as

$$MAPE_i^{l_a} = \frac{1}{h} \sum_{t=T}^{h} |\frac{y_{i,t}^{l_a} - \hat{y}_{i,t}^{l_a}}{y_{i,t}^{l_a}}|,$$

where $a = 0, 1$, represent the total and aggregate levels (if available), respectively.

MAPE is widely used, easy to interpret, and scale independent, which fits the enrollment data where the different racial/ethnic groups can be on very different scales. The forecast accuracy at level a is simply the average of the MAPEs (AvgMAPE) of all series at level a, which is defined as follows:

$$AvgMAPE^{l_a} = \frac{1}{i} \sum_{i} MAPE_i^{l_a}.$$

However, the AvgMAPE cannot be adopted at the disaggregate level because a few observations are zero, as well as some forecasted values. Since one way to prevent the problem is to use the sum of actual observations, which should not be zero, we use average weighted MAPE (AvgWMAPE) to measure the forecast accuracy at the disaggregate level. AvgWMAPE compares the sum of forecast errors over the forecast horizon and the sum of actual observations over the horizon to overcome the infinite error.

The weighted MAPE (WMAPE) of series i at the disaggregate level is defined as

$$WMAPE_i = \frac{\sum\limits_{t=T}^{h} |y_{i,t}^{l_2} - \hat{y}_{i,t}^{l_2}|}{\sum\limits_{t=T}^{h} |y_{i,t}^{l_2}|},$$

and AvgWMAPE is

$$AvgWMAPE = \frac{1}{i} \sum_i WMAPE_i$$

3.3.2. Fairness

Similarly to forecast accuracy, fairness will be evaluated at different levels (i.e., aggregate and disaggregate levels but not the total level because there is only one series at the total level). First, the statistical disparity (2) that was used in the fair aggregation model is applied once more. To evaluate fairness at one level, we take the average of the statistical disparity of all the pairs of the time series at that level. Following the previous notations, assuming there are N_a series at level a, $a = 1, 2$, and the attribute of being female is A, the average statistical disparity is:

$$\frac{\sum\limits_{1 \le i \le j \le N_a} g_{i,j}}{\binom{N_a}{2}} = \frac{\sum\limits_{1 \le i \le j \le N_a} |P(A|Y_i^{l_a}) - P(A|Y_j^{l_a})|}{\binom{N_a}{2}}$$

Besides the statistical disparity in some specific attributes, another perspective is the model performance for different racial/ethnic groups. To investigate fairness from this aspect, the differences of forecast accuracy across the groups at different levels are evaluated. Similarly to the average statistical disparity calculation, we consider the groups as pairs and obtain the difference of forecast accuracy for each pair. To evaluate the difference of MAPEs at the aggregate level, we take the average of the difference of MAPE between all the pairs. Then, the average difference of MAPEs at the aggregate level is:

$$\frac{\sum\limits_{1 \le i \le j \le N_1} |MAPE_i^{l_1} - MAPE_j^{l_1}|}{\binom{N_1}{2}}.$$

Similarly, the average difference of WMAPEs at the disaggregate level is defined as follows:

$$\frac{\sum\limits_{1 \le i \le j \le N_2} |WMAPE_i - WMAPE_j|}{\binom{N_2}{2}}.$$

We consider two groups are fairly treated by the model when the forecast accuracy of them are similar. Thus, smaller average difference of MAPEs or WMAPEs are, fairer the model is.

4. Analysis of results

4.1. Choice of μ

We let μ be a non-negative integer. As introduced in Section 3, the choice of μ depends on the distribution and scale of $d_{i,j}$ and $g_{i,j}$. To cover as many options as possible, we repeatedly increase μ by 1 in the fair aggregation models until the clustering results converge. Note that the small μ represents the aggregation is more based on similarity, and the greater μ represents the aggregation is more based on fairness. Respectively, the aggregation depends on only similarity when $\mu = 0$, and it depends on only fairness when μ is large enough to converge.

Figure 4 shows the number of different clustering results (compared to the previous result as μ increases) within the different ranges of μ. The clustering results converge at some large number; however, it appears to be very insensitive to large μ. In contrast, the models are more sensitive when μ is small. In general, the fair aggregation models are less sensitive to μ as it grows larger.

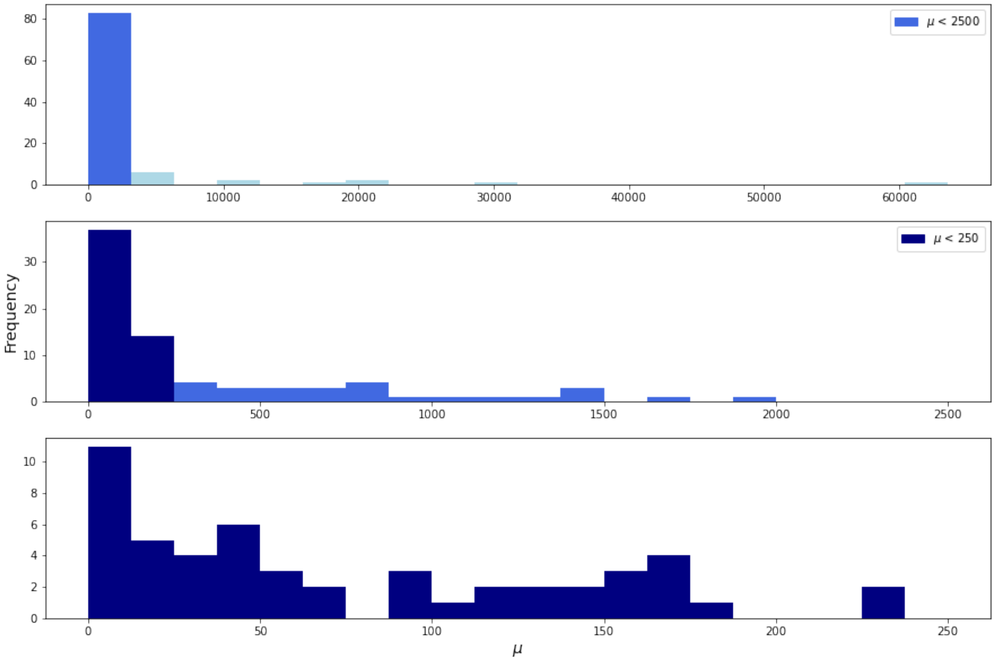

Figure 4: Number of different aggregation results in fair aggregation with different value of μ

The number of aggregate groups (clusters) in each different fair aggregation result is shown in

Figure 5. In this experiment, we set the maximum allowance of the number of disaggregate groups in each cluster as 15, then the amount of clusters generated by fair aggregation is within the range from 7 to 15 globally. The fair aggregation models that weight similarity more than fairness seem to generate more aggregate groups than those depends on fairness more. Such difference could be caused by the property of STS, which is sensitive to scales. Although we standardized the time series before calculating STS, the large differences of magnitudes of some disaggregate groups can still have significant effects, which may lead to sparse clustering results. Compared to STS, the statistical disparity could be more stable.

In our experiment, the aggregation results converge when μ approaches 63,577 while the minimum μ is 0. The fairness models with these two extreme values will be emphasized in the following sections.

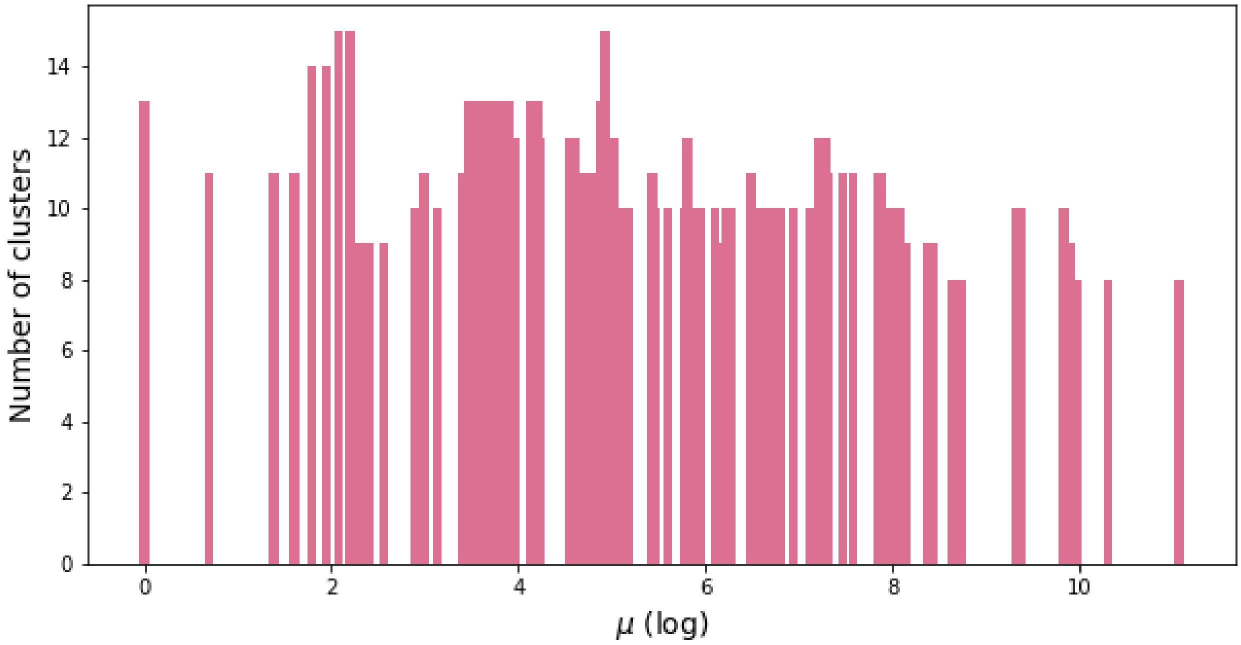

Figure 5: Number of clusters with different value of μ (log) in fair aggregation

4.2. Forecast accuracy

Regarding the models we proposed in the previous section (i.e., disaggregation model, standard aggregation model, and fair aggregation models), the forecast accuracy at the total and aggregate levels is calculated using MAPE and AvgMAPE, respectively, and the results are presented in Table 1, whereas the forecasting accuracy at the disaggregate level is calculated using AvgWMAPE, and the result is presented in Table 2.

First of all, the fair aggregation model with $\mu = 0$, indicating that we only consider the similarity based on STS but not fairness while aggregating, outperforms other models at all three levels, as we expected. This shows that in our case, the proposed hierarchical forecasting approach combined with the aggregation method can improve forecast performance, particularly at the aggregate level where aggregation occurs.

At the aggregate level, the fair aggregation model with $\mu = 0$ outperforms the standard aggregation model by approximately 55%, supporting our hypothesis that aggregating similar disaggregate racial/ethnic groups can improve forecast accuracy. The standard aggregation model outperforms the fair aggregation models with $\mu \geq 63,577$. A possible explanation of this phenomenon is that the standard aggregation model captures the similarity between the disaggregate groups better than fair aggregation models with $\mu \geq 63,577$.

Note that the standard aggregation model performs relatively well at all levels, implying that the standard aggregation of racial/ethnic groups in the enrollment context may be based on similarity; i.e., the historical enrollment of disaggregate groups belonging to the same aggregate group has similar patterns. As μ increases, the aggregation criteria shift to fairness rather than similarity, removing the advantage of aggregating similar series.

Although the fair aggregation model based on only STS has the best performance across all three levels, forecast accuracy at the total and disaggregate levels does not vary much as the model varied. We then investigate the relationship between forecast accuracy and μ at all three levels using fair aggregation models.

Model	Total level (MAPE)	Aggregate level (AvgMAPE)
Disaggregation model	0.027	N/A
Standard aggregation model	0.026	0.051
Fair aggregation model ($\mu = 0$)	**0.024**	**0.023**
Fair aggregation model ($\mu \geq 63,577$)	0.026	0.084

Table 1: Forecast accuracy at the total and aggregate levels for different models

Model	Disaggregate level (AvgWMAPE)
Disaggregation model	0.083
Standard aggregation model	0.079
Fair aggregation model ($\mu = 0$)	**0.078**
Fair aggregation model ($\mu \geq 63,577$)	0.081

Table 2: Forecast accuracy at the disaggregate level for different models

Figure 6 shows the relationship between the different values of μ and forecast accuracy at all three levels. For ease of presentation, the values of μ's on the x-axis are scaled on the basis of the number of different aggregation results. For example, when $29,762 < \mu \leq 63,577$, the aggregation results are the same, so the results are considered as one point on the plot. The blue and pink lines represent the forecast accuracy obtained from the benchmark models, which are the disaggregation and standard aggregation models, respectively.

It is shown that μ has the most significant impact at the aggregate level where the aggregation occurs. When μ is small, the forecasting performs well, as expected. Forecast accuracy decreases as μ increases, which is reasonable because the small μ represents greater emphasis on similarity than fairness during aggregation, and similarity is the most important factor in forecast accuracy. The forecast accuracy at the total and disaggregate levels are slightly affected by μ when compared with that at the aggregate level. At the total level, the forecast accuracy is very stable. At the disaggregate level, the forecast accuracy appears to be lower for large μ's.

We also notice that almost all fair aggregation models outperform the disaggregation model at the disaggregate level, and most of them outperform disaggregation models at the total level as well, even though the difference is not glaring at the total levels. We would argue that the advantage of aggregating the disaggregate data, which solves the problem of low-data quality and randomness at the disaggregate level, overcomes the disadvantages of information loss for the forecasts at the disaggregate level. It also could be exponential smoothing does not perform well for the disaggregate data. Other forecasting algorithms should be tested to verify this assumption.

At all levels, approximately half of the fair aggregation models outperform the standard aggregation model. We can conclude that the fair aggregation model combined with the hierarchical structure outperforms the disaggregate model and may outperform the standard aggregation model as well.

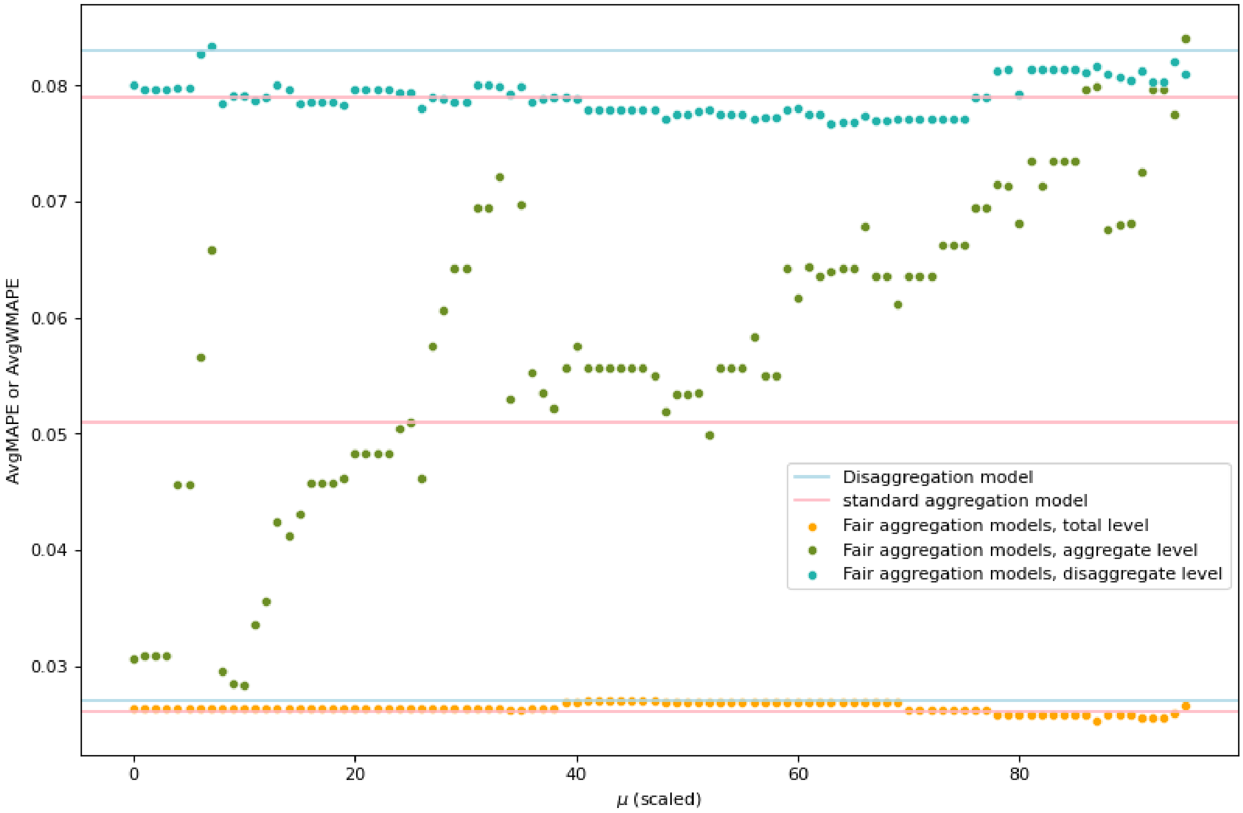

Figure 6: Fair aggregation parameter μ versus forecast accuracy at different levels

4.3. Fairness

4.3.1. Statistical disparity

In terms of fairness, we first evaluate the statistical disparity across the aggregate groups. Note that the smaller $g_{i,j}$ (statistical disparity) represents fairer groups; thus, the fair aggregation method aggregates the "fair" groups with relatively small $g_{i,j}$'s together. In this case, the statistical disparity between the aggregate groups (inter-aggregate fairness) is great whereas that within the aggregate groups (intra-aggregate distance) is small.

Model	Average statistical disparity between aggregate groups at the aggregate level
Standard aggregation model	0.153
Fair aggregation model ($\mu = 0$)	0.054
Fair aggregation model ($\mu \geq 63,577$)	0.404

Table 3: Fairness (average statistical disparity) at aggregate level for different models

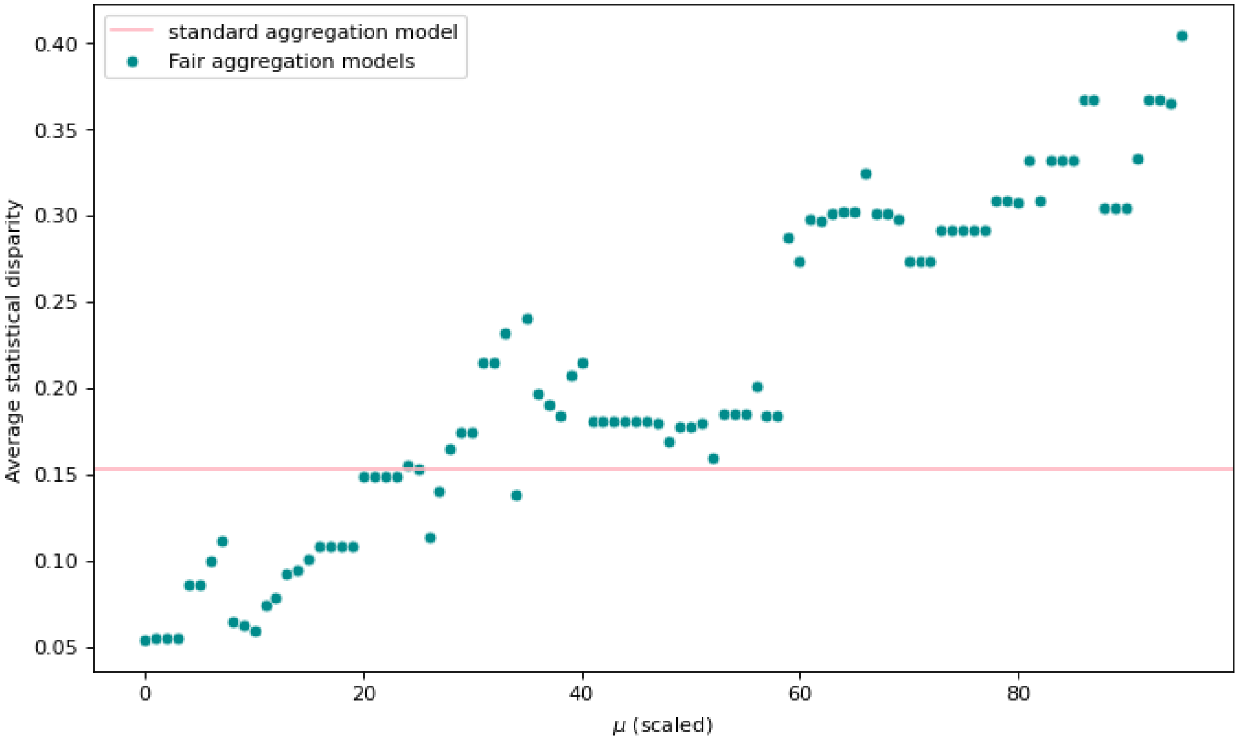

Figure 7: Fair aggregation parameter μ versus statistical disparity between aggregate groups

Table 3 and Figure 7 show how statistical disparity across aggregate groups changes along μ and how they compare with the standard aggregation. As expected, the statistical disparity increases as μ increases because the intra-aggregate statistical disparity is smaller and the inter-aggregate statistical disparity is larger. It shows that the statistical disparity can vary significantly on the basis of the different values of μ. Fairness aggregation models are more flexible and capable of presenting various pictures in terms of fairness than the standard aggregation model.

The fact that institutions can choose how to present fairness is an important indicator. According to the modeling results, it is easy to see what aggregation rule gives the best (or worst) statistical disparity. Moreover, similar groups in terms of fairness can be easily identified in the aggregation structure. This insightful information may be used by institutions to make plans.

For example, if we label all disaggregate racial/ethnic groups by indexes such as 1, 2, 3,...,68, and set the maximum number of groups in each cluster as 15, Figure 9 and Figure 8 shows the clustering (aggregation) tree obtained by fair aggregation with $\mu = 0$ and $\mu \geq 63,577$, respectively.

First, the institutions can choose their way to aggregate the disaggregate groups based on their purpose. For example, if the institution wants to present a fair picture, they can choose to aggregate the data based on the clustering tree with $\mu = 0$, where the inter-aggregate fairness is smaller. The analysis can start from the top to the bottom because the difference between the clusters is greater at the higher level than the lower level. Similarly, one can refer to Figure 8 for the aggregation rule with $\mu \geq 63577$ that leads to the greatest intra-aggregate fairness and the smallest inter-aggregate fairness.

Second, the relationships between the disaggregate groups are clearly shown in these clustering trees. One can find the most similar and the fairest groups, as well as the least similar and fair groups. For example, with respect to fairness, each of the four leaf nodes at the bottom level in Figure 8 includes the groups with small statistical disparity, and the two branches at the first level represents that the two groups are very different in terms of fairness. The institutions can analyze why the specific groups are fair/unfair and make appropriate plans targeting these groups to either increase or decrease the fairness. Similar analysis can also be performed with the focus of similarity or the combination of similarity and fairness by adjusting μ.

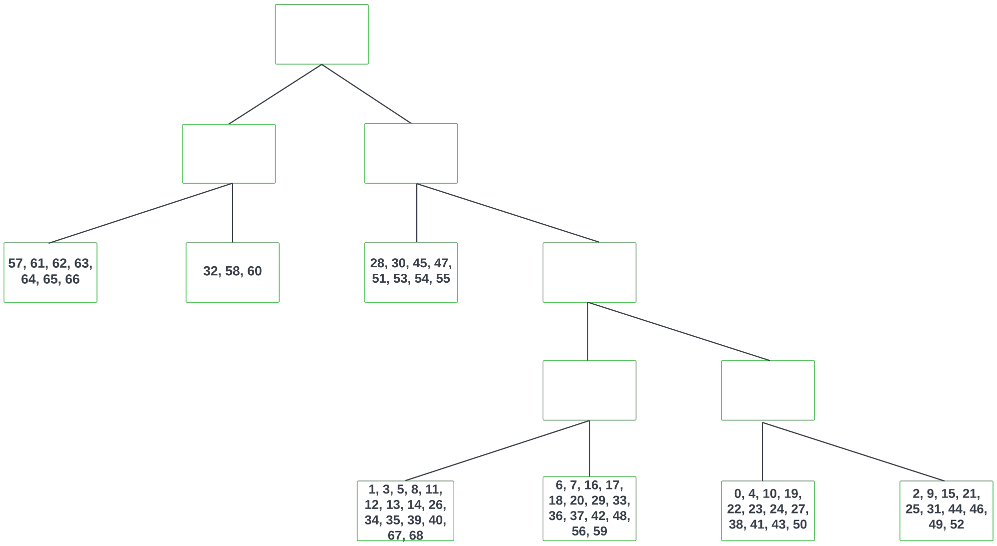

Figure 8: Clustering (aggregation) tree using fair aggregation with $\mu \geq 63,577$

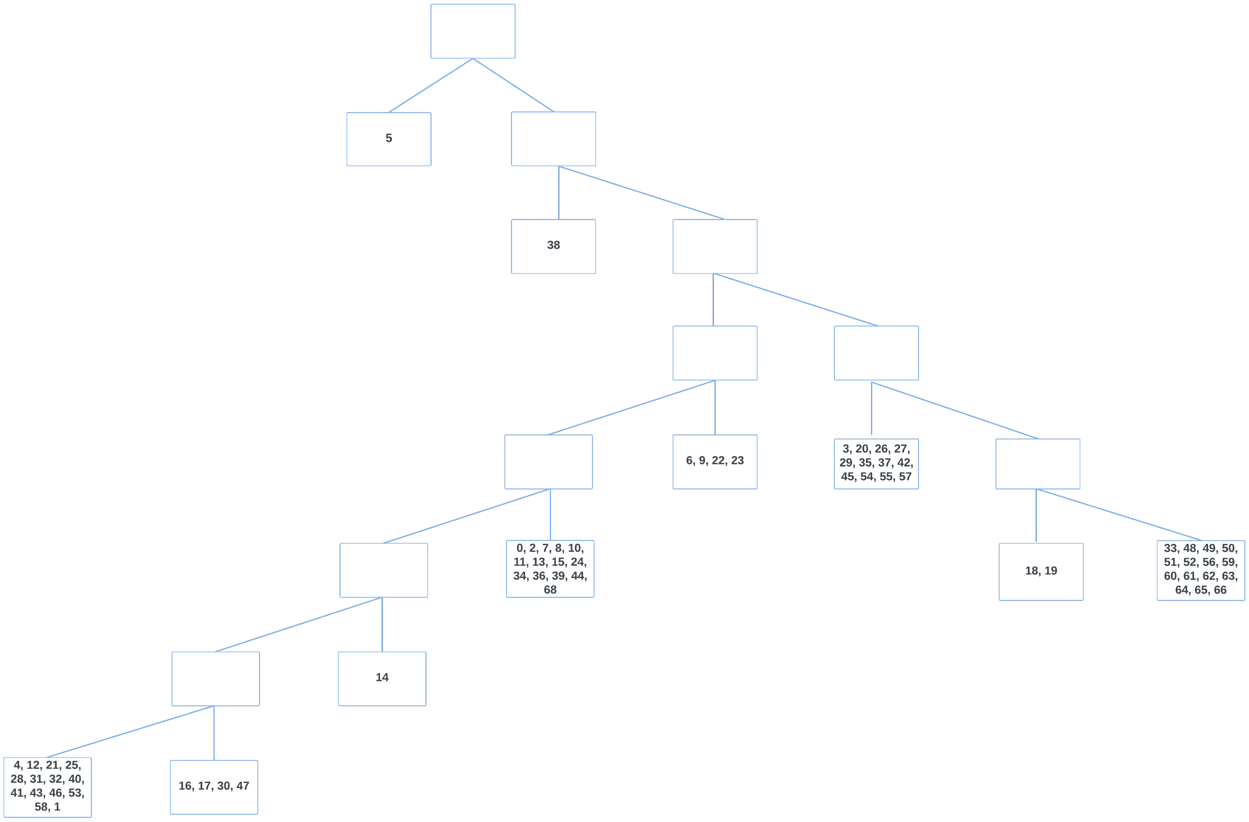

Figure 9: Clustering (aggregation) tree using fair aggregation with $\mu = 0$

Note that the clustering (aggregation) tree with $\mu \geq 63,577$ is more balanced than that with $\mu = 0$, which provides an evidence for our hypothesis in Section 4.1 that the statistical parity is more stable than STS in our analysis because STS is more sensitive to magnitudes.

In Figure 10, the relationship between average AvgMAPE and average statistical disparity at aggregate level is depicted. In general, the greater AvgMAPE is associated with greater statistical disparity, which corresponds to the less fair across aggregate groups. It meets our expectation because we would assume that the greater μ results more fairness within and less fairness across groups; therefore, the statistical disparity across groups increases as μ increases; meanwhile, as the aggregation weights more on fairness than similarity, the AvgMAPE increases as well. In this case, we can say that at the aggregate level, there is a trade-off between forecast accuracy and fairness at the aggregate level (i.e., forecast accuracy decreases as the fairness within groups increases). One can choose how to aggregate the disaggregate groups based on their specific needs.

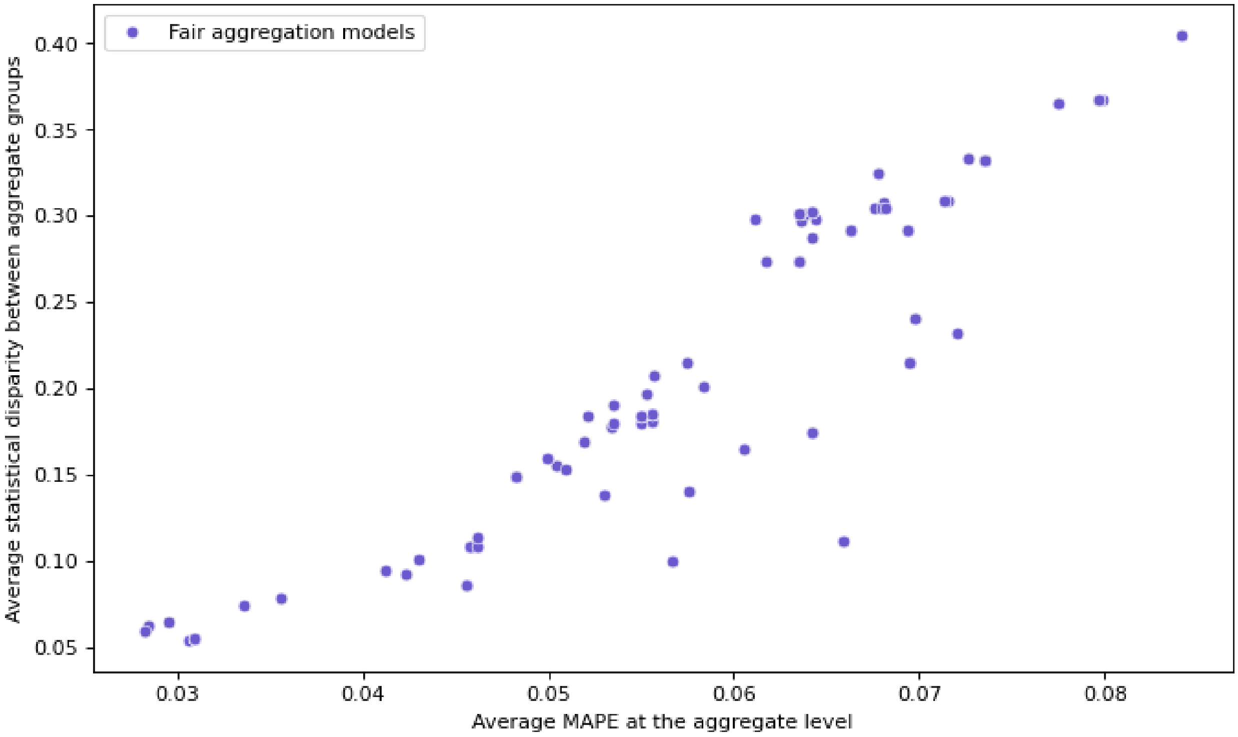

Figure 10: Average MAPE versus average statistical disparity at aggregate level

4.3.2. Difference of forecast accuracy

Figure 11 shows the relationship between μ's and the average difference of WMAPEs across aggregate groups at the aggregate level using fair aggregation models, where the values of μ's on the x-axis are scaled on the basis of the number of different aggregation results. The blue and pink lines represent the average difference of WMAPEs obtained from the disaggregation model and standard aggregation model, respectively. The average difference of WMAPEs tends to decrease when μ is relatively small, and increases after the scaled μ approaches 50, equivalent to μ approaches approximately 200. Compared to the disaggregation model and standard aggregation model, the fair aggregation models with relatively medium μ's treats the disaggregate groups more fairly and those with extreme μ's treats the disaggregate groups less fairly, although the difference of change is very slight in terms of magnitudes.

The average difference of MAPEs at the aggreagate level associated with fair aggregation models with different μ's are discribed in Figure 12, where the red line represents the average difference of MAPEs using standard aggregation model. As μ increases, the fair aggregation model treats the different groups less fairly. The trend is very similar to the trend in Figure 7, where the relationship between μ's and statistical parity is shown. We would argue that the greater μ leads to greater MAPEs, and the average difference across all aggregate groups increases is likely due to the increase of the magnitudes of all MAPEs at the aggregate level. Moreover, according to the average difference of MAPEs, the aggregate groups are treated more fairly by standard aggregation model than most fair aggregation models.

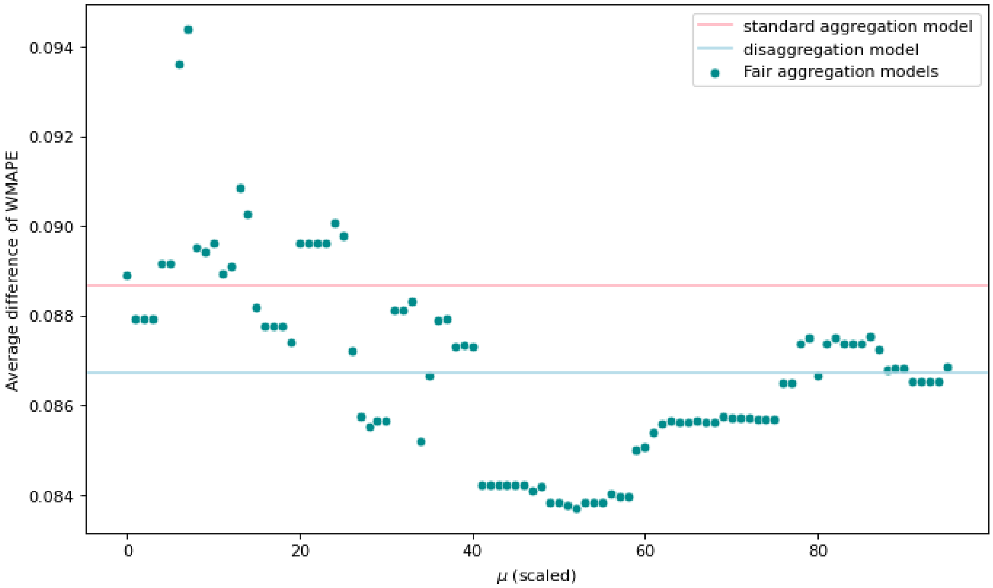

Figure 11: Aggregation parameter μ versus average differences of WMAPE at the disaggregate level

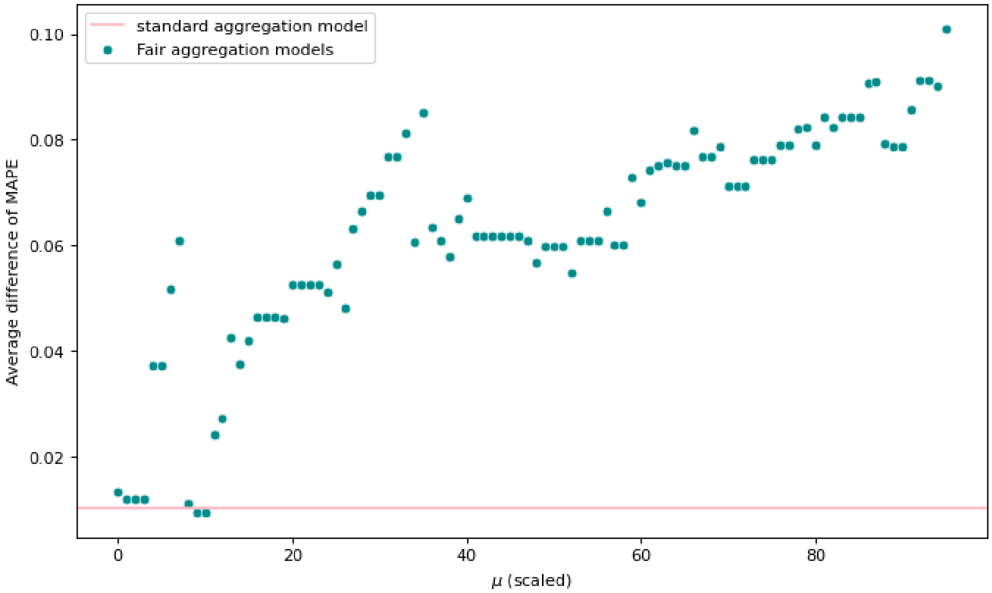

Figure 12: Aggregation parameter μ versus average differences of MAPE at the aggregate level

5. Conclusion and future work

5.1. Conclusions

As fairness has become an increasingly important social topic, it is essential to address it in a forecasting context. One prototypical example is enrollment forecasting, which is critical for institutions to plan policies, budgeting, staffing, and all enrollment-related activities. Considering that fairness across races/ethnicities in enrollment has been an important social topic, though there is a lack of research in a forecasting scheme, we use a hierarchical structure regarding race/ethnicity for enrollment forecasting, where we investigate the impact of the aggregation level on both forecast accuracy and fairness.

The aggregation levels are determined using the fair aggregation method, which is a binary hierarchical variant of K-medoid clustering that considers both similarity and fairness. We use STS to measure dissimilarity and statistical disparity to measure fairness. Alongside two benchmark models, we tested the new forecasting scheme with our flexible fairness aggregation models in terms of both forecast accuracy and fairness.

From the perspective of forecast accuracy, first, the hierarchical forecasting structure combined top-down and bottom-up approaches with an aggregate level can improve the forecast accuracy at all the levels compared to the disaggregate model. Second, fair aggregation results depend on its weight parameter for fairness. With some specifications of the parameter, fair aggregation can outperform standard aggregation model. Third, the forecast accuracy of fair aggregation model is sensitive to the parameter, especially at the aggregate level where the aggregation occurs. In general, at the aggregate level, the forecast accuracy decreases as the weight parameter for fairness increases. The hierarchical structure can be applied by institutions and enrollment forecasters to potentially improve the forecast accuracy.

In the aspect of fairness, the average statistical disparity across the aggregate groups at the aggregate level increases as the parameter increases, which corresponds to the decrease of the average statistical disparity within the groups. The fairness generated by fair aggregation models, in our case, statistical disparity, is very flexible and can be either smaller or greater than that generated by standard aggregation model. Moreover, the fair aggregation with different parameters can present a very detailed picture of similarity and/or fairness because the aggregation is shown in a hierarchical clustering structure.

For institutions, the flexibility of fairness representation indicates that they can present different fairness pictures by aggregating the disaggregate groups. Under some specific circumstances, institutions might want to show the data are fair/unfair, and fairness aggregation can be applied to direct such application. Furthermore, the fairness across and within groups can be clearly shown based on fair aggregation. The detailed information about fair/unfair groups can assist institutions in making future plans to achieve some specific fairness goal.

Fairness can also be evaluated as differences of forecast accuracy across groups. Our results showed that the fair aggregation models can treat the disaggregate groups either more or less fairly depending on the parameter. At the aggregate level, the standard aggregation model treats aggregate groups more fairly and the increase of the parameter is associated with the decrease of fairness in terms of forecast accuracy. Considering both statistical disparity and model treatment across groups as fairness indicator, we can conclude that there is a trade-off between fairness within aggregate groups and forecast accuracy. In the practice, the institutions and forecasters can adjust the parameter based on their needs in forecast accuracy and fairness.

In general, fair aggregation method can be applied in any practice that involves aggregation and fairness concerns. For example, the flexibility of fairness representation can be applied to show the strive businesses have made with diversity, equity and inclusion. Moreover, the fairness metrics can be adopted with different attributes rather than gender. The fair aggregation method itself is flexible, where the parameter, similarity, and fairness metrics, can all be further improved or adopted to suit the specific needs.

5.2. Limitations and future studies

Finally, we would like to discuss some drawbacks of this study and thus potential room for improvement.

First, the study may not be generally applicable because it is an investigation into a specific case. It would be beneficial to test different datasets, particularly enrollment data from different universities with different sizes, policies, categorization rules, etc. More generally, it would be interesting to see the fairness aggregation scheme used in other applications.

Second, regarding the aggregation method, both the clustering criteria and the aggregation algorithm can be investigated further. While we may consider another similarity/dissimilarity measurement rather than STS, the fairness metrics can be improved to have more desirable properties, such as the ability to include multiple attributes. Regarding the aggregation algorithm, the current binary hierarchical K-medoids algorithm can generate very sparse and unbalance tree because it is sensitive to the magnitudes of clustering criteria. Since we try to let each cluster has a reasonable size for forecasting, other algorithms that are likely generate clusters with similar sizes can be adopted. Moreover, the current algorithm tends to minimize the fairness within the cluster as the parameter increases. To have a more complete picture of how aggregate level affects fairness, other algorithm that can maximize the fairness within the cluster can be adopted.

Third, we only consider the difference of forecast accuracy across group as a fairness metric in the evaluation but not modeling stage. One may find a way to include this matter in the modeling stage.

Finally, other forecasting algorithms or techniques used in the literature, such as fuzzy time series and Markov chains, can be applied instead of exponential smoothing. Although the current experiment design focuses on how sequential aggregation would affect forecast performance, we can also compare the performance of different forecasting algorithms in this context.

References

Aghabozorgi, S., Shirkhorshidi, A. S., and Wah, T. Y. (2015). Time-series clustering–a decade review. *Information systems*, 53:16–38.

Anderson, N. and Svrluga, S. (2022). How is affirmative action used in college admissions?

Angwin, J., Larson, J., Mattu, S., and Kirchner, L. (2016). Machine bias. In *Ethics of Data and Analytics*, pages 254–264. Auerbach Publications.

Baker, R., Klasik, D., and Reardon, S. F. (2018). Race and stratification in college enrollment over time. *AERA Open*, 4(1):2332858417751896.

Bickel, P. J., Hammel, E. A., and O'Connell, J. W. (1975). Sex bias in graduate admissions: Data from berkeley: Measuring bias is harder than is usually assumed, and the evidence is sometimes contrary to expectation. *Science*, 187(4175):398–404.

Blyth, C. R. (1972). On simpson's paradox and the sure-thing principle. *Journal of the American Statistical Association*, 67(338):364–366.

Bousnguar, H., Najdi, L., and Battou, A. (2022). Forecasting approaches in a higher education setting. *Education and Information Technologies*, 27(2):1993–2011.

Brinkman, P. T. and McIntyre, C. (1997). Methods and techniques of enrollment forecasting. *New directions for institutional research*, 93:67–80.

Brown, R. G. and Meyer, R. F. (1961). The fundamental theorem of exponential smoothing. *Operations Research*, 9(5):673–685.

Buolamwini, J. and Gebru, T. (2018). Gender shades: Intersectional accuracy disparities in commercial gender classification. In *Conference on fairness, accountability and transparency*, pages 77–91. PMLR.

Chen, Y. A., Li, R., and Hagedorn, L. S. (2019). Undergraduate international student enrollment forecasting model: An application of time series analysis. *Journal of International Students*, 9(1):242–261.

College Factual (2021a). Statistics at University of California - Davis.

College Factual (2021b). Statistics at University of California - Davis, Department of Mathematics and Statistics.

De-Arteaga, M., Feuerriegel, S., and Saar-Tsechansky, M. (2022). Algorithmic fairness in business analytics: Directions for research and practice. *Production and Operations Management*, 31(10):3749–3770.

Duchemin, R. and Matheus, R. (2021). Forecasting customer churn: Comparing the performance of statistical methods on more than just accuracy. *Journal of Supply Chain Management Science*, 2(3-4):115–137.

Fei, H., Meskens, N., and Moreau, C.-H. (2009). Clustering of patient trajectories with an autostopped bisecting k-medoids algorithm. *IFAC Proceedings Volumes*, 42(4):355–360.

Flannery, R. (2022). American universities are losing chinese students to rivals: U.s.-china business forum.

Fu, R., Huang, Y., and Singh, P. V. (2020). Artificial intelligence and algorithmic bias: Source, detection, mitigation, and implications. In *Pushing the Boundaries: Frontiers in Impactful OR/OM Research*, pages 39–63. INFORMS.

Goehry, B., Goude, Y., Massart, P., and Poggi, J.-M. (2019). Aggregation of multi-scale experts for bottom-up load forecasting. *IEEE Transactions on Smart Grid*, 11(3):1895–1904.

Grip, R. S. (2009). Does projecting enrollments by race produce more accurate results in new jersey school districts? *Population research and policy review*, 28(6):747–771.

Hahn, H. W., Jenkins, R. W., and Paredes, R. A. (2015). Enrollment forecast 2015-2025 for texas colleges and universities.

Hussar, W. J. and Bailey, T. M. (2019). Projections of education statistics to 2027. nces 2019-001. *National Center for Education Statistics*.

Hussar, W. J. and Bailey, T. M. (2020). Projections of education statistics to 2028. nces 2020-024. *National Center for Education Statistics*.

Kahn, K. B. (1998). Revisiting top-down versus bottom-up forecasting. *The Journal of Business Forecasting*, 17(2):14.

Kashef, R. and Kamel, M. S. (2008). Efficient bisecting k-medoids and its application in gene expression analysis. In *International Conference Image Analysis and Recognition*, pages 423–434. Springer.

Kaufman, L. and Rousseeuw, P. J. (1990). Partitioning around medoids (program pam). *Finding groups in data: an introduction to cluster analysis*, 344:68–125.

Lapide, L. (2006). Top-down & bottom-up forecasting in s&op. *The Journal of Business Forecasting*, 25(2):14–16.

Liao, T. W. (2005). Clustering of time series data—a survey. *Pattern recognition*, 38(11):1857–1874.

Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., and Galstyan, A. (2021). A survey on bias and fairness in machine learning. *ACM Computing Surveys (CSUR)*, 54(6):1–35.

Mirowski, P., Chen, S., Ho, T. K., and Yu, C.-N. (2014). Demand forecasting in smart grids. *Bell Labs technical journal*, 18(4):135–158.

Möller-Levet, C. S., Klawonn, F., Cho, K.-H., and Wolkenhauer, O. (2003). Fuzzy clustering of short time-series and unevenly distributed sampling points. In *International symposium on intelligent data analysis*, pages 330–340. Springer.

Patki, N., Wedge, R., and Veeramachaneni, K. (2016). The synthetic data vault. In *2016 IEEE International Conference on Data Science and Advanced Analytics (DSAA)*, pages 399–410.

Ritov, Y., Sun, Y., and Zhao, R. (2017). On conditional parity as a notion of non-discrimination in machine learning. *arXiv preprint arXiv:1706.08519*.

Rostami-Tabar, B., Ali, M. M., Hong, T., Hyndman, R. J., Porter, M. D., and Syntetos, A. (2022). Forecasting for social good. *International Journal of Forecasting*, 38(3):1245–1257.

Simpson, E. H. (1951). The interpretation of interaction in contingency tables. *Journal of the Royal Statistical Society: Series B (Methodological)*, 13(2):238–241.

Song, Q. and Chissom, B. S. (1993). Forecasting enrollments with fuzzy time series—part i. *Fuzzy sets and systems*, 54(1):1–9.

Stallings, R. and Samanta, B. (2014). Prediction of university enrollment using computational intelligence. In *2014 IEEE Symposium on Swarm Intelligence*, pages 1–8. IEEE.

Torche, F. (2011). Is a college degree still the great equalizer? intergenerational mobility across levels of schooling in the united states. *American journal of sociology*, 117(3):763–807.

Tsai, T. C., Arik, S., Jacobson, B. H., Yoon, J., Yoder, N., Sava, D., Mitchell, M., Graham, G., and Pfister, T. (2022). Algorithmic fairness in pandemic forecasting: lessons from covid-19. *NPJ digital medicine*, 5(1):1–6.

University of California (2021). Disaggregated data.

Walczak, S. and Sincich, T. (1999). A comparative analysis of regression and neural networks for university admissions. *Information Sciences*, 119(1-2):1–20.

Weiler, W. C. (1987). An application of the nested multinomial logit model to enrollment choice behavior. *Research in Higher Education*, 27(3):273–282.

Yang, S., Chen, H.-C., Chen, W.-C., and Yang, C.-H. (2020). Student enrollment and teacher statistics forecasting based on time-series analysis. *Computational intelligence and neuroscience*, 2020.

Zhang, M. (2022). Google photos tags two african-americans as gorillas through facial recognition software.

Zotteri, G., Kalchschmidt, M., and Caniato, F. (2005). The impact of aggregation level on forecasting performance. *International Journal of Production Economics*, 93:479–491.