

AMERICAN UNIVERSITY OF BEIRUT

FAIRNESS NOTIONS IN CLUSTERING

by
MAYA TONY EL CHAKHTOURA

A thesis
submitted in partial fulfillment of the requirements
for the degree of Master of Engineering Management
to the Department of Industrial Engineering and Management
of Maroun Semaan Faculty of Engineering and Architecture
at the American University of Beirut



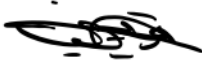
Beirut, Lebanon
May 2022

AMERICAN UNIVERSITY OF BEIRUT

FAIRNESS NOTIONS IN CLUSTERING

by
MAYA TONY EL CHAKHTOURA

Approved by:

	Signature
Dr. Maher Nouiehed, Assistant Professor Industrial Engineering and Management	Advisor
	Signature
Dr. Bacel Maddah, Professor and Chairperson Industrial Engineering and Management	Member of Committee
	Signature
Dr. Hussein Tarhini, Assistant Professor Industrial Engineering and Management	Member of Committee

Date of thesis defense: April 29, 2022


AMERICAN UNIVERSITY OF BEIRUT

THESIS RELEASE FORM

Student Name: El Chakhtoura Maya Tony
Last First Middle

I authorize the American University of Beirut, to: (a) reproduce hard or electronic copies of my thesis; (b) include such copies in the archives and digital repositories of the University; and (c) make freely available such copies to third parties for research or educational purposes:

- As of the date of submission
- One year from the date of submission of my thesis.
- Two years from the date of submission of my thesis.
- Three years from the date of submission of my thesis.



May 12, 2022

Signature

Date

ABSTRACT

OF THE THESIS OF

Maya Tony El Chakhtoura for Master of Engineering Management
Major: Engineering Management

Title: Fairness Notions in Clustering

Machine learning algorithms have been significantly integrated in the automated decision-making processes. Despite their wide practical success, these systems have demonstrated biases towards certain demographic groups. Such instances have motivated researchers to study fairness in machine learning. In this paper, we will focus on fairness in clustering, which is a well-studied unsupervised machine learning task. We propose a new fairness measure FM , Fairness Under Minorities, that is inspired by the Rényi correlation and which yields better fairness results whenever biases are present in minority groups. We outline some derived relations between our proposed notion and other fairness measures. Our experimental study illustrates the effectiveness of FM and proves that it better captures unfairness in minority groups, unlike other fairness measures. This paper also aims at demonstrating what fairness measures best fit certain datasets

TABLE OF CONTENTS

ABSTRACT	1
ILLUSTRATIONS	4
TABLES	5
INTRODUCTION	6
LITERATURE REVIEW	11
2.1. Fairness Notions in Machine Learning	11
2.2. A Fair Clustering Problem	12
2.3. Fairness Approaches in Clustering	14
2.4. Fairness Measures in Clustering	16
PROPOSED FAIRNESS MEASURE	20
RELATIONS	23
4.1. Case of 2 Clusters $K = 2$ with a Single Binary Sensitive Attribute	23
4.2. Case of 3 Clusters $K = 3$ with a Single Binary Sensitive Attribute	25
4.3. Generalized Case for a Single Binary Sensitive Attribute	27
ILLUSTRATIONS	29
5.1. Case of 2 Clusters with a Single Binary Sensitive Attribute	30
5.2. Case of 5 Clusters with a Single Binary Sensitive Attribute	35

EXPERIMENTAL STUDY	39
CONCLUSION AND FUTURE WORK.....	45
APPENDIX 1 CASE OF TWO CLUSTERS.....	46
APPENDIX 2 CASE OF THREE CLUSTERS.....	49
REFERENCES	52

ILLUSTRATIONS

Figure

1. Randomly generated data.....	13
2. Clustering.....	13
3. The approximation Q vs. our proposed notion FM for 2 clusters.....	21
4. Instance 1	33
5. Instance 2	33
6. Instance 3	33
7. Instances 4 and 5.....	34
8. Instance 6	34
9. Instance 7	34
10. Distribution of points in the clusters.....	40
11. Distribution of points with respect to gender.....	41
12. Distribution of points with respect to marital status	41
13. Distribution of points with respect to occupation.....	41
14. Balance with respect to each sensitive attribute	43
15. Weighted deviation with respect to each sensitive attribute.....	43
16. Approximation Q with respect to each sensitive attribute.....	43
17. FM with respect to each sensitive attribute	44

TABLES

Table

1. Instances for case 1	31
2. Fairness evaluation for case 1	32
3. Instances for case 2	36
4. Fairness evaluation for case 2	38
5. Non-sensitive attribute description	39
6. Distribution of points with respect to gender.....	40
7. Distribution of points with respect to marital status	40
8. Distribution of points with respect to occupation.....	40
9. Fairness measures	42
10. Summary of findings	45

CHAPTER 1

INTRODUCTION

Automated decision-making algorithms are recently being adopted in a wide range of applications that significantly affect our lives. Such applications include recommendation systems used by Netflix and Amazon, spam identification systems used to classify suspicious accounts on social media or used to identify spam emails, and vehicular automation systems (Araujo, Helberger, Kruikemeier, & de Vreese, 2020). Moreover, these automated decision-making processes have shown significant relevance in the judicial and law enforcement sector. For instance, such algorithms are being deployed in the US to recommend who is eligible for early release from jail (Dressel & Farid, 2018).

Despite their wide practical success, these models have demonstrated biases towards certain demographic groups when deployed in real systems. For instance, the automated selection program used by St. George Hospital Medical School (Lowry & Macpherson, 1988) to facilitate screening process of applicants was programmed to reject applications with grammatical and spelling mistakes as they indicated a poor English standing. As non-native English speakers were more prone to send applications with linguistic errors, the automated system started correlating the likelihood of getting accepted to race, birthplace, and address. Subsequently, as the English level of foreigners enhanced, outstanding applicants were still getting rejected because of their birthplace or address, and thus indicating an ethnic and racial biases in the automated system.

Another example is the judicial system used in the United States courts to assess the probability of a person to commit another crime. Judges used the software

Correctional Offender Management Profiling for Alternative Sanctions (COMPAS) to decide whether an offender should be kept in prison or get released. Dressel et al (2018) studied the effectiveness of this system and concluded that the results suffer from biases against certain protected groups. More specifically, its predictions were mainly in favor of Caucasians, i.e., an African-American and a Caucasian with the same profiles were not treated similarly. They even argue that a simple linear model provided only with two input features operates similarly to the software COMPAS which has 137 features.

Such instances have motivated researchers to study fairness in automated decision-making. Fairness in this context is defined as the *absence of any prejudice or favoritism toward an individual or a group based on their inherent or acquired characteristics* (Mehrabi, Morstatter, Saxena, Lerman, & Galstyan, 2019). Multiple fairness definitions were proposed in literature. Some focus on fairness of outcome which is concerned with egalitarian results, while others focus on fairness of process which is concerned with making the decision process as fair as possible (Trautmann & Van de Kuilen, 2016). In both cases presented earlier, the algorithms attempted a fair decision-making process by omitting all information related to race or ethnicity. However, these automated systems still suffered from prejudice since other attributes behaved as proxies for sensitive ones (such as race, gender, age, marital status, color, and religion).

In addition to the ethical point of view, a fair treatment for different demographic groups is legally required by many countries. The disparate impact doctrine, which refers to an unintentional disproportionate outcome that affects a protected group, was first recognized by the Supreme Court in the United States in 1971 (States, 1971). It started when African-American workers sued the Duke Power Company for requiring a high school diploma or passing an intelligence test to get promoted. They argued that they

were not illegible for that since North Carolina lacks proper education. Thus, the court ruled that employees should be hired or promoted based on their job performance. Accordingly, in employment, the disparate impact states that it is illegal for protected groups to suffer from discrimination in the hiring process, and it is informally known as the 80% rule. Similarly, other countries have implemented laws to evaluate fairness in certain fields under the disparate impact and disparate treatment notions, where the latter is defined as intentional discrimination against a protected group (Seiner, 2006).

In addition to these two doctrines, multiple notions have been recently proposed to impose fairness. Some of these notions that measure group fairness are demographic parity, equalized opportunity, and equalized odds (Mehrabi, Morstatter, Saxena, Lerman, & Galstyan, 2019). Demographic parity states that the positive outcome of a decision-making process should be independent of the sensitive attribute. For example, if both males and females apply to a certain university, demographic parity will ensure that both genders will be equally accepted regardless of whether one group is more eligible than the other. Equalized opportunity is another fairness metric that requires the positive outcome to be independent of the sensitive attribute given that the protected group is actually qualified. Referring to the same example, equalized opportunity ensures that qualified males and females have the same probability of getting accepted. Equalized odds operates similarly; however, it also ensures that unqualified protected groups will have equal negative outcomes.

These fairness notions are being also applied to machine learning algorithms which have been significantly integrated in the automated decision-making processes. These algorithms study hidden patterns in the data, and subsequently use these patterns to generate a certain outcome. Machine learning algorithms can be classified into two

categories: supervised and unsupervised (Alloghani, Al-Jumelly, Mustafina, Hussain, & Aljaaf, 2020). The main difference between these two classes is the presence of labels in the training data. Supervised learning tasks aim at generalizing information from available labeled data to be used for predicting unlabeled data. Unsupervised learning is defined as the process of grouping data that is not classified or categorized using automated algorithms that learn the underlying feature from the available data.

One popular well-studied unsupervised machine learning task is clustering. This task partitions data points into groups/clusters in a way that maximizes intra-cluster similarity and minimizes inter-cluster resemblance between objects, i.e., *given a set of data points, partition them into a set of groups which are as similar as possible* (Aggarwal & Reddy, 2013). To have a fair clustering assignment, the data points have to be partitioned in a way that would not be biased to the protected groups. Chierichetti et al. (2017) were the first to introduce the disparate impact notion of fairness to the clustering problem. They defined fairness as having balanced proportions of the demographic groups in each cluster. Similarly, Ziko et al. (2019), Abraham et al. (2018), and Baharlouei et al. (2019) quantified fairness as a measure of independence where a clustering assignment is said to be fair if the partitioning is independent of the sensitive attribute. These fairness measures will be further discussed in later sections.

As will be highlighted in later sections, minorities have been underrepresented in previously developed notions of fairness. Many instances have recorded discriminations against protected groups in small clusters. Among these instances is the anomaly detection model which identifies events that deviate from the majority of the dataset. For example, tech companies tend to block users who sign in with uncommon names. In some cultures, names are more frequently used and widespread; whereas, in other civilizations

people might have unique names, and thus identified as fake by the model. Accordingly, our main concern is to compare different fairness measures and propose a new measure that performs better in the presence of minorities. In addition to that, we will come up with conclusions on what fairness measures best fit certain datasets.

In this thesis, we will focus on studying fairness in minority groups in clustering. We will start by presenting some literature related to fairness in machine learning, specifically clustering. Then, we will introduce our proposed fairness measure *FM (Fairness under Minorities)* which will be theoretically compared to other measures by deriving some mathematical relations. To test the efficiency of *FM*, we will conduct an experimental study to compare its results with the different fairness measures. According, a summary of the outcomes will be presented along with some future work that still need to be performed.

CHAPTER 2

LITERATURE REVIEW

2.1. Fairness Notions in Machine Learning

In a typical machine learning problem, the algorithm is supplied by the feature X to generate the target label Y . Most of the fairness definitions studied in the literature of machine learning focus on three aspects of a binary classifier; the sensitive variable S for which fairness is measured, the target variable Y which can take two values, 0 or 1, in binary classification, and the score \hat{Y} which represents the predicted outcome, 0 or 1, for each observation. Accordingly, the following three fairness criteria can be categorized (Caton & Haas, 2020):

1. The Independence criterion which requires the score R to be independent of the sensitive attribute S : $\hat{Y} \perp S$
2. The Separation criterion which is an extension of the Independence property and requires the score R and the sensitive variable S to be independent while conditioning on the target variable Y : $\hat{Y} \perp S|Y$
3. The Sufficiency criterion which requires the target variable Y and the sensitive variable S to be independent while conditioning on the score R : $Y \perp S|\hat{Y}$

Under these properties, some previously studied notions of fairness can be defined. Among those that are related to the Independence criterion are the demographic parity and disparate impact. Demographic parity defines fairness as equal probabilities of being classified with the positive outcome, and it is represented by the following:

$$P(\hat{Y} = 1|S = 0) = P(\hat{Y} = 1|S = 1). \quad (1)$$

Disparate impact considers the ratio of the positive outcome between unprivileged and privileged groups, and it is represented by the following:

$$\frac{P(\hat{Y} = 1|S = 0)}{P(\hat{Y} = 1|S = 1)}. \quad (2)$$

Among the fairness notions that are related to the Separation criterion are the equalized opportunity and equalized odds. One advantage of such metrics is that they consider the underlying differences between the protected groups. Equalized opportunity defines fairness as equal probabilities of the true positive rates (TPR) across different groups, and it is represented by the following:

$$P(\hat{Y} = 1|S = 0, Y = 1) = P(\hat{Y} = 1|S = 1, Y = 1). \quad (3)$$

Similar to the equalized opportunity, equalized odds considers the false positive rates (FPR) in addition to the TPR, and it is represented by the following

$$P(\hat{Y} = 1|S = 0, Y = 1) = P(\hat{Y} = 1|S = 1, Y = 1)$$

and

$$P(\hat{Y} = 0|S = 0, Y = 0) = P(\hat{Y} = 0|S = 1, Y = 0). \quad (4)$$

2.2. A Fair Clustering Problem

Consider a set of N data points (p_1, \dots, p_N) associated with a sensitive attribute S . The clustering objective is to partition them into K clusters with corresponding centroids $C = [c_1, \dots, c_K]$ based on similarity where the random variable $A = [a_{ij}]$ takes the value of 1 if point p_j is assigned to cluster i , and zero otherwise where $i \in \{1, \dots, K\}$ and $j \in \{1, \dots, N\}$.

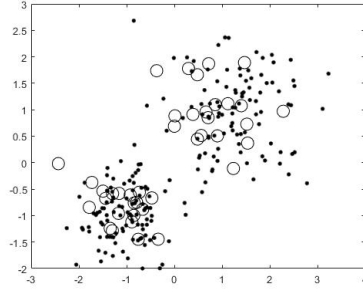


Figure 1: Randomly generated data

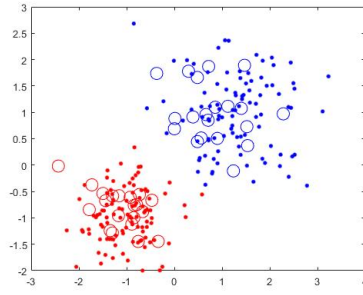


Figure 2: Clustering

Given the scattered dataset shown in figure 1, K -means, which is one of the most used partitioning algorithms, starts by choosing randomly K points as the initial centroids. Then, it assigns each data point to the closest centroid, and once the clusters are formed, the centroids of each cluster are updated. These two steps are iteratively repeated until the algorithm converges, i.e., the centroids no longer change (Reddy & Vinzamuri, 2014). Figure 2 demonstrates the final result of K -means clustering where the points are partitioned into two clusters ($K = 2$).

Mathematically, K -means algorithm aims at partitioning the points into K clusters by minimizing the following objective function:

$$\min_{A,C} \sum_{j=1}^N \sum_{i=1}^K a_{ij} \|p_j - c_i\|^2 \quad s.t. \quad \sum_{i=1}^K a_{ij} = 1 \quad \forall j, \quad a_{ij} \in \{0,1\}, \quad \forall i,j. \quad (5)$$

Now, taking into account the binary sensitive attribute S that can take one of two values X (represented by a dot) or Y (represented by a circle), the clustering assignment is said to be fair under the disparate impact doctrine if the protected groups have equal proportions in the clusters. In simpler words, considering discrete sensitive attributes, their distribution within clusters has to be proportional to their distribution in the dataset. Thus, if we consider the cluster assignment and the sensitive attribute to be two random variables, in the presence of independence, the conditional distribution of the sensitive attribute given the clustering assignment has to be equal to the distribution of the sensitive attribute in the dataset. Accordingly, for this clustering problem to be fair under the disparate impact doctrine, the random variable $A = [a_{ij}]$ has to be independent of the sensitive attribute S .

2.3. Fairness Approaches in Clustering

To impose such fairness on the clustering problem, several approaches have been developed which are classified into three categories: pre-processing, in-processing, and post-processing techniques (Mehrabi, Morstatter, Saxena, Lerman, & Galstyan, 2019). The pre-processing approach works on removing discrimination prior to the decision-making process, the in-processing technique works on imposing fairness simultaneously during the task of clustering, and the post-processing method adjusts the model output by removing discrimination from the classifier.

Among the pre-processing techniques is the one developed by Chierichetti et al. (2017). They introduced the concept of balance to impose fairness to the clustering problem with a single binary sensitive attribute, i.e., every point can take one of two labels. The balance of a cluster is then defined as the minimum between the fraction of

label 1 to label 2 and its inverse. The overall balance of the whole dataset is the minimum between the balances of all clusters. Accordingly, the higher the overall balance, the more proportional is the distribution of the sensitive attribute in the clusters, and thus the fairer the clustering process. To impose such fairness, the data points are first grouped into small fair subsets called fairlets which are comprised of only two points, one belonging to label 1 and the other to label 2, and then a balanced clustering is obtained by merging these fairlets into clusters. However, the running time for this method turned out to be quadratic, i.e., it takes so much time to be solved. Backurs et al. (2019) addressed this issue by proposing an algorithm that would compute the fairlet-decompositions in near-linear time. In addition to that, Schmidt et al. (2018) introduced the concept of corsets which are subsets in the dataset used to reduce the size of the input data. Fair clustering is then solved over these subsets which are afterwards combined to generate an approximately fair dataset. Their method covers single categorical sensitive attributes, i.e., the points in the dataset can take one of multiple values.

Among the in-processing techniques, many proposed methods add a regularization term to the clustering objective or add a fairness constraint to demonstrate proportional representation of protected groups in the clusters. Ziko et al. used the Kullback-Leibler loss term (KL-divergence) as a regularization term to penalize the difference between the probability of protected groups in each cluster and the desired target proportion. Accordingly, fairness is embedded simultaneously during the task of clustering through a maximization-minimization algorithm. This process aims at minimizing the clustering term and maximizing the fairness loss term. It is applied on single categorical sensitive attributes, handles large datasets, and guarantees convergence. Another approach that uses a regularization term in the objective function

is the FairKM outlined by Abraham et al. (2020). They used the weighted deviation fairness term to penalize the difference between the proportional representation of protected groups in each cluster and their respective proportions in the dataset. Fairness is imposed during the clustering process which is attained by applying k-means algorithm. The concept covers multiple binary, categorical, and numeric sensitive attributes, and maintains a decent clustering objective while imposing fairness. Baharlouei et al. (2019) used an approximation of the Hirschfeld-Gebelein-Rényi (HGR) fairness term to penalize the dependence between the clustering assignment A and the sensitive attribute S . This approximation represents an upper bound to the Rényi correlation and covers discrete sensitive attributes. This method ensures a decent clustering quality by adding a regularization term that minimizes the clustering objective and maximizes the fairness measure.

2.4. Fairness Measures in Clustering

Recently, researchers started studying fairness in clustering. They developed measures that quantify fairness under the disparate impact doctrine which requires proportionate representation of the protected groups in the outcome. Accordingly, a clustering problem is said to be fair if the proportions of the protected groups are equal among clusters. Motivated by that, Chierichetti et al. (2017) introduced the concept of balanced clusters.

Consider a set of N data points (p_1, \dots, p_N) that need to be partitioned into K clusters and that are associated with a single binary sensitive attribute S which can take one of two values X or Y . Let x_i be the number of points that are labeled X and belong to cluster i , y_i be the number points that are labeled Y and belong to cluster i , and n_i be the

number of points that belong to cluster i where $i \in \{1, \dots, K\}$. Let x be the number of points that have the X label in the dataset and y be the number of points that have the Y label in the dataset. (Note that this problem will be used in the relations and illustrations sections later on).

The balance of a cluster can be defined as follows:

$$Balance_i = \min \left\{ \frac{x_i}{y_i}, \frac{y_i}{x_i} \right\}. \quad (6)$$

The balance of the whole data set is then given as follows:

$$Balance = \min \{ Balance_1, \dots, Balance_K \}. \quad (7)$$

Therefore, the higher the balance, the fairer the clustering solution.

Another measure that quantifies fairness in clustering under the disparate impact doctrine is the weighted deviation (Abraham, Deepak, & Sundaram, 2018). It measures the squared difference between the distribution of the sensitive attribute S in each cluster i and its distribution in the dataset. Thus, the closer its value is to zero, the fairer the clustering solution. The weighted deviation of the dataset is given as follows:

$$\sum_{i=1}^K \left(\frac{n_i}{N} \right)^2 \left(\frac{\left(\frac{x_i}{n_i} - \frac{x}{N} \right)^2 + \left(\frac{y_i}{n_i} - \frac{y}{N} \right)^2}{n_i} \right). \quad (8)$$

Equation (8) shows that the deviation term is multiplied by the square of each cluster's fractional representation in the dataset. Thus, the value of the weighted deviation gets enlarged for large clusters.

In addition to these metrics, several independence measures have been used to quantify fairness in clustering. One of these measures is the Hirschfeld-Gebelein-Rényi correlation which was found to be superior to the Pearson correlation coefficient and the

Hilbert Schmidt independence criterion (HSIC) since it captures non-linear dependence between random variables. For two random variables A and S , the Rényi correlation is given by the following expression (Rényi, 1959):

$$HGR(A, S) = \sup_{f, g} \mathbb{E}\{f(A)g(S)\} \quad (9)$$

$$s. t. \quad \mathbb{E}\{f(A)\} = \mathbb{E}\{g(S)\} = 0 \quad \text{and} \quad \mathbb{E}\{f^2(A)\} = \mathbb{E}\{g^2(S)\} = 1. \quad (10)$$

The Rényi correlation is zero if and only if the two random variables are independent and 1 otherwise. However, computing it can be difficult, thus an upper bound can be used instead for the discrete case. Accordingly, Witsenhausen (1975) evaluated the HGR coefficient as the second highest eigenvalue of a well-defined matrix. This upper bound is exact whenever one of the random variables is binary.

Considering the two random variables to be the sensitive attribute S and the classifier A , the Rényi correlation can be rendered as follows:

$$Q(S, A) = \sum_{S=s} \sum_{A=a} \frac{P(S=s, A=a)^2}{P(S=s)P(A=a)} - 1 \quad (11)$$

where the random variable A is binary that takes the value of 1 if the data point is assigned to a certain cluster and 0 otherwise. Accordingly, a clustering assignment is said to be fair if and only if the classifier A is independent of the sensitive attribute S . Fairness is achieved whenever this upper bound term converges to 0, representing total independence, and whenever this value approaches 1, it highlights absolute dependence.

Considering the case of two clusters with a single binary sensitive attribute, equation (11) can be simplified as follows:

$$Q(S, A) = \sum_{S=s} \frac{P(S=s, A=0)^2}{P(S=s)P(A=0)} + \sum_{S=s} \frac{P(S=s, A=1)^2}{P(S=s)P(A=1)} - 1 \quad (12)$$

\Leftrightarrow

$$Q(S, A) = \sum_{S=s} \frac{P(S=s | A=0)^2 P(A=0)^2}{P(S=s)P(A=0)} + \sum_{S=s} \frac{P(S=s | A=1)^2 P(A=1)^2}{P(S=s)P(A=1)} - 1 \quad (13)$$

\Leftrightarrow

$$Q(S, A) = P(A = 0) \sum_{S=s} \frac{P(S=s | A=0)^2}{P(S=s)} + P(A = 1) \sum_{S=s} \frac{P(S=s | A=1)^2}{P(S=s)} - 1 \quad (14)$$

where $P(A = a)$ represents the chance a point belongs to a cluster a , $P(S = s | A = a)$ represents the fraction of points in cluster a with $S=s$, and $P(S = s)$ represents the ratio of points with $S=s$ in the dataset .

CHAPTER 3

PROPOSED FAIRNESS MEASURE

The approximation term Q presented earlier might however reflect fairness in cases where the clustering task suffers from discrimination. For example, suppose we are performing a statistical study to check whether job opportunities are distributed fairly among males and females in two different companies. Thus, we are considering gender as our sensitive attribute. Out of the 100 employees in Company A, 20 are females and 80 are males. Out of the 1,900 employees in Company B, 900 are females and 1,000 are males. By applying equation (14), the approximation measure Q is given by the following:

$$Q = \left(\frac{100}{2000}\right) \left(\frac{\left(\frac{20}{100}\right)^2}{\frac{920}{2000}} + \frac{\left(\frac{80}{100}\right)^2}{\frac{1080}{2000}}\right) + \left(\frac{1900}{2000}\right) \left(\frac{\left(\frac{900}{1900}\right)^2}{\frac{920}{2000}} + \frac{\left(\frac{1000}{1900}\right)^2}{\frac{1080}{2000}}\right) - 1 = 0.0143$$

It can be seen that this value is very close to 0, thus reflecting independence. This indicates that the job distributions for females and males are fair among both companies. However, we can see that Company A, which represents a minor group in the dataset, suffers from a major discrimination that was not captured by this fairness measure. Influenced by such instances, we modified the approximation Q to what is presented in equation (15).

$$FM(S, A) = \frac{1}{K} \sum_{S=s} \sum_{A=a} \frac{P(S=\frac{s}{A}=a)^2}{P(S=s)} - 1. \quad (15)$$

Our proposed measure FM , denoting *Fairness under Minorities*, accounts for minor groups in the dataset, and thus eliminates the downside of the previously stated metrics. The main difference between our proposed measure FM and the approximation Q is that the term $P(A = a)$, denoting the fractional representation of each cluster in the

dataset, has been replaced by $\frac{1}{K}$ where K represents the total number of clusters. By that, the fairness measure would not be skewed towards large clusters and would also highlight biases whenever they are present in minority groups.

Considering the same example stated before and applying equation (15), our proposed measure conveys the following:

$$FM = \left(\frac{1}{2}\right) \left(\frac{\left(\frac{20}{100}\right)^2}{\frac{920}{2000}} + \frac{\left(\frac{80}{100}\right)^2}{\frac{1080}{2000}} \right) + \left(\frac{1}{2}\right) \left(\frac{\left(\frac{900}{1900}\right)^2}{\frac{920}{2000}} + \frac{\left(\frac{1000}{1900}\right)^2}{\frac{1080}{2000}} \right) - 1 = 0.136$$

It can be seen that FM resulted in a higher value than Q , thus better capturing the gender discrimination present in Company A. To better visualize the difference, we considered the same sample but with all possible female job-distributions.

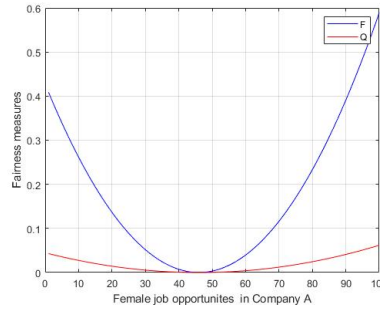


Figure 3: The approximation Q vs. our proposed notion FM for 2 clusters

Figure 3 represents the approximation Q and our proposed fairness measure FM with respect to the female distribution in Company A. It can be seen that FM represents an upper-bound to the approximation Q , and thus better represents minorities. The red curve representing the approximation Q is almost flat and thus not demonstrating the bias in the minority group. Whereas, the blue curve representing our fairness measure FM has a u-shape and thus is more sensitive to biases in the minority group. Also, notice that these two measures are equal to zero whenever the sensitive attribute is equally distributed. However, this is the case of only 2 clusters. In later sections, we will test the

efficiency of FM for a generalized case and compare it to other fairness measures besides this approximation Q .

CHAPTER 4

RELATIONS

As mentioned previously, we will be comparing our fairness measure FM to the balance, the weighted deviation, and the approximation Q of the Rényi model to illustrate the functionality of FM in fair clustering. First, the mathematical formulations for each measure will be presented and then they will be derived in terms of one another.

The relations will be computed based on the previously defined problem in the literature review section for three different cases: the first case covering two clusters ($K = 2$), the second case covering three clusters ($K = 3$), and the third one covering generalized relations for K clusters.

4.1. Case of 2 Clusters ($K = 2$) with a Single Binary Sensitive Attribute

In accordance with equations (6) and (7), the balance measure can be defined as follows:

$$Balance_1 = \min\left\{\frac{x_1}{y_1}, \frac{y_1}{x_1}\right\}, \quad (16)$$

$$Balance_2 = \min\left\{\frac{x_2}{y_2}, \frac{y_2}{x_2}\right\}, \quad (17)$$

and
$$Overall\ balance = \min\{Balance_1, Balance_2\}. \quad (18)$$

In accordance with equation (8), the weighted deviation can be defined as follows:

$$Weighted\ Deviation = \left(\frac{n_1}{N}\right)^2 \left(\frac{\left(\frac{x_1 - x}{n_1 - N}\right)^2 + \left(\frac{y_1 - y}{n_1 - N}\right)^2}{n_1}\right) + \left(\frac{n_2}{N}\right)^2 \left(\frac{\left(\frac{x_2 - x}{n_2 - N}\right)^2 + \left(\frac{y_2 - y}{n_2 - N}\right)^2}{n_2}\right). \quad (19)$$

Through mathematical derivations, equation (19) can be simplified to the following:

$$\text{Weighted Deviation} = \frac{2(x_1N - xn_1)^2}{N^3n_1n_2}. \quad (20)$$

In accordance with equation (11), the approximation Q can be defined as follows:

$$Q = \left(\frac{n_1}{N}\right) \left(\frac{\left(\frac{x_1}{n_1}\right)^2}{\frac{x}{N}} + \frac{\left(\frac{y_1}{n_1}\right)^2}{\frac{y}{N}}\right) + \left(\frac{n_2}{N}\right) \left(\frac{\left(\frac{x_2}{n_2}\right)^2}{\frac{x}{N}} + \frac{\left(\frac{y_2}{n_2}\right)^2}{\frac{y}{N}}\right) - 1. \quad (21)$$

Through mathematical derivations, equation (21) can be simplified to the following:

$$Q = \frac{(x_1N - xn_1)^2}{xyn_1n_2}. \quad (22)$$

In accordance with equation (15), our proposed measure FM can be defined as follows:

$$FM = \left(\frac{1}{K}\right) \left(\frac{\left(\frac{x_1}{n_1}\right)^2}{\frac{x}{N}} + \frac{\left(\frac{y_1}{n_1}\right)^2}{\frac{y}{N}}\right) + \left(\frac{1}{K}\right) \left(\frac{\left(\frac{x_2}{n_2}\right)^2}{\frac{x}{N}} + \frac{\left(\frac{y_2}{n_2}\right)^2}{\frac{y}{N}}\right) - 1. \quad (23)$$

Through mathematical derivations, equation (23) can be simplified to the following:

$$FM = \frac{(x_1N - xn_1)^2(n_1^2 + (N - n_1)^2)}{2xyn_1^2n_2^2}. \quad (24)$$

Lemma 1: Suppose that $K = 2$ and S is a binary sensitive attribute, then the following relations between fairness notions exist:

$$Q = \frac{N^3}{2xy} [\text{WeightedDeviation}], \quad (25)$$

$$FM = \frac{1}{2} \left[\frac{n_1}{n_2} + \frac{n_2}{n_1} \right] [Q], \quad (26)$$

and

$$FM = \frac{N^3 \left(\frac{n_1 + n_2}{n_2 n_1} \right)}{4xy} [\text{WeightedDeviation}]. \quad (27)$$

Equation (25) represents the approximation Q in terms of the weighted deviation.

It shows that the approximation Q can be attained by multiplying the weighted deviation

by the factor $\frac{N^3}{2xy}$. Analyzing this equation, it can be deduced that the former gets amplified whenever we have a larger dataset and whenever the points are not distributed evenly in the dataset.

Equation (26) represents FM in terms of the approximation Q . It shows that FM can be attained by multiplying the approximation Q by the factor $\frac{1}{2} \left[\frac{n_1}{n_2} + \frac{n_2}{n_1} \right]$. Analyzing this equation, it can be deduced that this factor will always augment the value of our measure FM , thus representing an upper bound to the approximation Q . Also, notice that FM will have the same value of Q only if the clusters have the same distribution of points.

Equation (27) represents FM in terms of the weighted deviation. It shows that FM can be attained by multiplying the weighted deviation by the factor $\frac{N^3 \left(\frac{n_1+n_2}{n_2 n_1} \right)}{4xy}$. Analyzing this equation, it can be deduced that the former gets amplified whenever we have a larger dataset and whenever the points are not distributed evenly in the dataset.

4.2. Case of 3 Clusters ($K = 3$) with a Single Binary Sensitive Attribute

In accordance with equations (6) and (7), the balance measure can be defined as follows:

$$Balance_1 = \min \left\{ \frac{x_1}{y_1}, \frac{y_1}{x_1} \right\}, \quad (28)$$

$$Balance_2 = \min \left\{ \frac{x_2}{y_2}, \frac{y_2}{x_2} \right\}, \quad (29)$$

$$Balance_3 = \min \left\{ \frac{x_3}{y_3}, \frac{y_3}{x_3} \right\}, \quad (30)$$

and $Overall\ balance = \min\{Balance_1, Balance_2, Balance_3\}$. (31)

In accordance with equation (8), the weighted deviation can be defined as follows:

Weighted Deviation =

$$\left(\frac{n_1}{N}\right)^2 \left(\frac{\left(\frac{x_1-x}{n_1-N}\right)^2 + \left(\frac{y_1-y}{n_1-N}\right)^2}{n_1} \right) + \left(\frac{n_2}{N}\right)^2 \left(\frac{\left(\frac{x_2-x}{n_2-N}\right)^2 + \left(\frac{y_2-y}{n_2-N}\right)^2}{n_2} \right) + \left(\frac{n_3}{N}\right)^2 \left(\frac{\left(\frac{x_3-x}{n_3-N}\right)^2 + \left(\frac{y_3-y}{n_3-N}\right)^2}{n_3} \right). \quad (32)$$

Through mathematical derivations, equation (32) can be simplified to the following:

$$\text{Weighted Deviation} = \frac{2[n_1(x_2N-xn_2)^2+n_2(x_1N-xn_1)^2-N(x_1n_2-x_2n_1)^2]}{N^3n_1n_2n_3}. \quad (33)$$

In accordance with equation (11), the approximation Q can be defined as follows:

$$Q = \left(\frac{n_1}{N}\right) \left(\frac{\left(\frac{x_1}{n_1}\right)^2}{\frac{x}{N}} + \frac{\left(\frac{y_1}{n_1}\right)^2}{\frac{y}{N}} \right) + \left(\frac{n_2}{N}\right) \left(\frac{\left(\frac{x_2}{n_2}\right)^2}{\frac{x}{N}} + \frac{\left(\frac{y_2}{n_2}\right)^2}{\frac{y}{N}} \right) + \left(\frac{n_3}{N}\right) \left(\frac{\left(\frac{x_3}{n_3}\right)^2}{\frac{x}{N}} + \frac{\left(\frac{y_3}{n_3}\right)^2}{\frac{y}{N}} \right) - 1. \quad (34)$$

Through mathematical derivations, equation (34) can be simplified to the following:

$$Q = \frac{n_1(x_2N-xn_2)^2+n_2(x_1N-xn_1)^2-N(x_1n_2-x_2n_1)^2}{xyn_1n_2n_3}. \quad (35)$$

In accordance with equation (15), our proposed measure FM can be defined as follows:

$$FM = \left(\frac{1}{K}\right) \left(\frac{\left(\frac{x_1}{n_1}\right)^2}{\frac{x}{N}} + \frac{\left(\frac{y_1}{n_1}\right)^2}{\frac{y}{N}} \right) + \left(\frac{1}{K}\right) \left(\frac{\left(\frac{x_2}{n_2}\right)^2}{\frac{x}{N}} + \frac{\left(\frac{y_2}{n_2}\right)^2}{\frac{y}{N}} \right) + \left(\frac{1}{K}\right) \left(\frac{\left(\frac{x_3}{n_3}\right)^2}{\frac{x}{N}} + \frac{\left(\frac{y_3}{n_3}\right)^2}{\frac{y}{N}} \right) - 1. \quad (36)$$

Through mathematical derivations, equation (36) can be simplified to the following:

$FM =$

$$\frac{N[n_1(x_2N-xn_2)^2+n_2(x_1N-xn_1)^2-N(x_1n_2-x_2n_1)^2]}{3xyn_1n_2n_3^2} + \frac{N^2[(n_3-n_1)\left(\frac{x_1-x}{n_1-N}\right)^2+(n_3-n_2)\left(\frac{x_2-x}{n_2-N}\right)^2]}{3xyn_3}. \quad (37)$$

Lemma 2: Suppose that $K = 3$ and S is a binary sensitive attribute, then the following relations between fairness notions exist:

$$Q = \frac{N^3}{2xy} [WeightedDeviation], \quad (38)$$

$$FM = \frac{N[Q]}{3n_3} + \frac{N^2[(n_3 - n_2) \left(\frac{x_2}{n_2} - \frac{x}{N}\right)^2 + (n_3 - n_1) \left(\frac{x_1}{n_1} - \frac{x}{N}\right)^2]}{3xyn_3}, \quad (39)$$

and

$$FM = \frac{N^2}{3xyn_3} \left[\frac{N^2[WeightedDeviation]}{2} + (n_3 - N_1) \left(\frac{x_1}{n_1} - \frac{x}{N}\right)^2 + (n_3 - n_2) \left(\frac{x_2}{n_2} - \frac{x}{N}\right)^2 \right]. \quad (40)$$

It can be seen that the relation between the approximation Q and the weighted deviation, represented by equation (38), is similar to that of the case of two clusters. However, equations (39) and (40), relating our measure FM to the approximation Q and to the weighted deviation respectively, are no longer linear, thus indicating that they might sometimes present an upper bound to FM .

4.3. Generalized Case for a Single Binary Sensitive Attribute

$$Let \ n_i^* = \max\{n_1, \dots, n_k\} \quad (41)$$

Lemma 3: Suppose that K can take any value and S is a binary sensitive attribute, then the following relations between fairness notions exist:

$$Q = \frac{N^3}{2xy} [WeightedDeviation], \quad (42)$$

$$FM = \frac{N[Q]}{kn_i^*} + \frac{N^2}{kxyn_i^*} \sum_{i=1}^k (n_i^* - n_i) \left(\frac{x_i}{n_i} - \frac{x}{N}\right)^2, \quad (43)$$

$$and \quad FM = \frac{N^2}{kxyn_i^*} \left[\frac{N^2[WeightedDeviation]}{2} + \sum_{i=1}^k (n_i^* - n_i) \left(\frac{x_i}{n_i} - \frac{x}{N}\right)^2 \right]. \quad (44)$$

These equations are similar to the ones computed for the case of three clusters, and thus a similar analysis can be followed. To illustrate the significance of these relations, some illustrations and an experimental study will be presented in the following sections.

CHAPTER 5

ILLUSTRATIONS

In this section, we will demonstrate the effectiveness of our proposed fairness measure FM by experimentally comparing it to the balance, weighted deviation, and the approximation Q .

The dataset mentioned in the literature review section will again be considered, but now, the first case covers two clusters ($K = 2$) and the second case covers five clusters ($K = 5$).

Two tables will be presented for each case; the first presenting the different instances studied and the second demonstrating the values obtained from each of the fairness measures considered.

The goal behind this experiment is computing and comparing the values obtained from the four different fairness measures (the balance, the weighted deviation, the approximation Q , and FM) while considering different instances that highlight two factors; the presence of bias in the distribution of the points and the presence of minority groups in the dataset. The points are said to have a bias distribution if the sensitive attribute is not represented proportionally within clusters as in the dataset. Minority groups are present whenever the number of points in one of the clusters is relatively minimal compared to the number of points in other clusters. A (+) sign indicates the presence of bias or minority groups, and a (-) sign indicates their absence each instance considered.

5.1. Case of 2 Clusters with a Single Binary Sensitive Attribute

In this case, 1,000 data points, which are associated with a single binary sensitive attribute, were considered. They were distributed into two clusters, i.e., $K = 2$.

Instance 1 represents an absolute fair distribution under the absence of minority groups where the total number of points is evenly divided between both clusters that have proportional representation of the sensitive attribute as in the dataset.

Instance 2 deviates from instance 1 in that it considers different cluster distributions.

Instance 3 represents minority groups where the number of points in cluster 1 is relatively small compared to the number of points in cluster 2, but with somewhat a fair distribution, i.e., the proportion of points having label X in the dataset is somewhat equivalent to the proportion of points having label X in both clusters, and the proportion of points having label Y in the dataset is somewhat equivalent to the proportion of points having label Y in both clusters.

Instance 4 represents a dataset where the number of points having label X is relatively small to the number of points having label Y ; however, the distribution of the points in the two clusters is equivalent to the distribution of the points in the dataset, i.e. the proportion of points having label X in the dataset is equivalent to the proportion of points having label X in both clusters, and the proportion of points having label Y in the dataset is equivalent to the proportion of points having label Y in both clusters.

Instance 5 is similar to instance 4; however, the distribution of the points in both clusters is not equivalent to the distribution of the points in the dataset.

Instance 6 represents a bias distribution of the points in the clusters where the proportion of points having label X in both clusters is not equivalent to the proportion of

points having label X in the dataset, and the proportion of points having label Y in both clusters is not equivalent to the proportion of points having label Y in the dataset.

Instance 7 represents the presence of minority groups where the number of points in cluster 1 is relatively small compared to the number of points in cluster 2. In addition to that, cluster 1 suffers from bias in the distribution of its points.

Table 1: Instances for case 1

Instances	Bias	Minorities	N	x	y	n_1	x_1	y_1	n_2	x_2	y_2
1	-	-	1,000	500	500	500	250	250	500	250	250
2	-	-	1,000	500	500	400	150	250	600	350	250
3	-	+	1,000	500	500	100	40	60	900	460	440
4	-	-	1,000	100	900	500	50	450	500	50	450
5	+	-	1,000	100	900	500	90	410	500	10	490
6	+	-	1,000	400	600	600	20	580	400	380	20
7	+	+	1,000	400	600	100	10	90	900	390	510

The values obtained from the balance, the approximation Q , and FM range between 0 and 1, where a value of 1 demonstrates complete fairness for the balance measure; whereas, a value of 0 denotes absolute fairness for the approximation Q and FM . The weighted deviation indicates complete fairness when it results in 0; however, it does not have an upper level. That is why its values were normalized to the range of $[0,1]$.

Table 2 represents the values obtained from the four different fairness measures.

It can be deduced that for the case of 2 clusters, the balance was able to capture unfairness whenever we had bias (instances 5 to 7); however, at a greater cost, i.e., it sometimes resulted in exaggerated values (values closer to 0) when we had insignificant difference between the distribution of points in the dataset and that in the clusters (instances 2 and 3).

The weighted deviation and our fairness measure FM resulted in values close to 0 whenever the clusters had a somewhat fair distribution (instances 2 and 3), and they both resulted in values close to 1 whenever we had bias in both clusters (instance 6). However, the weighted deviation resulted in a value of 0.73 whenever we had biased large clusters (instance 5) and in a value of 0.25 whenever a biased minority group; thus, proving that it is skewed towards large clusters. Whereas, our fairness measure FM resulted in a value of 0.2691 for instance 5 and in a value of 0.1898 for instance 7, proving that it treats minority groups and large clusters similarly.

The computed values for the approximation Q were so close to our fairness measure FM , with the latter being an upper bound, and thus capturing unfairness better than the initial approximation to the Rényi model.

As mentioned previously, instance 4 represents a dataset where the proportion of points having label X is relatively small to the proportion of points having label Y ; however, the distribution of the points in the two clusters is equivalent to the distribution of the points in the dataset. For that case, the weighted deviation, the approximation Q , and FM resulted in values of 0, thus indicating that they measure fairness as a means of proportional representation of the sensitive attribute within clusters and the dataset, unlike the balance measure which resulted in an exaggerated value of 0.111.

Table 2: Fairness evaluation for case 1

Instances	Balance	Weighted Deviation	Approximation Q	FM
1	1	0	0	0
2	0.6	0.0625	0.0417	0.0451
3	0.667	0.0399	0.0044	0.0202
4	0.111	0	0	0
5	0.2041	0.73	0.2691	0.2691
6	0.0345	0.84	0.8403	0.9103
7	0.1111	0.25	0.0417	0.1898

To better visualize these results, the values obtained from the weighted deviation, the approximation Q , and our fairness measure FM were plotted for each instance covering all the values that x_i can take.

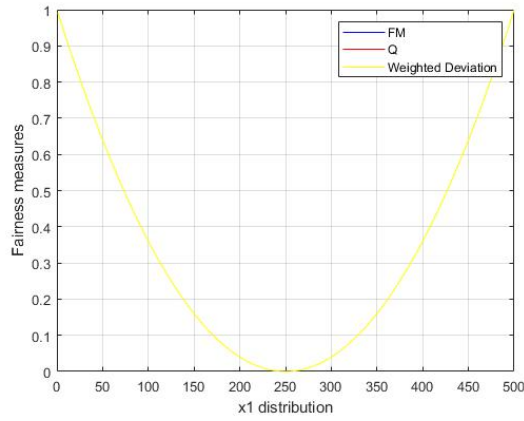


Figure 4: Instance 1

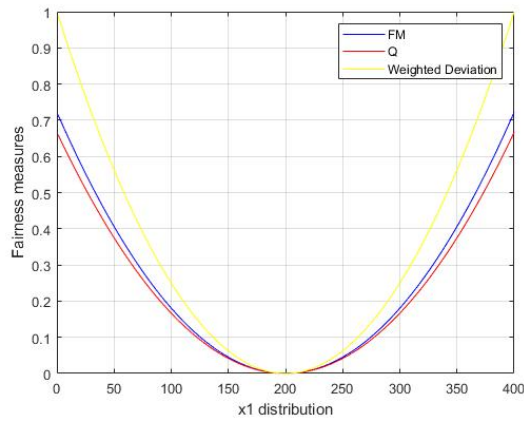


Figure 5: Instance 2

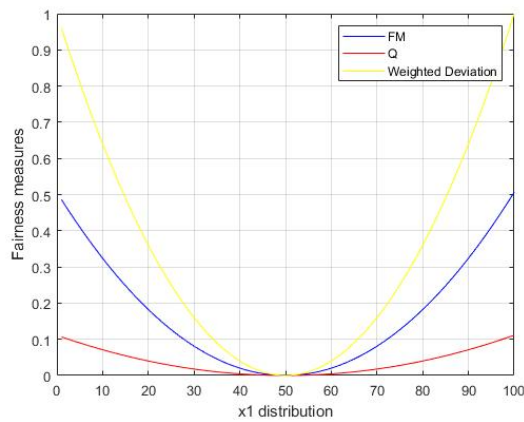


Figure 6: Instance 3

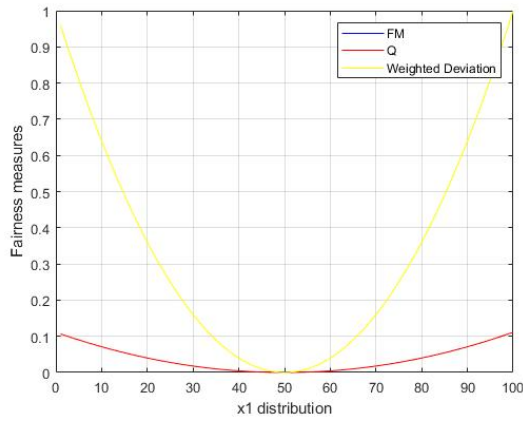


Figure 7: Instances 4 and 5

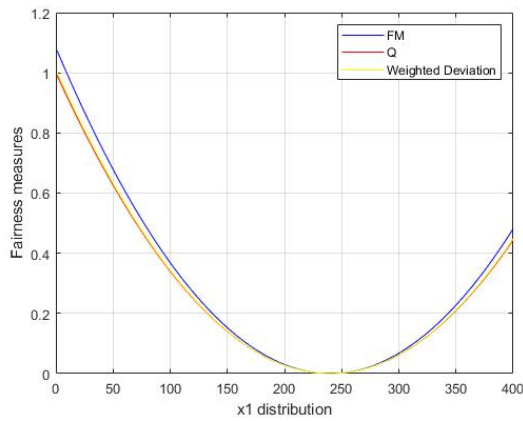


Figure 8: Instance 6

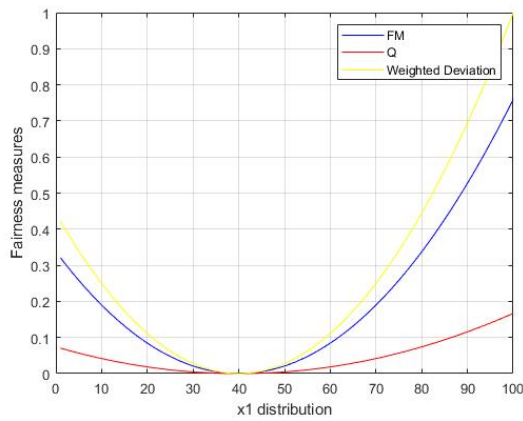


Figure 9: Instance 7

The figures above represent the values obtained from the weighted deviation, the approximation Q , and FM for all the instances covering all possible data distributions. It can be seen that in most cases our fairness model showed better results than the

approximation Q in capturing unfairness; however, the weighted deviation outdid both fairness measures since it represented an upper bound for the case of 2 clusters.

5.2. Case of 5 Clusters with a Single Binary Sensitive Attribute

In this case, 2,000 data points, which are associated with a single binary sensitive attribute, were considered. They were distributed into five clusters, i.e., $K = 5$.

Instance 1 represents an absolute fair distribution under the absence of minority groups where the total number of points is evenly divided between all clusters that have proportional representation of the sensitive attribute as in the dataset.

Instance 2 represents the presence of minority groups where one of the clusters represents a small fraction of the dataset compared to others. However, all the distributions in all the clusters are somewhat proportional to that in the dataset.

Instance 3 represents a dataset where the number of points having label X is relatively small to the number of points having label Y ; however, the distribution of the points in all clusters is equivalent to the distribution of the points in the dataset, i.e. the proportion of points having label X in the dataset is equivalent to the proportion of points having label X in all clusters, and the proportion of points having label Y in the dataset is equivalent to the proportion of points having label Y in all clusters.

Instance 4 is similar to instance 3; however, the distribution of the points in the clusters is not equivalent to the distribution of the points in the dataset.

Instance 5 represents a bias distribution of the points in the clusters where the proportion of points having label X in the clusters is not equivalent to the proportion of points having label X in the dataset, and the proportion of points having label Y in both clusters is not equivalent to the proportion of points having label Y in the dataset.

Instance 6 represents a minority group that has a bias distribution. The remaining clusters have somewhat a fair distribution, i.e., the proportion of points having label X in the dataset is somewhat equivalent to the proportion of points having label X in the remaining four clusters, and the proportion of points having label Y in the dataset is somewhat equivalent to the proportion of points having label Y in the remaining four clusters.

Instance 7 represents a minority group and a large cluster with a fair distribution; however, three other large clusters suffer from bias in the distribution of their points.

Instance 8 represents two minority groups with bias distributions. The remaining clusters have somewhat a fair distribution.

Instance 9 represents the presence of minority groups where two clusters represent a small fraction of the dataset compared to others. However, all the distributions in all the clusters are proportional to that in the dataset.

Note that for simplification, the values of y and y_i 's in table 8 were eliminated where $y = N - x$ and $y_i = n_i - x_i$.

Table 3: Instances for case 2

Instances	Bias	Minorities	N	x	n_1	x_1	n_2	x_2	n_3	x_3	n_4	x_4	n_5	x_5
1	-	-	2,000	1,000	400	200	400	200	400	200	400	200	400	200
2	-	+	2,000	1,000	100	45	500	245	450	250	550	270	400	190
3	-	-	2,000	200	400	40	400	40	400	40	400	40	400	40
4	+	-	2,000	200	400	10	400	60	400	30	400	40	400	60
5	+	-	2,000	800	400	90	300	250	350	50	450	50	500	360
6	+	+	2,000	900	100	10	500	200	450	250	550	250	400	190
7	+	+	2,000	1100	100	50	500	200	450	450	550	0	400	400
8	+	+	2,000	900	100	10	100	20	550	250	600	300	650	320
9	-	+	2,000	900	100	50	100	50	550	250	600	300	650	250

The values obtained from the balance, the approximation Q , and FM range between 0 and 1, where a value of 1 demonstrates complete fairness for the balance

notion; whereas, a value of 0 denotes absolute fairness for the approximation Q and FM . The weighted deviation indicates complete fairness when it results in 0; however, it does not have an upper level. That is why its values were normalized to the range of $[0,1]$.

Table 4 represents the values obtained from the four different fairness measures.

Similar to the case of 2 clusters, the balance was able to capture unfairness whenever we had bias (instances 4 to 8), but also resulting in some exaggerated values (values closer to 0) when we had insignificant difference between the distribution of points in the dataset and that in the clusters (instances 2 and 3). However, compared to the values obtained from the balance notion for the case of 2 clusters, this exaggeration seemed to diminish whenever more clusters were involved.

Our fairness measure FM performed well in capturing unfairness whenever we had bias whether in minority groups or in large clusters, and it highlighted this prejudice better than the weighted deviation and the approximation Q , i.e., our fairness measure FM resulted in values closer to 1 whenever we had bias in the dataset, large clusters, or minority groups. However, for instance 7, which represents fairness in two clusters out of which one represents a minority group, and bias in the three other large clusters, the weighted deviation and the approximation Q resulted in 0.946 and 0.7071 respectively; whereas, our fairness measure FM resulted in 0.5919. This proves that the weighted deviation and the approximation Q account for large clusters in their calculations more than they account for minority groups.

As mentioned previously, instance 3 represents a dataset where the number of points having label X is relatively small to the number of points having label Y ; however, the distribution of the points in all clusters is equivalent to the distribution of the points in the dataset, i.e. the proportion of points having label X in the dataset is equivalent to

the proportion of points having label X in all clusters, and the proportion of points having label Y in the dataset is equivalent to the proportion of points having label Y in all clusters. For that case, the weighted deviation, the approximation Q , and FM resulted in values of 0, thus indicating that they measure fairness as a means of proportional representation of the sensitive attribute within clusters and the dataset, unlike the balance notion which resulted in an exaggerated value of 0.111.

Table 4: Fairness evaluation for case 2

Instances	Balance	Weighted Deviation	Approximation Q	FM
1	1	0	0	0
2	0.8	0.00397	0.004	0.0051
3	0.1111	0	0	0
4	0.0257	0.05625	0.025	0.025
5	0.125	0.376	0.376	0.392
6	0.1111	0.0399	0.0379	0.1105
7	0	0.946	0.7071	0.5919
8	0.1111	0.0554	0.0428	0.153
9	0.625	0.0116	0.0097	0.0095

In conclusion, for the case of more than 2 clusters, a similar interpretation can be deduced for the balance. However, in that case, our fairness measure FM represented an upper bound for both the weighted deviation and the approximation Q , and thus exemplifying a more suitable fairness measure, with the exception of when we had fair minorities but biased large clusters, which proves that the weighted deviation and the approximation Q do not account for minor groups.

CHAPTER 6

EXPERIMENTAL STUDY

We now detail a real-world experimental study to quantify the effectiveness of our fairness measure FM . We first describe the dataset considered, followed by our results and analysis.

We consider the Customer Segmentation dataset from Kaggle repository. It contains data from a supermarket mall, and has been mainly used to identify unsatisfied customer needs. The dataset has 2,000 instances represented by 7 attributes. Among these attributes, 3 were chosen as non-sensitive attributes {age, educational level, and income}, and thus they were used for the clustering assignment. The following table illustrates some details for each of these non-sensitive attributes.

Table 5: Non-sensitive attribute description

Non-sensitive attributes	Type
Age	Continuous
Educational level	Continuous
Income	Continuous

All non-sensitive attributes were normalized prior to the clustering assignment which was performed using K -means algorithm. The data were grouped into 3 clusters ($K = 3$), and to test the fairness of this clustering assignment, the balance, weighted deviation, approximation Q , and FM were computed based on 3 sensitive attributes {gender, marital status, and occupation}, all of which were changed to binary features. The following tables illustrate the distribution of the points in the three clusters according to each sensitive attribute.

Table 6: Distribution of points with respect to gender

Gender	Total	Male	Female	Ratio Female
Cluster 1	976	431	545	0.558402
Cluster 2	158	113	45	0.28481
Cluster 3	866	538	328	0.378753
Total	2000	1082	918	0.459

Table 7: Distribution of points with respect to marital status

Marital Status	Total	Married	Single	Ratio Single
Cluster 1	976	530	446	0.456967
Cluster 2	158	67	91	0.575949
Cluster 3	866	396	470	0.542725
Total	2000	993	1007	0.5035

Table 8: Distribution of points with respect to occupation

Occupation	Total	Working	Not Working	Ratio Not Working
Cluster 1	976	444	532	0.545082
Cluster 2	158	155	3	0.018987
Cluster 3	866	768	98	0.113164
Total	2000	1367	633	0.3165

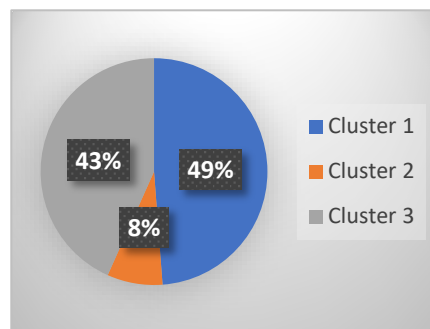


Figure 10: Distribution of points in the clusters

It can be seen that this clustering assignment represents a minority group which is illustrated in cluster 2 having only 158 points (representing 8% of the whole dataset)

compared to clusters 1 and 3 that have 976 (49% of the whole dataset) and 866 (49% of the whole dataset) points respectively.

Now, to better visualize the distribution of the data points according to each sensitive attribute, the following bar charts have been developed.

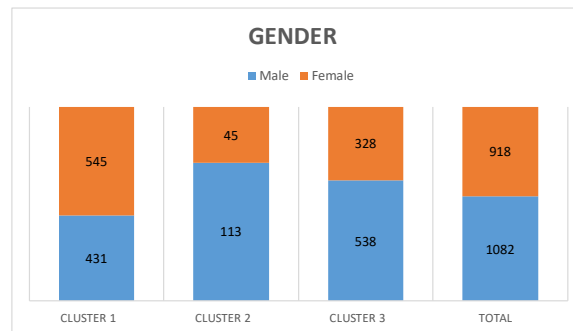


Figure 11: Distribution of points with respect to gender

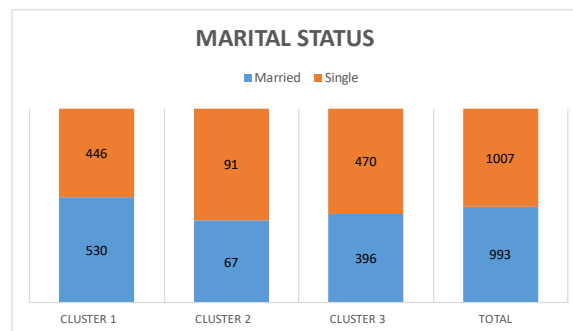


Figure 12: Distribution of points with respect to marital status

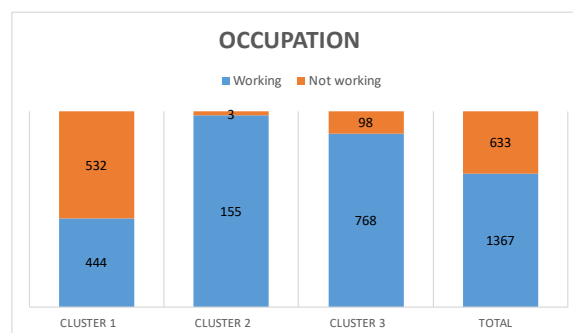


Figure 13: Distribution of points with respect to occupation

Gender as a sensitive attribute:

Figure 11 shows that the distribution of the points in each of the clusters does not perfectly reflect their distribution in the dataset, and thus representing minor biases.

Marital status as a sensitive attribute:

It can be seen in figure 12 that the distribution of points in all three clusters is approximately proportional to their distribution in the dataset, and thus representing fairness of clustering.

Occupation as a sensitive attribute:

Figure 13 reflects complete biases in the clusters, especially in the minority group.

Table 9 represents the values obtained from each fairness measure with respect to each sensitive attribute. Note again that the weighted deviation, approximation Q , and FM approach zero whenever fairness is attained, unlike the balance which approaches 1.

Table 9: Fairness measures

Fairness measure	Sensitive Attribute		
	Gender	Marital status	Occupation
Balance	0.3982	0.7363	0.0194
Deviation	1.00E-05	2.14E-06	5.04E-05
Q	0.0403	0.0086	0.2329
FM	0.0626	0.0119	0.2806

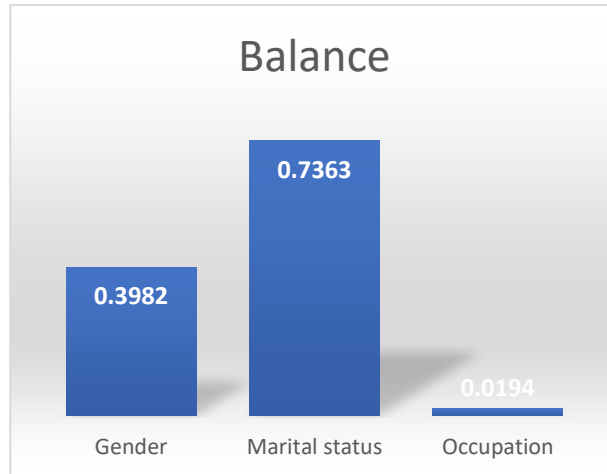


Figure 14: Balance with respect to each sensitive attribute

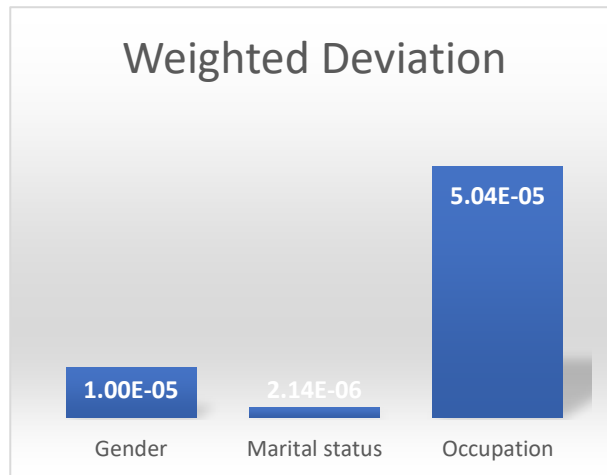


Figure 15: Weighted deviation with respect to each sensitive attribute

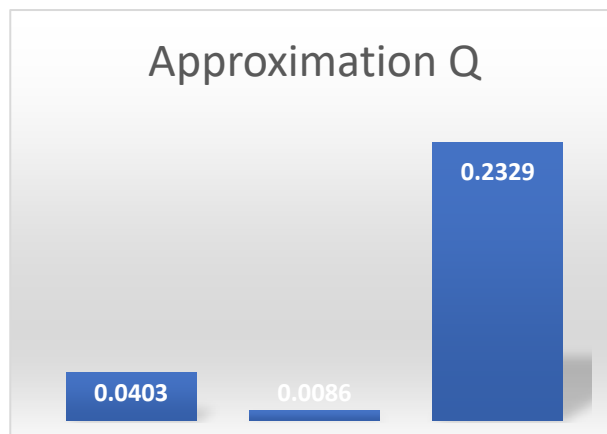


Figure 16: Approximation Q with respect to each sensitive attribute

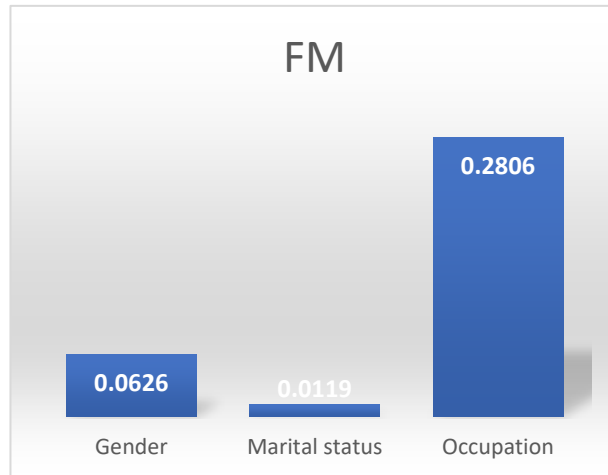


Figure 17: FM with respect to each sensitive attribute

The values obtained from each fairness measure were plotted with respect to each sensitive attribute to visualize the feature that mostly suffers from biases. It can be deduced that discrimination prevails in the sensitive attribute *occupation* since it attained the highest values in the weighted deviation, approximation Q , and FM , and the lowest value in the balance measure, thus proving that these fairness measures reflect the distribution of the points in the clusters to that in the dataset.

CHAPTER 7

CONCLUSION AND FUTURE WORK

A summary of what has been presented earlier is illustrated in the following table.

Table 10: Summary of findings

Fairness Measures	Conclusion
Balance	Too strict and not a good representation when dealing with large number of clusters
Weighted deviation	Skewed towards large clusters and does not capture unfairness in minority groups especially when dealing with large number of clusters
Approximation Q	Does not capture biases in minority groups
FM	Treats large clusters and minorities similarly where unfairness can be captured for both especially when dealing with large number of clusters

Concerning our future work, we can study relations between the fairness measures for single categorical sensitive attributes.

APPENDIX 1

CASE OF TWO CLUSTERS

$$\begin{aligned}
\mathbf{Q} &= \binom{n_1}{N} \left(\frac{\left(\frac{x_1}{n_1}\right)^2}{\frac{x}{N}} + \frac{\left(\frac{y_1}{n_1}\right)^2}{\frac{y}{N}} \right) + \binom{n_2}{N} \left(\frac{\left(\frac{x_2}{n_2}\right)^2}{\frac{x}{N}} + \frac{\left(\frac{y_2}{n_2}\right)^2}{\frac{y}{N}} \right) - \mathbf{1} \\
&= \frac{n_1}{N} \left(\frac{x_1^2 N}{x n_1^2} + \frac{y_1^2 N}{y n_1^2} \right) + \frac{n_2}{N} \left(\frac{x_2^2 N}{x n_2^2} + \frac{y_2^2 N}{y n_2^2} \right) - 1 = \frac{x_1^2}{x n_1} + \frac{y_1^2}{y n_1} + \frac{x_2^2}{x n_2} + \frac{y_2^2}{y n_2} - 1 \\
&= \frac{x_1^2}{x n_1} + \frac{(n_1 - x_1)^2}{(N - x) n_1} + \frac{(x - x_1)^2}{x(N - n_1)} + \frac{(y - y_1)^2}{(N - x)(N - n_1)} - 1 \\
&= \frac{x_1^2}{x n_1} + \frac{(n_1 - x_1)^2}{(N - x) n_1} + \frac{(x - x_1)^2}{x(N - n_1)} + \frac{(N - x - n_1 + x_1)^2}{(N - x)(N - n_1)} - 1 \\
&= \frac{1}{x n_1 (N - x)(N - n_1)} (x_1^2 (N - x)(N - n_1) + x(n_1 - x_1)^2 (N - n_1) \\
&\quad + n_1 (x - x_1)^2 (N - x) + x n_1 (N - x - n_1 + x_1) - x n_1 (N - x))(N - n_1) \\
&= \frac{1}{x n_1 (N - x)(N - n_1)} (x_1^2 (N^2 - N n_1 - x N + n_1 x) + (n_1^2 - 2 x_1 n_1 + x_1^2)(x N - x n_1) \\
&\quad + (x^2 - 2 x x_1 + x_1^2)(n_1 N - x n_1) + x n_1 (N^2 - 2 N x + x^2 - 2 N n_1 + 2 N x_1 + 2 x n_1 - \\
&\quad 2 x x_1 + n_1^2 - 2 n_1 x_1 + x_1^2) - x n_1 (N^2 - N n_1 - x N + x n_1)) \\
&= \frac{1}{x n_1 (N - x)(N - n_1)} (x_1^2 N^2 - x_1^2 N n_1 - x_1^2 x N + x_1^2 n_1 x + n_1^2 N x - x n_1^3 - 2 x_1 n_1 x N + \\
&\quad 2 x_1 x n_1^2 + x_1^2 x N - x_1^2 x n_1 + x^2 n_1 N - x^3 n_1 - 2 x x_1 n_1 N + 2 x_1^2 x n_1 + x_1^2 n_1 N - x_1^2 x n_1 + \\
&\quad x n_1^2 N - 2 N x^2 n_1 + x^3 n_1 - 2 N n_1^2 x + 2 N x_1 x n_1 + 2 x^2 n_1^2 - 2 x^2 x_1 n_1 + n_1^3 x - \\
&\quad 2 n_1^2 x_1 x + x_1^2 x n_1 - x n_1 N^2 + x n_1^2 N + x^2 n_1 N - x^2 n_1^2) = \frac{x_1^2 N^2 - 2 x_1 n_1 x N + x^2 n_1^2}{x n_1 (N - x)(N - n_1)} \\
&= \frac{(x_1 N - x n_1)^2}{x n_1 (N - x)(N - n_1)}
\end{aligned}$$

$$\begin{aligned}
\mathbf{FM} &= \binom{1}{K} \left(\frac{\left(\frac{x_1}{n_1}\right)^2}{\frac{x}{N}} + \frac{\left(\frac{y_1}{n_1}\right)^2}{\frac{y}{N}} \right) + \binom{1}{K} \left(\frac{\left(\frac{x_2}{n_2}\right)^2}{\frac{x}{N}} + \frac{\left(\frac{y_2}{n_2}\right)^2}{\frac{y}{N}} \right) - \mathbf{1} \\
&= \frac{1}{2} \left(\frac{x_1^2 N}{x n_1^2} + \frac{y_1^2 N}{y n_1^2} \right) + \frac{1}{2} \left(\frac{x_2^2 N}{x n_2^2} + \frac{y_2^2 N}{y n_2^2} \right) - 1 = \frac{x_1^2 N}{2 x n_1^2} + \frac{y_1^2 N}{2 y n_1^2} + \frac{x_2^2 N}{2 x n_2^2} + \frac{y_2^2 N}{2 y n_2^2} - 1 \\
&= \frac{x_1^2 N}{2 x n_1^2} + \frac{N(n_1 - x_1)^2}{2 n_1^2 (N - x)} + \frac{N(x - x_1)^2}{2 x (N - n_1)^2} + \frac{N(y - y_1)^2}{2 (N - x)(N - n_1)^2} - 1 \\
&= \frac{x_1^2 N}{2 x n_1^2} + \frac{N(n_1 - x_1)^2}{2 n_1^2 (N - x)} + \frac{N(x - x_1)^2}{2 x (N - n_1)^2} + \frac{N(y - y_1)^2}{2 (N - x)(N - n_1)^2} - 1 \\
&= \frac{x_1^2 N}{2 x n_1^2} + \frac{N(n_1 - x_1)^2}{2 n_1^2 (N - x)} + \frac{N(x - x_1)^2}{2 x (N - n_1)^2} + \frac{N(N - x - n_1 + x_1)^2}{2 (N - x)(N - n_1)^2} - 1 \\
&= \frac{1}{2 x n_1^2 (N - x)(N - n_1)^2} (x_1^2 N (N - x)(N - n_1)^2 + x N (n_1 - x_1)^2 (N - n_1)^2 + \\
&\quad n_1^2 N (x - x_1)^2 (N - x) + x n_1^2 N (N - x - n_1 + x_1)^2 - 2 x n_1^2 (N - x)(N - n_1)^2) \\
&= \frac{1}{2 x n_1^2 (N - x)(N - n_1)^2} (x_1^2 N (N - x)(N^2 - 2 N n_1 + n_1^2) + x N (N^2 - 2 N n_1 + n_1^2)(n_1^2 - \\
&\quad 2 n_1 x_1 + x_1^2) + (n_1^2 N (N - x)(x^2 - 2 x x_1 + x_1^2) + x n_1^2 N (N^2 - 2 x N + x^2 - 2 n_1 N + \\
&\quad 2 N x_1 + 2 x n_1 - 2 x x_1 + n_1^2 - 2 n_1 x_1 + x_1^2) - 2 x n_1^2 (N - x)(N^2 - 2 n_1 N + n_1^2))
\end{aligned}$$

$$\begin{aligned}
&= \frac{1}{2xn_1^2(N-x)(N-n_1)^2} (x_1^2N(N^3 - 2N^2n_1 + n_1^2N - xN^2 + 2xNn_1 - xn_1^2) + \\
&xN(N^2n_1^2 - 2n_1N^2x_1 + x_1^2N^2 - 2Nn_1^3 + 4Nn_1^2x_1 - 2x_1^2Nn_1 + n_1^4 - 2n_1^3x_1 + \\
&x_1^2n_1^2) + n_1^2N(Nx^2 - 2Nxx_1 + x_1^2N - x^3 + 2x^2x_1 - xx_1^2) + xn_1^2N^3 - 2x^2n_1^2N^2 + \\
&x^3n_1^2N - 2xn_1^3N^2 + 2N^2n_1^2x_1x + 2x^2n_1^3N - 2x^2n_1^2N^2x_1 + xn_1^4N - 2n_1^3N^2x_1x + \\
&x_1^2xn_1^2N - 2xn_1^2(N^3 - 2N^2n_1 + n_1^2N - xN^2 + 2Nn_1x - n_1^2x)) \\
&= \frac{1}{2xn_1^2(N-x)(N-n_1)^2} (x_1^2N^4 - 2x_1^2N^3n_1 + x_1^2N^2n_1^2 - x_1^2xN^3 + 2x_1^2xN^2n_1 - x_1^2xNn_1^2 + \\
&xN^3n_1^2 - 2x_1xN^3n_1 + x_1^2xN^3 - 2xN^2n_1^3 + 4x_1xN^2n_1^2 - 2x_1^2xN^2n_1 + xNn_1^4 - \\
&2x_1xNn_1^3 + x_1^2xNn_1^2 + x^2N^2n_1^2 - 2x_1xN^2n_1^2 + x_1^2N^2n_1^2 - x^3Nn_1^2 + 2x_1x^2Nn_1^2 - \\
&x_1^2xNn_1^2 + xN^3n_1^2 - 2x^2N^2n_1^2 + x^3Nn_1^2 - 2xN^2n_1^3 + 2x_1xN^2n_1^2 + 2x^2Nn_1^3 - \\
&2x_1x^2Nn_1^2 + xNn_1^4 - 2x_1xNn_1^3 + x_1^2xNn_1^2 - 2xN^3n_1^2 + 4xN^2n_1^3 - 2xNn_1^4 + \\
&2x^2N^2n_1^2 - 4x^2Nn_1^3 + 2x^2n_1^4) \\
&= \frac{2x_1^2N^2n_1^2 + x_1^2N^4 - 2x_1^2N^3n_1 - 4x_1xNn_1^3 + 4x_1xN^2n_1^2 - 2x_1xN^3n_1 - 2x^2Nn_1^3 + 2x^2n_1^4 + x^2N^2n_1^2}{2xn_1^2(N-x)(N-n_1)^2} \\
&= \frac{x_1^2N^2(2n_1^2 + N^2 - 2Nn_1) + x^2n_1^2(2n_1^2 + N^2 - 2Nn_1) - 2x_1xNn_1(2n_1^2 - 2Nn_1 + N^2)}{2xn_1^2(N-x)(N-n_1)^2} \\
&= \frac{(x_1^2N^2 - 2x_1xNn_1 + x^2n_1^2)(n_1^2 + n_1^2 - 2Nn_1 + N^2)}{2xn_1^2(N-x)(N-n_1)^2} \\
&= \frac{(x_1N - xn_1)^2(n_1^2 + (N - n_1)^2)}{2xn_1^2(N-x)(N-n_1)^2}
\end{aligned}$$

$$\begin{aligned}
\text{Weighted deviation} &= \left(\frac{n_1}{N}\right)^2 \left(\frac{\left(\frac{x_1}{n_1} - \frac{x}{N}\right)^2 + \left(\frac{y_1}{n_1} - \frac{y}{N}\right)^2}{n_1} \right) + \left(\frac{n_2}{N}\right)^2 \left(\frac{\left(\frac{x_2}{n_2} - \frac{x}{N}\right)^2 + \left(\frac{y_2}{n_2} - \frac{y}{N}\right)^2}{n_2} \right) \\
&= \frac{n_1}{N^2} \left(\frac{x_1^2}{n_1^2} - \frac{2xx_1}{Nn_1} + \frac{x^2}{N^2} + \frac{y_1^2}{n_1^2} - \frac{2yy_1}{Nn_1} + \frac{y^2}{N^2} \right) + \frac{n_2}{N^2} \left(\frac{x_2^2}{n_2^2} - \frac{2xx_2}{Nn_2} + \frac{x^2}{N^2} + \frac{y_2^2}{n_2^2} - \frac{2yy_2}{Nn_2} + \frac{y^2}{N^2} \right) \\
&= \frac{x_1^2}{n_1N^2} - \frac{2xx_1}{N^3} + \frac{x^2n_1}{N^4} + \frac{y_1^2}{N^2n_1} - \frac{2yy_1}{N^3} + \frac{y^2n_1}{N^4} + \frac{x^2}{N^2n_2} - \frac{2xx_2}{N^3} + \frac{x^2n_2}{N^4} + \frac{y_2^2}{N^2n_2} - \frac{2yy_2}{N^3} + \frac{y^2n_2}{N^4} \\
&= \frac{1}{N^4n_1n_2} (x_1^2N^2n_2 - 2xx_1Nn_1n_2 + x^2n_1^2n_2 + y_1^2N^2n_2 - 2yy_1Nn_1n_2 + y^2n_1^2n_2 + \\
&x_2^2N^2n_1 - 2xx_2Nn_1n_2 + x^2n_1n_2^2 + y_2^2N^2n_1 - 2yy_2Nn_1n_2 + y^2n_1n_2^2) \\
&= \frac{1}{N^4n_1n_2} (x_1^2N^2(N - n_1) - 2xx_1Nn_1(N - n_1) + xn_1^2(N - n_1) + N^2(n_1 - x_1)^2(N - \\
&n_1) - 2Nn_1(N - x)(n_1 - x_1)(N - n_1) + n_1^2(N - x)^2(N - n_1) + N^2n_1(x - x_1)^2 - \\
&2xNn_1(x - x_1)(N - n_1) + x^2n_1(N - n_1)^2 + N^2n_1(N - x - n_1 + x_1)^2 - \\
&2Nn_1(N - x)(N - x - n_1 + x_1)(N - n_1) + n_1(N - x)^2(N - n_1)^2) \\
&= \frac{2(x_1N - xn_1)^2}{N^3n_1n_2}
\end{aligned}$$

Lemma 1:

$$\frac{FM}{Q} = \frac{\frac{(x_1N - xn_1)^2(n_1^2 + (N - n_1)^2)}{2xn_1^2(N-x)(N-n_1)^2}}{\frac{(x_1N - xn_1)^2}{xn_1(N-x)(N-n_1)}} = \frac{n_1^2 + (N - n_1)^2}{2n_1(N - n_1)} = \frac{n_1^2}{2n_1(N - n_1)} + \frac{(N - n_1)}{2n_1(N - n_1)} = \frac{1}{2} \left[\frac{n_1}{n_2} + \frac{n_2}{n_1} \right]$$

$$FM = \frac{1}{2} \left[\frac{n_1}{n_2} + \frac{n_2}{n_1} \right] [Q]$$

$$\frac{Q}{\text{Weighted deviation}} = \frac{\frac{(x_1N - xn_1)^2}{xn_1(N-x)(N-n_1)}}{\frac{\frac{2(x_1N - xn_1)^2}{N^3n_1n_2}}{\frac{(x_1N - xn_1)^2}{xn_1yn_2}}} = \frac{(x_1N - xn_1)^2}{xn_1yn_2} = \frac{N^3}{2xy}$$

$$Q = \frac{N^3}{2xy} [\text{WeightedDeviation}]$$

$$\frac{FM}{\text{Weighted Deviation}} = \frac{\frac{(x_1N - xn_1)^2(n_1^2 + (N - n_1)^2)}{2xn_1^2(N-x)(N-n_1)^2}}{\frac{\frac{2(x_1N - xn_1)^2}{N^3n_1n_2}}{\frac{(x_1N - xn_1)^2}{xn_1yn_2}}} = \frac{N^3 \left(\frac{n_1}{n_2} + \frac{n_2}{n_1} \right)}{4xy}$$

$$FM = \frac{N^3 \left(\frac{n_1}{n_2} + \frac{n_2}{n_1} \right)}{4xy} [\text{WeightedDeviation}]$$

APPENDIX 2

CASE OF THREE CLUSTERS

$$\begin{aligned}
Q &= \binom{n_1}{N} \left(\frac{\left(\frac{x_1}{n_1}\right)^2}{\frac{x}{N}} + \frac{\left(\frac{y_1}{n_1}\right)^2}{\frac{y}{N}} \right) + \binom{n_2}{N} \left(\frac{\left(\frac{x_2}{n_2}\right)^2}{\frac{x}{N}} + \frac{\left(\frac{y_2}{n_2}\right)^2}{\frac{y}{N}} \right) + \binom{n_3}{N} \left(\frac{\left(\frac{x_3}{n_3}\right)^2}{\frac{x}{N}} + \frac{\left(\frac{y_3}{n_3}\right)^2}{\frac{y}{N}} \right) - 1 \\
&= \frac{n_1}{N} \left(\frac{x_1^2 N}{x n_1^2} + \frac{y_1^2 N}{y n_1^2} \right) + \frac{n_2}{N} \left(\frac{x_2^2 N}{x n_2^2} + \frac{y_2^2 N}{y n_2^2} \right) + \frac{n_3}{N} \left(\frac{x_3^2 N}{x n_3^2} + \frac{y_3^2 N}{y n_3^2} \right) - 1 \\
&= \frac{x_1^2}{x n_1} + \frac{(n_1 - x_1)^2}{n_1(N-x)} + \frac{x_2^2}{x n_2} + \frac{(n_2 - x_2)^2}{n_2(N-x)} + \frac{x_3^2}{x n_3} + \frac{(n_3 - x_3)^2}{n_3(N-x)} - 1 \\
&= \frac{1}{x n_1 n_2 n_3 (N-x)} \left(x_1^2 n_2 n_3 (N-x) + x n_2 n_3 (n_1 - x_1)^2 + x_2^2 n_1 n_3 (N-x) + \right. \\
&\quad \left. x n_1 n_3 (n_2 - x_2)^2 + x_3^2 n_1 n_2 (N-x) + x n_1 n_2 (n_3 - x_3)^2 - x n_1 n_2 n_3 (N-x) \right) \\
&= \frac{1}{x n_1 n_2 n_3 (N-x)} \left(x_1^2 N n_2 n_3 - x x_1^2 n_2 n_3 + x n_2 n_3 (n_1^2 - 2 n_1 x_1 + x_1^2) + x_2^2 N n_1 n_3 - \right. \\
&\quad \left. x x_2^2 n_1 n_3 + x n_1 n_3 (n_2^2 - x n_2 x_2 + x_2^2) + x_3^2 N n_1 n_2 - x x_3^2 n_1 n_2 + x n_1 n_2 (n_3^2 - 2 x_3 n_3 + \right. \\
&\quad \left. x_3^2) - x N n_1 n_2 n_3 + x^2 n_1 n_2 n_3 \right) \\
&= \frac{1}{x n_1 n_2 n_3 (N-x)} \left(x_1^2 N n_2 n_3 - x x_1^2 n_2 n_3 + x n_1^2 n_2 n_3 - 2 x x_1 n_1 n_2 n_3 + x x_1^2 n_2 n_3 + \right. \\
&\quad \left. x_2^2 N n_1 n_3 - x x_2^2 n_1 n_3 + x n_1 n_2^2 n_3 - 2 x x_2 n_1 n_2 n_3 + x x_2^2 n_1 n_3 + x x_3^2 n_1 n_2 - \right. \\
&\quad \left. x N n_1 n_2 n_3 + x^2 n_1 n_2 n_3 \right) \\
&= \frac{1}{x n_1 n_2 n_3 (N-x)} \left(x_1^2 N n_2 (N - n_1 - n_2) + x n_1^2 n_2 (N - n_1 - n_2) - 2 x x_1 n_1 n_2 (N - \right. \\
&\quad \left. n_1 - n_2) + x_2^2 N n_1 (N - n_1 - n_2) + x n_1 n_2^2 (N - n_1 - n_2) - 2 x x_2 n_1 n_2 (N - n_1 - n_2) + \right. \\
&\quad \left. (N n_1 n_2 (x - x_1 - x_2)^2 + x n_1 n_2 (N - n_1 - n_2)^2 - 2 x n_1 n_2 (x - x_1 - x_2) (N - n_1 - \right. \\
&\quad \left. n_2) - x N n_1 n_2 (N - n_1 - n_2) + x^2 n_1 n_2 (N - n_1 - n_2)) \right) \\
&= \frac{1}{x n_1 n_2 n_3 (N-x)} \left(x_1^2 N^2 n_2 - x_1^2 N n_1 n_2 - x_1^2 N n_2^2 + x N n_1^2 n_2 - x n_1^3 n_2 - x n_1^2 n_2^2 - \right. \\
&\quad \left. 2 x x_1 N n_1 n_2 + 2 x x_1 n_1^2 n_2 + 2 x x_1 n_1 n_2^2 + x_2^2 N^2 n_1 - x_2^2 N n_1^2 - x_2^2 N n_1 n_2 + x N n_1 n_2^2 - \right. \\
&\quad \left. x n_1^2 n_2^2 - x n_1 n_2^3 - 2 x x_2 N n_1 n_2 + 2 x x_2 n_1^2 n_2 + 2 x x_2 n_1 n_2^2 + N n_1 n_2 (x^2 + x_1^2 + x_2^2 - \right. \\
&\quad \left. 2 x x_1 - 2 x x_2 + 2 x_1 x_2) + x n_1 n_2 (N^2 + n_1^2 + n_2^2 - 2 N n_1 - 2 N n_2 + 2 n_1 n_2) - \right. \\
&\quad \left. 2 x^2 n_1 n_2 (N - n_1 - n_1) + 2 x x_1 n_1 n_2 (N - n_1 - n_2) + 2 x x_2 n_1 n_2 (N - n_1 - n_2) - \right. \\
&\quad \left. x N^2 n_1 n_2 + x N n_1^2 n_2 + x N n_1 n_2^2 + x^2 N n_1 n_2 - x^2 n_1^2 n_2 - x^2 n_1 n_2^2 \right) \\
&= \frac{1}{x n_1 n_2 n_3 (N-x)} \left(x_1^2 N^2 n_2 - x_1^2 N n_1 n_2 - x_1^2 N n_2^2 + x N n_1^2 n_2 - x n_1^3 n_2 - x n_1^2 n_2^2 - \right. \\
&\quad \left. 2 x x_1 N n_1 n_2 + 2 x x_1 n_1^2 n_2 + 2 x x_1 n_1 n_2^2 + x_2^2 N^2 n_1 - x_2^2 N n_1^2 - x_2^2 N n_1 n_2 + x N n_1 n_2^2 - \right. \\
&\quad \left. x n_1^2 n_2^2 - x n_1 n_2^3 - 2 x x_2 N n_1 n_2 + 2 x x_2 n_1^2 n_2 + 2 x x_2 n_1 n_2^2 + x^2 N n_1 n_2 + x_1^2 N n_1 n_2 + \right. \\
&\quad \left. x_2^2 N n_1 n_2 - 2 x x_1 N n_1 n_2 - 2 x x_2 N n_1 n_2 + 2 x_1 x_2 N n_1 n_2 + x N^2 n_1 n_2 + x n_1^3 n_2 + \right. \\
&\quad \left. x n_1 n_2^3 - 2 x N n_1^2 n_2 - 2 x N n_1 n_2^2 + 2 x n_1^2 n_2^2 - 2 x^2 N n_1 n_2 + 2 x^2 n_1^2 n_2 + 2 x^2 n_1 n_2^2 + \right. \\
&\quad \left. 2 x x_1 N n_1 n_2 - 2 x x_1 n_1^2 n_2 - 2 x x_1 n_1 n_2^2 + 2 x x_2 N n_1 n_2 - 2 x x_2 n_1^2 n_2 - 2 x x_2 n_1 n_2^2 - \right. \\
&\quad \left. x N^2 n_1 n_2 + x N n_1^2 n_2 + x N n_1 n_2^2 + x^2 N n_1 n_2 - x^2 n_1^2 n_2 - x^2 n_1 n_2^2 \right) \\
&= \frac{1}{x n_1 n_2 n_3 (N-x)} \left(x_1^2 N^2 n_2 - x_1^2 N n_2^2 + x_2^2 N^2 n_1 - x_2^2 N n_1^2 - 2 x x_1 N n_1 n_2 - \right. \\
&\quad \left. 2 x x_2 N n_1 n_2 + 2 x_1 x_2 N n_1 n_2 + x^2 n_1^2 n_2 + x^2 N n_2^2 \right)
\end{aligned}$$

$$\begin{aligned}
&= \frac{1}{xn_1n_2n_3(N-x)} (n_1(x_2^2N^2 - 2xx_2Nn_2 + x^2n_2^2) + n_2(x_1^2N^2 - 2xx_1Nn_1 + x^2n_1^2 - \\
&N(x_1^2n_2^2 - 2x_1x_2n_1n_2 + x_2^2n_1^2)) \\
&= \frac{1}{xn_1n_2n_3(N-x)} (n_1(x_2N - xn_2)^2 + n_2(x_1N - xn_1)^2 - N(x_1n_2 - x_2n_1)^2) \\
\mathbf{FM} &= \left(\frac{1}{K}\right) \left[\left(\frac{\left(\frac{x_1}{n_1}\right)^2}{\frac{x}{N}} + \frac{\left(\frac{y_1}{n_1}\right)^2}{\frac{y}{N}} \right) + \left(\frac{\left(\frac{x_2}{n_2}\right)^2}{\frac{x}{N}} + \frac{\left(\frac{y_2}{n_2}\right)^2}{\frac{y}{N}} \right) + \left(\frac{\left(\frac{x_3}{n_3}\right)^2}{\frac{x}{N}} + \frac{\left(\frac{y_3}{n_3}\right)^2}{\frac{y}{N}} \right) \right] - \mathbf{1} \\
&= \frac{1}{3} \left(\frac{x_1^2N}{xn_1^2} + \frac{y_1^2N}{yn_1^2} \right) + \frac{1}{3} \left(\frac{x_2^2N}{xn_2^2} + \frac{y_2^2N}{yn_2^2} \right) + \frac{1}{3} \left(\frac{x_3^2N}{xn_3^2} + \frac{y_3^2N}{yn_3^2} \right) - 1 = \frac{x_1^2N}{3xn_1^2} + \frac{(n_1-x_1)^2N}{2(N-x)n_1^2} + \frac{x_2^2N}{3xn_2^2} + \\
&\frac{(n_2-x_2)^2N}{2(N-x)n_2^2} + \frac{x_3^2N}{3xn_3^2} + \frac{(n_3-x_3)^2N}{2(N-x)n_3^2} - 1 \\
&= \frac{1}{3xn_1^2n_2^2n_3^2(N-x)} (x_1^2Nn_2^2n_3^2(N-x) + (n_1-x_1)^2Nn_2^2n_3^2x + x_2^2Nn_1^2n_3^2(N-x) + \\
&(n_2-x_2)^2Nn_1^2n_3^2x + x_3^2Nn_1^2n_2^2(N-x) + (n_3-x_3)^2Nn_1^2n_2^2x - 3xn_1^2n_2^2n_3^2(N-x)) \\
&= \frac{1}{3xn_1^2n_2^2n_3^2(N-x)} (x_1^2Nn_2^2(N-n_1-n_2)^2(N-x) + (n_1-x_1)^2Nn_2^2(N-n_1-n_2)^2x + \\
&x_2^2Nn_1^2(N-n_1-n_2)^2(N-x) + (n_2-x_2)^2Nn_1^2(N-n_1-n_2)^2x + (x-x_1-x_2)^2Nn_1^2n_2^2(N-x) + \\
&(N-n_1-n_2-x+x_1+x_2)^2Nn_1^2n_2^2x - 3xn_1^2n_2^2(N-n_1-n_2)^2(N-x)) \\
&= \frac{N[n_1(x_2N-xn_2)^2+n_2(x_1N-xn_1)^2-N(x_1n_2-x_2n_1)^2]}{3xy n_1 n_2 n_3^2} + \frac{N^2 \left[(n_3-n_1) \left(\frac{x_1-x}{n_1} \frac{x}{N} \right)^2 + (n_3-n_2) \left(\frac{x_2-x}{n_2} \frac{x}{N} \right)^2 \right]}{3xy n_3}
\end{aligned}$$

$$\begin{aligned}
\mathbf{Weighted\ deviation} &= \left(\frac{n_1}{N}\right)^2 \left(\frac{\left(\frac{x_1-x}{n_1} \frac{x}{N}\right)^2 + \left(\frac{y_1-y}{n_1} \frac{y}{N}\right)^2}{n_1} \right) + \left(\frac{n_2}{N}\right)^2 \left(\frac{\left(\frac{x_2-x}{n_2} \frac{x}{N}\right)^2 + \left(\frac{y_2-y}{n_2} \frac{y}{N}\right)^2}{n_2} \right) + \\
&\left(\frac{n_3}{N}\right)^2 \left(\frac{\left(\frac{x_3-x}{n_3} \frac{x}{N}\right)^2 + \left(\frac{y_3-y}{n_3} \frac{y}{N}\right)^2}{n_3} \right) \\
&= \frac{n_1}{N^2} \left(\frac{x_1^2}{n_1^2} - \frac{2xx_1}{Nn_1} + \frac{x^2}{N^2} + \frac{y_1^2}{n_1^2} - \frac{2yy_1}{Nn_1} + \frac{y^2}{N^2} \right) + \frac{n_2}{N^2} \left(\frac{x_2^2}{n_2^2} - \frac{2xx_2}{Nn_2} + \frac{x^2}{N^2} + \frac{y_2^2}{n_2^2} - \frac{2yy_2}{Nn_2} + \frac{y^2}{N^2} \right) + \\
&\frac{n_3}{N^2} \left(\frac{x_3^2}{n_3^2} - \frac{2xx_3}{Nn_3} + \frac{x^2}{N^2} + \frac{y_3^2}{n_3^2} - \frac{2yy_3}{Nn_3} + \frac{y^2}{N^2} \right) \\
&= \frac{x_1^2}{n_1N^2} - \frac{2xx_1}{N^3} + \frac{x^2n_1}{N^4} + \frac{y_1^2}{N^2n_1} - \frac{2yy_1}{N^3} + \frac{y^2n_1}{N^4} + \frac{x^2}{N^2n_2} - \frac{2xx_2}{N^3} + \frac{x^2n_2}{N^4} + \frac{y_2^2}{N^2n_2} - \frac{2yy_2}{N^3} + \\
&\frac{y^2n_2}{N^4} + \frac{x_3^2}{n_3N^2} - \frac{2xx_3}{N^3} + \frac{x^2n_3}{N^4} + \frac{y_3^2}{N^2n_3} - \frac{2yy_3}{N^3} + \frac{y^2n_3}{N^4} \\
&= \frac{1}{N^4n_1n_2n_3} (x_1^2N^2n_2n_3 - 2xx_1Nn_1n_2n_3 + x^2n_1^2n_2n_3 + y_1^2N^2n_2n_3 - 2yy_1Nn_1n_2n_3 + \\
&y^2n_1^2n_2n_3 + x_2^2N^2n_1n_3 - 2xx_2Nn_1n_2n_3 + x^2n_1n_2^2n_3 + y_2^2N^2n_1n_3 - \\
&2yy_2Nn_1n_2n_3 + y^2n_1n_2^2n_3 + x_3^2N^2n_1n_2 - 2xx_3Nn_1n_2n_3 + x^2n_1n_2n_3^2 + \\
&y_3^2N^2n_1n_2 - 2yy_3Nn_1n_2n_3 + y^2n_1n_2n_3^2) \\
&= \frac{1}{N^4n_1n_2n_3} (x_1^2N^2n_2(N-n_1-n_2) - 2xx_1Nn_1n_2(N-n_1-n_2) + x^2n_1^2n_2(N-n_1-n_2) \\
&+ (n_1-x_1)N^2n_2(N-n_1-n_2) - 2(N-x)(n_1-x_1)Nn_1n_2(N-n_1-n_2) + \\
&(N-x)^2n_1^2n_2(N-n_1-n_2) + x_2^2N^2n_1(N-n_1-n_2) - 2xx_2Nn_1n_2(N-n_1-n_2) + \\
&x^2n_1n_2^2(N-n_1-n_2) + (n_2-x_2)^2N^2n_1(N-n_1-n_2) - 2(N-x)(n_2-x_2)Nn_1n_2(N-n_1-n_2) \\
&+ (N-x)^2n_1n_2^2(N-n_1-n_2) + (x-x_1-x_2)^2N^2n_1n_2 -
\end{aligned}$$

$$\begin{aligned}
& 2x(x - x_1 - x_2)Nn_1n_2(N - n_1 - n_2) + x^2n_1n_2(N - n_1 - n_2)^2 + (n_3 - x_3)^2N^2n_1n_2 - 2(N - x)(n_3 - x_3)Nn_1n_2(N - n_1 - n_2) + (N - x)^2n_1n_2(N - n_1 - n_2)^2 \\
& = \frac{2[n_1(x_2N - xn_2)^2 + n_2(x_1N - xn_1)^2 - N(x_1n_2 - x_2n_1)^2]}{N^3n_1n_2n_3}
\end{aligned}$$

Lemma 2:

$$\begin{aligned}
FM &= \frac{1}{3xyn_1^2n_2^2n_3^2} (n_2^2(x_1N - xn_1)^2[(N - n_1 - n_2)^2 + n_1n_2 + n_1^2] + n_1^2(x_2N - xn_2)^2[(N - n_1 - n_2)^2 + n_1n_2 + n_2^2] - N^2(x_1n_2 - x_2n_1)^2[n_1n_2]) \\
&= \frac{1}{3xyn_1^2n_2^2n_3^2} (n_2^2(x_1N - xn_1)^2[N^2 - 3Nn_1 - 2Nn_2 + n_2^2 + 2n_1^2 + 3n_1n_2] + n_1^2(x_2N - xn_2)^2[N^2 - 2Nn_1 - 3Nn_2 + n_1^2 + 2n_2^2 + 3n_1n_2] - N^2n_1n_2(x_1n_2 - x_2n_1)^2 + Nn_1n_2^2(x_1N - xn_1)^2 + Nn_1^2n_2(x_2N - xn_2)^2) \\
&= \frac{1}{3xyn_1^2n_2^2n_3^2} (n_2^2(x_1N - xn_1)^2(n_3^2 - Nn_1 + n_1^2 + n_1n_2) + n_1^2(x_2N - xn_2)^2(n_3^2 - Nn_2 + n_2^2 + n_1n_2) + Nn_1n_2[-N(x_1n_2 - x_2n_1)^2 + n_2(x_1N - xn_1)^2 + n_1(x_2N - xn_2)^2]) \\
&= \frac{1}{3xyn_1^2n_2^2n_3^2} (n_2^2(x_1N - xn_1)^2(n_3^2 - n_1(N - n_1 - n_2)) + n_1^2(x_2N - xn_2)^2(n_3^2 - n_2(N - n_2 - n_1)) + Nn_1n_2[-N(x_1n_2 - x_2n_1)^2 + n_2(x_1N - xn_1)^2 + n_1(x_2N - xn_2)^2]) \\
&= \frac{1}{3xyn_1^2n_2^2n_3^2} (n_2^2(x_1N - xn_1)^2(n_3^2 - n_1n_3) + n_1^2(x_2N - xn_2)^2(n_3^2 - n_2n_3) + Nn_1n_2[-N(x_1n_2 - x_2n_1)^2 + n_2(x_1N - xn_1)^2 + n_1(x_2N - xn_2)^2]) \\
&= \frac{1}{3xyn_1^2n_2^2n_3^2} (n_2^2n_3(n_3 - n_1)(x_1N - xn_1)^2 + n_1^2n_3(n_3 - n_2)(x_2N - xn_2)^2 + Nn_1n_2[-N(x_1n_2 - x_2n_1)^2 + n_2(x_1N - xn_1)^2 + n_1(x_2N - xn_2)^2])
\end{aligned}$$

$$FM = \frac{N[Q]}{3n_3} + \frac{N^2[(n_3 - n_2)\left(\frac{x_2 - x}{n_2} - \frac{x}{N}\right)^2 + (n_3 - n_1)\left(\frac{x_1 - x}{n_1} - \frac{x}{N}\right)^2]}{3xyn_3}$$

$$FM = \frac{N^2}{3xyn_3} \left[\frac{N^2[\text{WeightedDeviation}]}{2} + (n_3 - N_1) \left(\frac{x_1}{n_1} - \frac{x}{N}\right)^2 + (n_3 - n_2) \left(\frac{x_2}{n_2} - \frac{x}{N}\right)^2 \right]$$

$$Q = \frac{N^3}{2xy} [\text{WeightedDeviation}]$$

REFERENCES

- Abraham, S. S., D. P., & Sundaram, S. S. (2018). Fairness in Clustering with Multiple Sensitive Attributes.
- Aggarwal, C. C., & Reddy, C. K. (2013). *Data clustering: Algorithms and applications*. CRC Press.
- Alloghani, M., Al-Jumeily, D., Mustafina, J., Hussain, A., & Aljaaf, A. J. (2019). *A systematic review on supervised and unsupervised machine learning algorithms for data science*. (M. W. Berry, A. Mohamed, & B. W. Yap, Eds.) Springer International Publishing. Retrieved from https://link.springer.com/chapter/10.1007%2F978-3-030-22475-2_1
- Alloghani, M., Al-Jumelly, D., Mustafina, J., Hussain, A., & Aljaaf, A. J. (2020). A Systematic Review on Supervised and Unsupervised Machine Learning Algorithms for Data Science. In M. W. Berry, M. Azlinah, & B. Yap, *Supervised and Unsupervised Learning for Data Science* (pp. 3-21). Springer Nature Switzerland AG.
- Araujo, T., Helberger, N., Kruijkemeier, S., & de Vreese, C. H. (2020). In AI we trust? Perceptions about automated decision-making by artificial intelligence. *AI & SOCIETY*, 35, 611-623. Retrieved from <https://doi.org/10.1007/s00146-019-00931-w>
- Backurs, A., Indyk, P., Onak, K., Schieber, B., Vakilian, A., & Wagner, T. (2019). Scalable Fair Clustering.
- Baharlouei, S., Nouiehed, M., Beirami, A., & Razaviyayn, M. (2019). Rényi Fair Inference.
- Caton, S., & Haas, C. (2020). Fairness in Machine Learning: A Survey.

- Chierichetti, F., Kumar, R., Lattanzi, S., & Vassilvitskii, S. (2017). Fair Clustering Through Fairlets.
- Dressel, J., & Farid, H. (2018, January 17). The accuracy, fairness, and limits of predicting recidivism. *Science Advances*. Retrieved from <https://www.science.org/doi/10.1126/sciadv.aao5580>
- Lowry, S., & Macpherson, G. (1988, March 5). A blot on the profession. *British Medical Journal*, *296(6623)*, 657-658. Retrieved from <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2545288/pdf/bmj00275-0003.pdf>
- Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., & Galstyan, A. (2019). A survey on bias and fairness in machine learning. Retrieved from <https://arxiv.org/pdf/1908.09635.pdf>
- Reddy, C. K., & Vinzamuri, B. (2014). A Survey of Partitional and Hierarchical Clustering Algorithms. In *Data Clustering* (pp. 87-106). CRC Press.
- Rényi, A. (1959). On measures of dependence. *Acta mathematica hungarica*, *10(3-4)*, 441-451.
- Schmidt, M., Schwiegelshohn, C., & Sohler, C. (2018). Fair Corsets and Streaming Algorithms for Fair k-means.
- Seiner, J. A. (2006). Disentangling disparate impact and separate treatment: Adapting the Canadian approach. *Yale Law & Policy Review*, *25(1)*, 95-142. Retrieved from <https://digitalcommons.law.yale.edu/cgi/viewcontent.cgi?article=1534&context=yldr>

States, S. C. (1971, March 8). *Griggs v. Duke Power Co.* 401 U.S. 424. Retrieved from <https://supreme.justia.com/cases/federal/us/401/424/>

Trautmann, S. T., & Van de Kuilen, G. (2016, December). Process fairness, outcome fairness, and dynamic consistency: Experimental evidence for risk and ambiguity. *Journal of Risk and Uncertainty*, 53(2/3), 75-88. Retrieved from <https://www-jstor-org.ezproxy.aub.edu.lb/stable/pdf/45158451.pdf?refreqid=excelsior%3Aa1c134f948cf188d438db21f0021011e>

Witsenhausen, H. S. (1975). On sequences of pairs of dependent random variables. *SIAM Journal on Applied Mathematics*, 28(1), 100-113. Retrieved from <https://www.proquest.com/docview/917552191/fulltextPDF/D859F1E3FFA044EEPQ/1?accountid=8555>

Ziko, I. M., Granger, E., Yuan, J., & Ayed, I. B. (2019). Clustering with Fairness Constraints: A Flexible and Scalable Approach .