FACROC: A Fairness Measure for Fair Clustering Through ROC Curves

Tai Le Quy[1]✉, Long Le Thanh[2], Lan Luong Thi Hong[3], and Frank Hopfgartner[1,4]

[1] University of Koblenz, Germany
tailequy@uni-koblenz.de
[2] Hanoi University of Science and Technology, Vietnam
attentionocr@gmail.com
[3] Hanoi University of Industry, Vietnam
lanlhbk@haui.edu.vn
[4] University of Sheffield, United Kingdom
hopfgartner@uni-koblenz.de

Abstract. Fair clustering has attracted remarkable attention from the research community. Many fairness measures for clustering have been proposed; however, they do not take into account the clustering quality w.r.t. the values of the protected attribute. In this paper, we introduce a new visual-based fairness measure for fair clustering through ROC curves, namely FACROC. This fairness measure employs AUCC as a measure of clustering quality and then computes the difference in the corresponding ROC curves for each value of the protected attribute. Experimental results on several popular datasets for fairness-aware machine learning and well-known (fair) clustering models show that FACROC is a beneficial method for visually evaluating the fairness of clustering models.

Keywords: clustering · fair clustering · fairness measure · ROC curve · fairness-aware datasets.

1 Introduction

Clustering is a fundamental problem in unsupervised learning, and fairness in clustering has garnered significant attention within the machine learning (ML) community, starting with the foundational work of Chierichetti et al. [7]. Fair clustering techniques aim to ensure equitable representation or treatment of groups or individuals within clusters. Researchers have focused on two main challenging problems in fair clustering: defining and enforcing fairness constraints [6]. Hence, a number of fairness notions and techniques were introduced to ensure fairness constraints in clustering and can be applied in many domains, such as healthcare [4], education [13], etc.

There are more than 20 fairness notions in fair clustering [6]. Since fairness notions can be turned into measures [17], we will use the terms "fairness notion" and “fairness measure” interchangeably. Fairness notions are defined based on the group-level, individual-level, algorithm agnostic, and algorithm specific. At the group-level, the algorithms should not discriminate against or unfairly favor any group of individuals in the predictions. For instance, *balance* notion [7], the most popular fairness notion used for fair clustering, requires a ratio balance between the protected group, e.g., female, and the non-protected group, e.g., male. Unlike group-level fairness, individual-level fairness ensures that similar individuals are treated similarly by the clustering model. The fairness notion of *proportionality* for centroid clustering [5] is an example of individual-level fairness. However, to our knowledge, all the defined fairness notions do not consider the clustering quality w.r.t. values of the protected attribute.

Regarding the clustering quality, it is determined by several measures, such as the silhouette coefficient, sum of squared error (SSE), Dunn index (DI), and others [13,15]. However, most of these metrics are difficult to visualize for comparison between clustering models. To address this issue, AUCC (Area Under the Curve for Clustering) [10] has been introduced as a measure of clustering quality. This is a visual-based measure that utilizes the Receiver Operating Characteristics (ROC) analysis of clustering results. In addition, ABROCA, a visual-based fairness measure, has been proposed for the classification problem [8]. This method evaluates the fairness of classification models through slicing analysis based on the ROC curves. ABROCA measures the absolute value of the area between the ROC curves of the protected and non-protected groups.

To this end, we propose a new fairness measure for FAir Clustering through ROC curves (shortly: FACROC) which takes into account the clustering quality on each value of the protected attribute. In particular, we use AUCC to measure the clustering quality and then compute the FACROC by the deviation between the ROC curves corresponding to each value of the protected attribute. Afterward, we perform the experiments on five popular datasets used in fairness-aware ML with three prevalent fair clustering models to evaluate the performance of FACROC versus other popular fairness measures.

The rest of our paper is structured as follows: Section 2 overviews the related work. The computation of the FACROC measure is described in Section 3. Section 4 presents the details of our experiments on various datasets and clustering models. Finally, the conclusion and outlook are summarized in Section 5.

2 Related work

In this work, we focus on two main types of fairness measures for clustering including group-level fairness and individual-level fairness. Regarding group-level, *balance* is the first group fairness measure introduced by [7]. Next, the *bounded representation* measure was proposed with the aim of reducing imbalances in cluster representations of protected attributes [1]. This measure was generalized with two parameters α and β in the study of [3]. Subsequently, *social fairness* notion was introduced by [9], which aims to provide equitable costs for different clusters. Additionally, *diversity-aware* fairness was initiated by [16] which ensures a minimum number of cluster centers in the clustering are selected from each group. Based on the summarization task, the *fair summaries* measure was used to ensure that the data summary for each group is represented equally [12].

In contrast to group-level fairness, individual-level fairness notions aim to ensure that similar individuals are treated similarly. For example, Chen et al. [5] proposed the individual-level fairness notion of proportionality for centroid clustering to guarantee that points are treated equally. An individual-level concept that establishes a fair radius for clusters in center-based clustering objectives was presented by Jung et al. [11]. Chakrabarti et al. [4] provided algorithms for the k-center objective and proposed the idea of individual fairness, which guarantees that points receive comparable quality of service.

However, unlike other fairness measures for clustering based on the representation of protected attribute values in clusters, we propose a new visual-based fairness measure that takes into account the difference in the clustering performance for each group w.r.t. the protected attribute.

3 FACROC: a fairness measure for fair clustering

Given a dataset $\mathcal{X} = \{x_1, \ldots, x_n\}$ with n data points and a clustering $\mathcal{C} = \{C_1, C_2, ..., C_k\}$ with k clusters, the corresponding AUCC value is computed with the following steps [10]:

1. Compute a similarity matrix of the objects in the original dataset.
2. Acquire two arrays that present the pairwise relationship for each pair of objects:
   (a) Similarity: get from the similarity matrix.
   (b) Clustering: 1 if the pair belongs to the same cluster; otherwise, 0.
3. AUCC is obtained from the ROC analysis procedure with two above arrays, in which pairwise clustering memberships correspond to the "true classes" concept in classification.

Given a binary protected attribute $P = \{p, \bar{p}\}$; e.g., gender={female, male}, we inherit the concept of fairness from the ABROCA fairness measure for classification [8]: "*equal model performance across subgroups*". To this end, FACROC is defined as the absolute value of the area between the protected (ROC_p) and non-protected group ($ROC_{\overline{p}}$) curves across all possible thresholds $t \in [0, 1]$ of False Positive Rate (FPR) and True Positive Rate (TPR). The absolute difference between the two curves is measured to capture the case that the curves may cross each other.

$$\int_0^1 \mid ROC_p(t) - ROC_{\overline{p}}(t) \mid \, dt. \tag{1}$$

The value range: $FACROC \in [0, 1]$. The lower value indicates a lower difference in the clustering quality between the two groups and, therefore, a fairer model. Fig. 1 visualizes the FACROC slice plot on the COMPAS dataset.

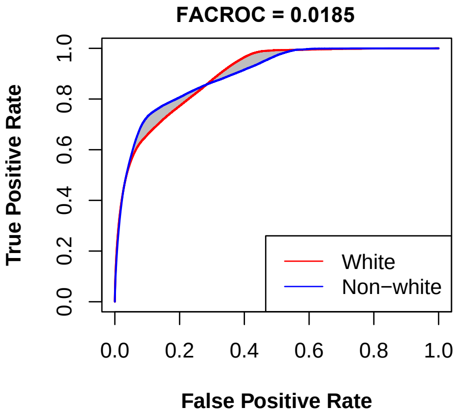


Fig. 1: An example of the FACROC slice plot.

4 Evaluation

In this section, we evaluate the performance of (fair) clustering models on well-known fairness measures for clustering and our proposed FACROC measure on prevalent fairness-aware datasets.

4.1 Datasets

We perform the experiments on five popular datasets for fairness-aware ML [14], as summarized in Table 1[5].

Table 1: An overview of the datasets

Datasets	#Instances	#Instances (cleaned)	#Attributes	Protected attribute	k*
Adult	48,842	45,222	15	Gender (F: 14,695; M: 30,527)	2
COMPAS	4,743	4,020	51	Race (NW: 2,561; W: 1,459)	7
Credit card	30,000	30,000	24	Gender (F: 18,112; M: 11,888)	2
German credit	1000	1000	21	Gender (F: 310; M:690)	2
Student-Mat.	395	395	33	Gender (F: 208, M: 187)	9
Student-Por.	649	649	33	Gender (F: 383; M: 266)	9

In particular, the ***Adult dataset***[6] is one of the most prevalent datasets for fairness-aware ML research. The class attribute (whether the income is greater than 50,000$) is removed because this study uses the dataset for clustering tasks. The ***COMPAS dataset***[7] is used for crime recidivism risk prediction. We convert the protected attribute *Race* into a binary attribute with values *{White, Non-White}*. We remove datetime attributes (*compas_ screening_ date,*

[5] Abbreviations: F (Female), M (Male), W (White), NW (Non-White)

[6] `https://archive.ics.uci.edu/dataset/2/adult`

[7] `https://github.com/propublica/compas-analysis`

dob, etc.), defendants' name/ID, and the class label *two years recidivism*. The ***Credit card clients dataset***[8] contains information about 30,000 customers in Taiwan in October 2005. The prediction task is to forecast whether a customer will default in the next month. We remove the class label for clustering. The ***German credit dataset***[9] contains information about 1000 customers, with the goal of predicting whether a customer has good or bad credit. We also eliminate this class label in our experiments. The ***Student performance dataset***[10] details students' academic performance in secondary education at two Portuguese schools in the 2005–2006 school year, covering two subsets: Mathematics (shortly: Student-Mat.) and Portuguese (shortly: Student-Por.).

4.2 Experimental setups

Clustering models. We evaluate the performance of traditional k-means and hierarchical clustering, as well as well-known fair clustering models.

- **Fair clustering through fairlets** [7] (shortly: Fairlet): This is the first work on fair clustering at the group-level to ensure an equal representation of each value of the protected attribute in each cluster. A two-phase approach was proposed: 1) Fairlet decomposition: grouping all instances into "fairlets" which are small clusters that satisfy the fairness constraint; 2) Clustering on fairlets: applying standard clustering methods, such as k-center, k-median, to these fairlets to produce the final fair clusters.
- **Scalable fair clustering** [2] (shortly: Scalable): This is an extended investigation of the fair k-median clustering problem [7], with a new practical approximate fairlet decomposition algorithm that runs in nearly linear time. Therefore, this proposed approach can be applied to large datasets.
- **Proportionally fair clustering** [5] (shortly: Proportionally): The authors define proportional fairness: any group of n/k points should have the right to form their cluster if there exists a center closer for all n/k points. The goal is to find clustering where no subset of points has a justified complaint about their assigned cluster without assuming predefined protected groups.

Fairness measures. We compare the proposed FACROC measure with the following well-known fairness measures:

- **Balance** [7]: Given a clustering $\mathcal{C} = \{C_1, C_2, ..., C_k\}$ with k clusters, $\mathcal{X}$ be a set of data points, $P = \{p, \bar{p}\}$ be the protected attribute, $\psi : \mathcal{X} \to P$ denotes the demographic group to which the data point belongs, i.e., male or female, the balance of clustering $\mathcal{C}$ is computed by:

$$balance(\mathcal{C}) = min_{i=1}^{k} balance(C_i). \tag{2}$$

where:

$$balance(C_i) = \min\left(\frac{|\{x \in C_i \mid \psi(x) = p\}|}{|\{x \in C_i \mid \psi(x) = \bar{p}\}|}, \frac{|\{x \in C_i \mid \psi(x) = \bar{p}\}|}{|\{x \in C_i \mid \psi(x) = p\}|}\right). \tag{3}$$

[8] `https://archive.ics.uci.edu/ml/datasets/default+of+credit+card+clients`

[9] `https://archive.ics.uci.edu/dataset/144/statlog+german+credit+data`

[10] `https://archive.ics.uci.edu/dataset/320/student+performance`

- **Proportionality** [5]: Given $\mathcal{X}$ be a set of n data points, $\mathcal{Y}$ be a set of feasible cluster centers m and a number $\rho > 1$, we call a clustering $\mathcal{C} \subseteq \mathcal{Y}$ ($|\mathcal{C}| = k$) is *ρ-proportional* if $\forall S \subseteq \mathcal{X}$ with $|S| \geq \lceil \frac{n}{k} \rceil$ and for all $y \in \mathcal{Y}$, there exists $i \in S$ with $\rho \cdot d(i, y) \geq D_i(\mathcal{C})$, where $d(i, y)$ is the distance between points i and y, and $D_i(\mathcal{C}) = min_{x \in \mathcal{C}} d(i, x)$. When $\rho = 1$, we call this proportional fairness.

Parameter selection. We denote the optimal number of clusters as k^*, which is determined based on the AUCC values using the implementation provided by [10]. Fig. 2 illustrates the value of k^* for the COMPAS dataset; other values of k^* are presented in Table 1. The minimum balance threshold is set to 0.4 because the minimum balance score of all datasets is 0.4493 (corresponding to the German credit dataset).

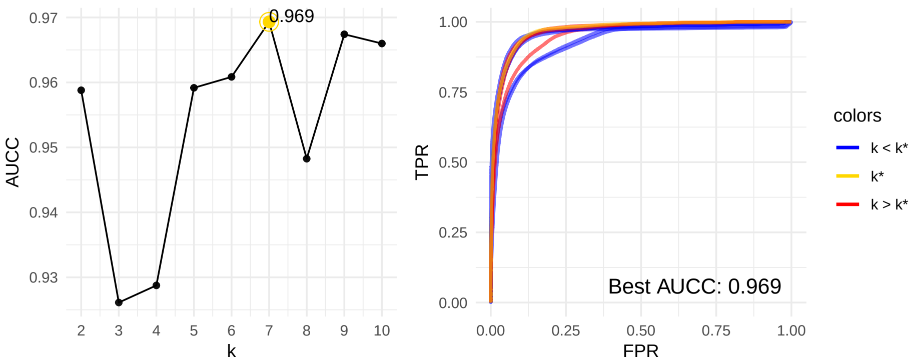


Fig. 2: COMPAS: selecting the optimal number of cluster (k^*) by AUCC.

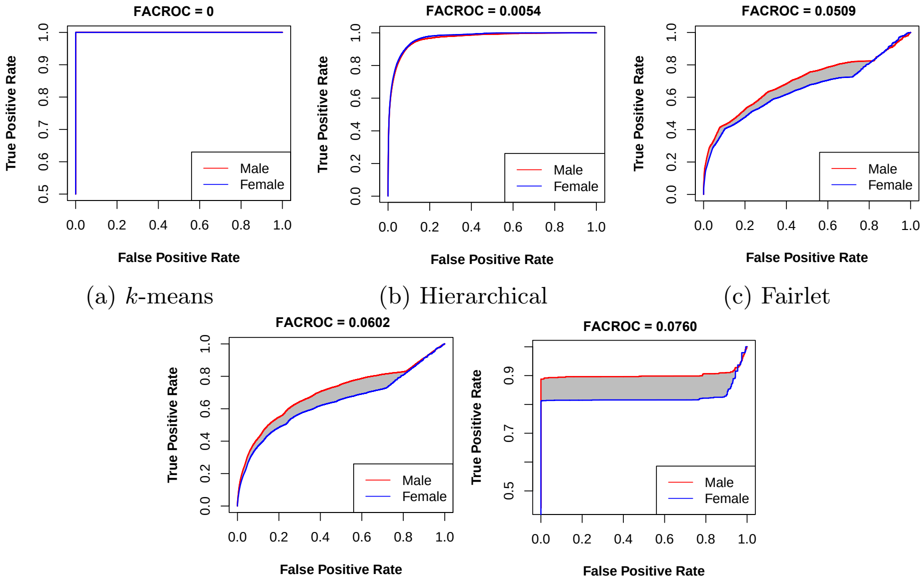
4.3 Experimental results

We present the results of clustering models regarding performance (Silhouette coefficient and AUCC) and fairness (Balance, Proportionality, and FACROC). The source code is available at `https://github.com/tailequy/FACROC`.

Adult dataset. The performance of the clustering models is shown in Table 2 and Fig. 3, with the best results highlighted in **bold**. In terms of clustering performance, obviously, k-means outperforms fair clustering methods. However, regarding fairness constraint, fair clustering models outperform k-means w.r.t. the fairness measure that they optimize. In detail, *Fairlet* and *Scalable* have better balance scores, while the *Proportionally* model is better in terms of the proportionality measure. This is explained by the fact that the definition of fairness is different in all models. Interestingly, the FACROC value of k-means is perfect, while the *Proportionally* model shows the worst result, i.e., the performances of observed fair clustering models are biased toward groups of people.

Table 2: Adult: performance of (fair) clustering models

Measures	k-means	Hierarchical	Fairlet	Scalable	Proportionally
Silhouette coefficient	**0.9861**	**0.9861**	0.4062	0.4377	0.3711
AUCC	**1.0000**	0.9998	0.6607	0.6569	0.8503
Balance	0.1684	0.1926	**0.5001**	0.4396	0.2966
Proportionally	1.0000	1.3274	1.4701	1.5994	**1.6321**
FACROC	**0.0000**	0.0054	0.0509	0.0602	0.0760



(a) k-means (b) Hierarchical (c) Fairlet

(d) Scalable (e) Proportionally

Fig. 3: Adult: FACROC slice plots.

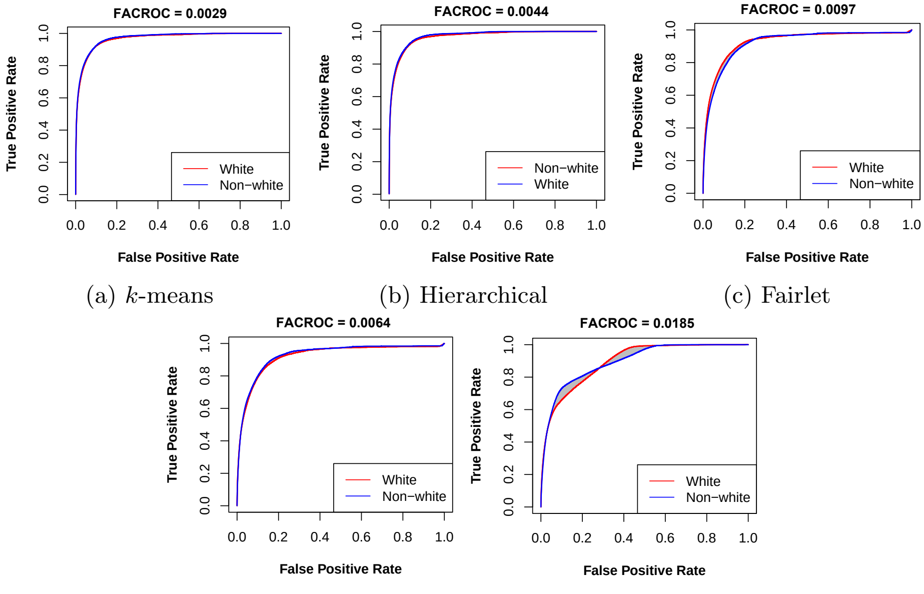
COMPAS dataset. An interesting trend is also observed in the results of the clustering models in the COMPAS data set (Table 3 and Fig. 4) when the FACROC of the k-means algorithm is significantly better than that of other models. This can be attributed to fair clustering models prioritizing fairness constraints, which can result in variations in clustering quality among groups.

Credit card clients dataset. Compared with fair clustering models, k-means is still the model with the highest clustering quality (Table 4 and Fig. 5). However, in terms of the FACROC measure, *Propotionally* and k-means share the top rank with a very low value.

German credit. In this dataset, a similar trend is observed in Table 5 and Fig. 6. k-means achieves the best results according to the FACROC measure, while fair clustering models have better results on the fairness measure they are designed to optimize.

Table 3: COMPAS: performance of (fair) clustering models

Measures	k-means	Hierarchical	Fairlet	Scalable	Proportionally
Silhouette coefficient	**0.6110**	0.6082	0.3827	0.3583	0.3903
AUCC	**0.9690**	0.9689	0.9233	0.9236	0.8933
Balance	0.0000	0.1017	0.4186	**0.4291**	0.3226
Proportionality	0.8554	1.1728	0.9443	1.0937	**1.2861**
FACROC	**0.0029**	0.0044	0.0097	0.0064	0.0185



(a) k-means (b) Hierarchical (c) Fairlet

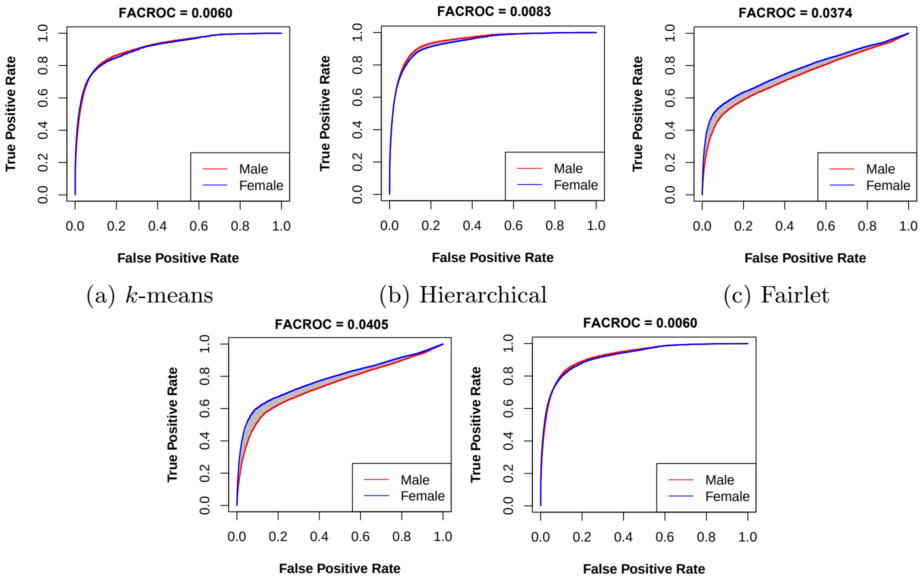
(d) Scalable (e) Proportionally

Fig. 4: COMPAS: FACROC slice plots.

Table 4: Credit card clients: performance of (fair) clustering models

Measures	k-means	Hierarchical	Fairlet	Scalable	Proportionally
Silhouette coefficient	0.5778	**0.6014**	0.3390	0.3476	0.5528
AUCC	0.9247	**0.9422**	0.7456	0.7607	0.9127
Balance	0.6145	0.5927	**0.6957**	0.6440	0.6477
Proportionality	1.0003	0.9939	1.0909	1.0828	**1.4935**
FACROC	**0.0060**	0.0083	0.0374	0.0405	**0.0060**

Student performance dataset. In the Student-Mat subset (Table 6 and Fig. 7), in terms of clustering performance, k-means outperforms other clustering models, although its silhouette coefficient is low. Interestingly, *Scalable* fair clustering achieves the highest balance score due to having the lowest AUCC score. Moreover, the k-means model once again achieves the highest FACROC value, followed by *Hierarchical clustering* and *Proportionally* fair clustering. However,



(a) k-means (b) Hierarchical (c) Fairlet

(d) Scalable (e) Proportionally

Fig. 5: Credit card: FACROC slice plots.

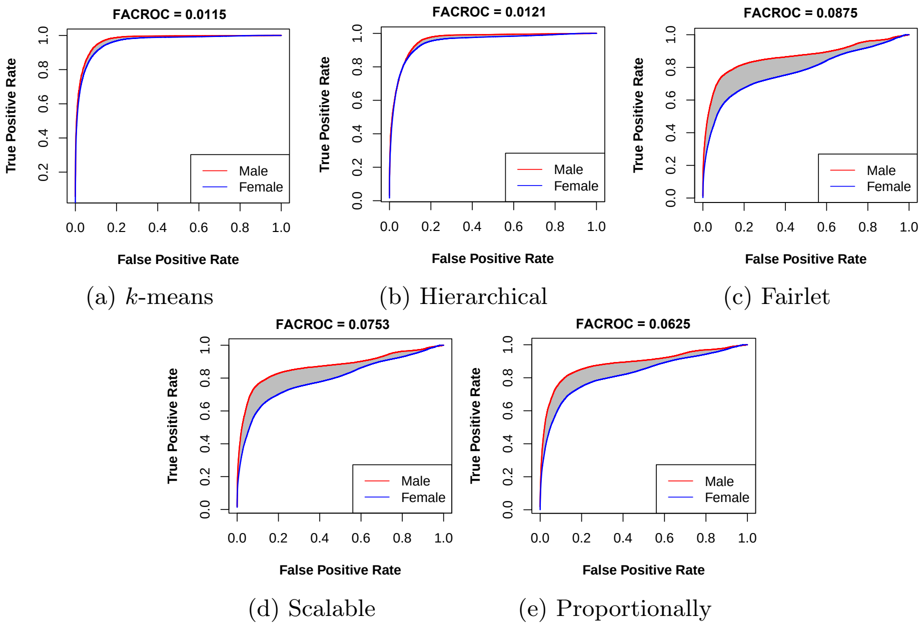
Table 5: German credit: performance of (fair) clustering models

Measures	**k-means**	**Hierarchical**	**Fairlet**	**Scalable**	**Proportionally**
Silhouette coefficient	**0.7222**	0.6963	0.1387	0.3053	0.5302
AUCC	**0.9672**	0.9523	0.8046	0.8192	0.8440
Balance	0.3008	0.3314	**0.4545**	0.4045	0.3669
Proportionality	0.9469	0.9874	1.2964	1.2975	**1.3829**
FACROC	**0.0115**	0.0121	0.0875	0.0753	0.0625

a different trend emerges in the Student-Por subset (Table 7 and Fig. 8), where *Fairlet* outperforms other models in terms of FACROC, while *Proportionally* achieves the best balance score and proportionality.

Table 6: Student-Mat: performance of (fair) clustering models

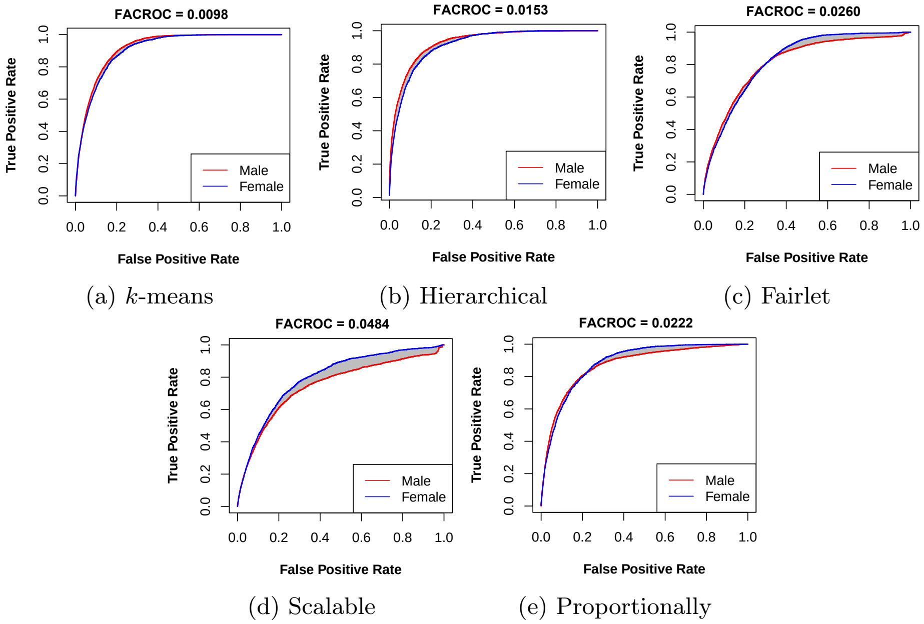
Measures	**k-means**	**Hierarchical**	**Fairlet**	**Scalable**	**Proportionally**
Silhouette coefficient	**0.1814**	0.1526	0.0456	0.0552	0.1111
AUCC	**0.9117**	0.8931	0.8201	0.7705	0.8823
Balance	0.0000	0.1294	0.4231	**0.6176**	0.3182
Proportionality	0.9970	1.1295	1.2782	1.1510	**1.4167**
FACROC	**0.0098**	0.0153	0.0260	0.0484	0.0222



(a) k-means (b) Hierarchical (c) Fairlet

(d) Scalable (e) Proportionally

Fig. 6: German credit: FACROC slice plots.



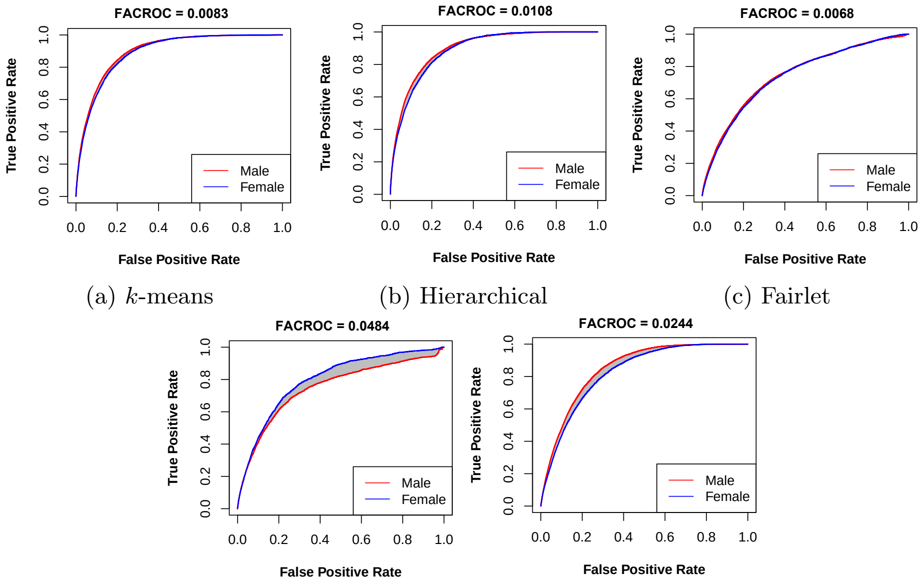
(a) k-means (b) Hierarchical (c) Fairlet

(d) Scalable (e) Proportionally

Fig. 7: Student-Mat: FACROC slice plots.

Table 7: Student-Por: performance of (fair) clustering models

Measures	k-means	Hierarchical	Fairlet	Scalable	Proportionally
Silhouette coefficient	**0.1345**	0.1136	0.0242	0.0591	0.0618
AUCC	**0.8935**	0.8641	0.7389	0.8270	0.8459
Balance	0.3881	0.3972	0.4184	0.4231	**0.4483**
Proportionality	1.0341	1.3728	1.2142	1.1623	**1.4320**
FACROC	0.0083	0.0108	**0.0068**	0.0484	0.0244



(a) k-means (b) Hierarchical (c) Fairlet

(d) Scalable (e) Proportionally

Fig. 8: Student-Por: FACROC slice plots.

5 Conclusions and outlook

In this work, we introduced FACROC, a new fairness notion for fair clustering that leverages the ROC curves of clustering analysis w.r.t. a protected attribute. We evaluated our proposed fairness measure on several datasets and (fair) clustering models. The results demonstrate that our measure is effective in visualizing and analyzing the fairness of clustering models using ROC curves. Furthermore, the evaluation highlights significant variations among fairness measures due to differences in their definitions and objective functions. In the future, we plan to extend this work by developing a method to optimize the FACROC value and adapt it for multiple protected attributes.

References

1. Ahmadian, S., Epasto, A., Kumar, R., Mahdian, M.: Clustering without over-representation. In: KDD 2019. pp. 267–275 (2019)
2. Backurs, A., Indyk, P., Onak, K., Schieber, B., Vakilian, A., Wagner, T.: Scalable fair clustering. In: ICML 2019. pp. 405–413. PMLR (2019)
3. Bera, S., Chakrabarty, D., Flores, N., Negahbani, M.: Fair algorithms for clustering. Advances in Neural Information Processing Systems **32** (2019)
4. Chakrabarti, D., Dickerson, J.P., Esmaeili, S.A., Srinivasan, A., Tsepenekas, L.: A new notion of individually fair clustering: α-equitable k-center. In: AISTATS 2022. pp. 6387–6408. PMLR (2022)
5. Chen, X., Fain, B., Lyu, L., Munagala, K.: Proportionally fair clustering. In: ICML 2019. pp. 1032–1041. PMLR (2019)
6. Chhabra, A., Masalkovaitė, K., Mohapatra, P.: An overview of fairness in clustering. IEEE Access **9**, 130698–130720 (2021)
7. Chierichetti, F., Kumar, R., Lattanzi, S., Vassilvitskii, S.: Fair clustering through fairlets. Advances in neural information processing systems **30** (2017)
8. Gardner, J., Brooks, C., Baker, R.: Evaluating the fairness of predictive student models through slicing analysis. In: LAK 2019. pp. 225–234 (2019)
9. Ghadiri, M., Samadi, S., Vempala, S.: Socially fair k-means clustering. In: ACM FAccT 2021. pp. 438–448 (2021)
10. Jaskowiak, P.A., Costa, I.G., Campello, R.J.: The area under the roc curve as a measure of clustering quality. Data Mining and Knowledge Discovery **36**(3), 1219–1245 (2022), `https://github.com/pajaskowiak/clusterConfusion`
11. Jung, C., Kannan, S., Lutz, N.: Service in your neighborhood: Fairness in center location. Foundations of Responsible Computing (FORC) (2020)
12. Kleindessner, M., Awasthi, P., Morgenstern, J.: Fair k-center clustering for data summarization. In: ICML 2019. pp. 3448–3457. PMLR (2019)
13. Le Quy, T., Friege, G., Ntoutsi, E.: A review of clustering models in educational data science toward fairness-aware learning. Educational data science: Essentials, approaches, and tendencies: Proactive education based on empirical big data evidence pp. 43–94 (2023)
14. Le Quy, T., Roy, A., Iosifidis, V., Zhang, W., Ntoutsi, E.: A survey on datasets for fairness-aware machine learning. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery **12**(3), e1452 (2022)
15. Saxena, A., Prasad, M., Gupta, A., Bharill, N., Patel, O.P., Tiwari, A., Er, M.J., Ding, W., Lin, C.T.: A review of clustering techniques and developments. Neurocomputing **267**, 664–681 (2017)
16. Thejaswi, S., Ordozgoiti, B., Gionis, A.: Diversity-aware k-median: Clustering with fair center representation. In: ECML PKDD 2021. pp. 765–780. Springer (2021)
17. Žliobaitė, I.: Measuring discrimination in algorithmic decision making. Data Mining and Knowledge Discovery **31**(4), 1060–1089 (2017)