IEEE Access
Multidisciplinary : Rapid Review : Open Access Journal

Received January 16, 2021, accepted February 4, 2021, date of publication February 10, 2021, date of current version February 24, 2021.

Digital Object Identifier 10.1109/ACCESS.2021.3058577

Histogram Ordering

MATEU SBERT[1], (Member, IEEE), COSMIN ANCUTI[2], CODRUTA O. ANCUTI[2], (Member, IEEE), JORDI POCH[1], SHUNING CHEN[3], AND MARIUS VILA[1]

[1]Institute of Informatics and Applications, University of Girona, 17004 Girona, Spain
[2]Facultatea de Electronică, Telecomunicaţii şi Tehnologii Informaţionale (ETcTI), Universitatea Politehnica Timişoara, 300006 Timişoara, Romania
[3]Urban Institute and the Department of Civil Engineering, School of Engineering, Kyushu University, Fukuoka 819-0395, Japan

Corresponding author: Mateu Sbert (mateu.sbert@gmail.com)

This work was supported in part by the Romanian Ministry of Education and Research, CNCS—UEFISCDI (PNCDI III), under Project PN-III-P1-1.1-TE-2019-1111 and Project PN-III-P2-2.1-PED-2019-2805. The work of Mateu Sbert and Jordi Poch was supported in part by the Spanish Ministry for Science and Innovation under Grant PID2019-106426RB-C31.

ABSTRACT Frequency histograms are ubiquitous, being practically used in any field of science. In this paper, we present a partial order for frequency histograms and, to our knowledge, no order of this kind has been yet defined. This order is based on the stochastic order of discrete probability distributions and it has invariance properties that make it unique. First, we model a frequency histogram as a sequence of bins associated with a discrete probability (or relative frequency) distribution. Then, we consider that two histograms are ordered if they are defined on the same sequence of bins and their respective frequency distributions are stochastically ordered. The ordering can be easily spotted because the respective cumulative distribution functions of the frequencies of two ordered histograms do not cross each other. Finally, with each bin we can associate a representative value of the bin, and for two ordered histograms it holds that all quasi-arithmetic means (such as arithmetic, harmonic, and geometric mean) of the representative values weighted by the frequencies are ordered in the same direction than the histograms are. Our theoretical study is supported by three experiments in the fields of image processing, traffic flow, and income distribution.

INDEX TERMS Histograms, ordering, quasi-arithmetic mean, traffic flow, histogram equalization, income distribution.

I. INTRODUCTION

Histograms were introduced by Karl Pearson in [1]. Since then, they have become ubiquitous in all branches of knowledge, to visually represent the distribution of occurrences, or frequency, between different values of a variable. To each value there corresponds a rectangular "bin", which height gives the corresponding frequency. If the variable is continuous, its range is usually divided into equal intervals, and a representative value for each segment is given. When we consider relative frequencies we speak of a frequency histogram. We will consider in this paper frequency histograms representing the distribution of a numerical variable.

We present in this paper a partial order for frequency histograms. To our knowledge, no order has been yet defined on frequency histograms, neither on histograms in general. The order is based on the underlying order of the frequency distributions of the histograms. This order is such that, for any two ordered histograms, any quasi-arithmetic mean [2], [3] calculated on the same sequence of representative values and weighted by the frequencies, keep the same order. For instance, if one histogram precedes another one, then the arithmetic, harmonic and geometric means will be greater for this histogram than for the second histogram. This property will hold whichever representative values we use for the bins or intervals. And the order is preserved if the bins are clustered.

Quasi-arithmetic, also called Kolmogorov or Kolmogorov-Nagumo, means are ubiquitous in many branches of science. In information theory, Rényi [4] defined axiomatically the entropy of a probability distribution as a Kolmogorov mean of the information conveyed by the results, where $-\log p_k$ is the information conveyed by result k with probability p_k. In economics, since long time it has been discussed about which mean was more accurate for a particular problem [5], and some problems are represented by harmonic mean rather than arithmetic mean, like the price earning ratio, P/E. Also, the power mean, depending on a parameter $\eta > 0$ which gives

The associate editor coordinating the review of this manuscript and approving it for publication was Badri Narayan Subudhi.

the elasticity of substitution among different types of labor, is implicitly used for the aggregate labor demand (power mean with power equal to $\frac{\eta-1}{\eta}$), and for its corresponding wage (power mean with power equal to $1-\eta$) [6], pag. 30.

Sometimes a problem is represented by more than one mean. Interesting examples of two mean representations, arithmetic and harmonic, come from Physics, where springs added in series combine harmonically, and in parallel arithmetically, while resistors in parallel combine harmonically, and in series arithmetically. Another example is traffic flow [7] where two different means for speeds are meaningful for the problem at hand, the arithmetic mean, or time mean speed, and harmonic mean, or space mean speed. In [8] both geometric and harmonic mean are used in addition to arithmetic mean to improve noise source maps. And weighted quasi-arithmetic means have been shown [9] to be related to the optimal variance of the multiple importance sampling technique in monte carlo. See also [10] for an interesting discussion on Kolmogorov means.

Recently, an invariance property for quasi-arithmetic means of a sequence of numbers when weights, or frequencies, changed was described in [11]. When frequencies were ordered [12], all means were ordered in the same direction too. But to apply, the invariance property requires the sequence of numbers to be in increasing order. This apparent restriction does not happen in a frequency histogram, where by construction the representative values in the bins are in increasing order. This will allow us to define the histogram order by inheriting the order of the underlying frequencies. We will show three applications of this order, to traffic flow [7], [13], where histogram order explains inconsistencies in the variation of the two different means used to measure mean speed between two spots; to image processing, where we will consider the order between the original image intensity histogram and its equalized version [14]; and to income distribution, where we will compare histogram ordering with the classic Lorentz order [15] and Gini coefficient [16] to measure income inequalities.

This paper is organized in the following way. After the introduction, Section II recalls the definition of quasi-arithmetic mean and introduces the invariance properties. In Section III we define the order between histograms and give its properties. Section IV-A recalls the time speed and space speed means used in traffic flow. In Section IV-B we show the relationship of our order with histogram equalization in image processing. In Section IV-C we show the application to the study of the income per capita.

II. PREVIOUS WORK: INVARIANCE PROPERTY OF QUASI-ARITHMETIC MEANS

A. QUASI-ARITHMETIC, KOLMOGOROV OR KOLMOGOROV-NAGUMO MEAN

The weighted Kolmogorov, or Kolmogorov-Nagumo, or quasi-arithmetic mean [2], [3] of sequence $\{b_k\}$ is defined as

$$h^{-1}\Big(\sum_{k=1}^{n}\alpha_k h(b_k)\Big), \tag{1}$$

where $h(x)$, called the Kolmogorov-Nagumo function, is a strictly monotonic function (and thus invertible) with range a subset of $\mathcal{R}$, with inverse function $h^{-1}(x)$, and the $\{\alpha_k\}_{k=1}^{n}$ are positive weights adding to 1. Observe that, for the mean to be defined, for all indexes k the values $h(b_k)$ have to be defined. The function $h(x)$ is called the Kolmogorov-Nagumo function.

The Kolmogorov mean contains all classic means, for instance, for $h(x)=x$ we have the weighted arithmetic mean, $h(x)=1/x$ the harmonic mean, $h(x)=\log(x)$ the geometric mean, and for $h(x)=x^r$ the family of power means.

B. QUASI-ARITHMETIC MEAN INVARIANCE PROPERTY

The fact that a problem can be represented by different weighted means makes us consider whether the order between means is kept when the weights change. The following invariance property was obtained in Sbert and Poch [11] (see Appendix for a direct proof):

Theorem 1: Let $h(x)$ be an invertible strictly monotonic function, with inverse function $h^{-1}(x)$. Consider a sequence of positive weights $\{\alpha_k\}_{k=1}^{n}$ and $\{\alpha'_k\}_{k=1}^{n}$, $\sum_{k=1}^{n}\alpha_k = \sum_{k=1}^{n}\alpha'_k = 1$. Then conditions a) and b) are equivalent:
a)

$$\begin{aligned}
\alpha'_1 &\geq \alpha_1\\
\alpha'_1+\alpha'_2 &\geq \alpha_1+\alpha_2\\
&\vdots\\
\alpha'_1+\ldots+\alpha'_{n-1} &\geq \alpha_1+\ldots+\alpha_{n-1}\\
\alpha'_1+\ldots+\alpha'_{n-1}+\alpha'_n &= \alpha_1+\ldots+\alpha_{n-1}+\alpha_n
\end{aligned} \tag{2}$$

b) For any increasing sequence of real numbers $\{b_k\}_{k=1}^{n}$ it holds:

$$h^{-1}\Big(\sum_{k=1}^{n}\alpha'_k h(b_k)\Big) \leq h^{-1}\Big(\sum_{k=1}^{n}\alpha_k h(b_k)\Big). \tag{3}$$

Condition a) defines first stochastic dominance or first stochastic order between distributions [12], [17] and we write it as $\{\alpha_k\} \succ_{FSD} \{\alpha'_k\}$ or $\{\alpha'_k\} \prec_{FSD} \{\alpha_k\}$.
In the Appendix we prove the following corollary

Corollary 1: If $\{\alpha_k\} \succ_{FSD} \{\alpha'_k\}$ then $\alpha'_1 \geq \alpha_1$ and $\alpha'_n \leq \alpha_n$

The following corollary tells us that the order invariance extends to all means.

Corollary 2: If $\{\alpha_k\} \succ_{FSD} \{\alpha'_k\}$ then all quasi-arithmetic means of an increasing sequence of numbers $\{b_k\}_{k=1}^{n}$ with weights $\{\alpha_k\}$ will precede the mean with weights $\{\alpha'_k\}$.

Proof: If $\{\alpha_k\} \succ_{FSD} \{\alpha'_k\}$ then condition a) of Theorem 1 is true independently of the mean considered, then for any mean such that it is defined over the sequence $\{b_k\}_{k=1}^{n}$ condition b) holds. □

Eq. 2 means that the cumulative distribution function (CDF) of $\{\alpha_k\}_{k=1}^n$ will be always under the CDF of $\{\alpha'_k\}_{k=1}^n$. Geometrically, the CDF of the dominating distribution $\{\alpha_k\}_{k=1}^n$ lies under the CDF of $\{\alpha'_k\}_{k=1}^n$.

The order $\succ_{FSD}$ defines a partial order between the distributions, as it can be easily seen from Eq. 2 that this order is a reflexive, antisymmetric and transitive relation. It is a partial order in general, and only a total order for $n = 2$, where $\{\alpha'_k\} \prec_{FSD} \{\alpha_k\}$ *iff* $\alpha'_1 \geq \alpha_1$. For $n > 2$, consider the distributions $\{\alpha'_k\} = \{\frac{1}{n} - \epsilon, \frac{1}{n} + \epsilon, \ldots, \frac{1}{n}, \frac{1}{n} + \epsilon, \frac{1}{n} - \epsilon\}$ and $\{\alpha_k\} = \{\frac{1}{n} + \epsilon, \frac{1}{n} - \epsilon, \ldots, \frac{1}{n}, \frac{1}{n} - \epsilon, \frac{1}{n} + \epsilon\}$, where $\frac{1}{n} > \epsilon > 0$. According to Corollary 1, neither $\{\alpha'_k\} \prec_{FSD} \{\alpha_k\}$ nor $\{\alpha_k\} \prec_{FSD} \{\alpha'_k\}$.

Observe that the weights $\{\alpha_k\}_{k=1}^n$ can be interpreted as a discrete probability distribution, or probability mass function (PMF), or also as a frequency distribution. Inversely, a PMF or a frequency distribution can be used as weights. This double interpretation, as weights and as frequencies, will be used in next Section.

III. HISTOGRAM ORDERING

We first define formally a frequency histogram (for economy of notation we will use just "histogram" within the context of this paper). Given a real variable X with values in $D = [l_{min}, l_{max}]$, we divide D into a sequence of n intervals, $[l_{min} = l_0, l_1[, [l_1, l_2[, \ldots, [l_{n-1}, l_{max} = l_n]$. We also allow for the possibility of $l_0 = -\infty$ and/or $l_n = +\infty$, and we do not require the intervals to be of equal length, although they can be graphically represented as of equal length, see Fig. 11(a).

Given a distribution of frequencies $\boldsymbol{\alpha} = \{\alpha_k\}_{k=1}^n$, we associate interval $[l_{i-1}, l_i[$ with frequency α_i.

Definition 1: Given a sequence of intervals $\mathbf{l} = \{l_k\}_{k=0}^n$ *and a frequency distribution* $\boldsymbol{\alpha} = \{\alpha_k\}_{k=1}^n$, *histogram* $\mathcal{H}(\mathbf{l}, \boldsymbol{\alpha})$ *is a mapping* $\mathcal{H} : \mathbf{l} \to \alpha$ *such that for all* $1 \leq i \leq n$, $\mathcal{H}([l_{i-1}, l_i[) = \alpha_i$.

Observe that we set the upper boundaries of the intervals open so they do not overlap with each other.

Definition 2: We say that sequence of real numbers $\mathbf{b} = \{b_k\}_{k=1}^n$ *is a representative sequence for histogram* $\mathcal{H}(\mathbf{l}, \boldsymbol{\alpha})$ *when for all i,* $b_i \in [l_{i-1}, l_i[$.

Observe that the $\{b_k\}_{k=1}^n$ numbers form by definition an increasing sequence, and we do not restrict them to be positive. Using $\mathbf{b}$ we can define the mean of a histogram, extending the quasi-arithmetic transform from sequences to histograms.

Definition 3: Given a strictly monotonic function $h(x)$, *with inverse function* $h(x)^{-1}$, *the quasi-arithmetic mean of histogram* $\mathcal{H}(\mathbf{l}, \boldsymbol{\alpha})$, $\mathbf{l} = \{l_k\}_{k=0}^n$, $\boldsymbol{\alpha} = \{\alpha_k\}_{k=1}^n$, *with representative values* $\mathbf{b} = \{b_k\}_{k=1}^n$, *is the quasi arithmetic mean of sequence* $\mathbf{b}$ *with weights* $\boldsymbol{\alpha}$, *this is,* $h^{-1}\left(\sum_{k=1}^n \alpha_k h(b_k)\right)$.

We will compare two histograms based on the same sequence of intervals $\mathbf{l}$:

Definition 4: Two histograms $\mathcal{H}(\mathbf{l}, \boldsymbol{\alpha}), \mathcal{H}(\mathbf{l}, \boldsymbol{\alpha'})$, *will be ordered, and we write* $\mathcal{H}(\mathbf{l}, \boldsymbol{\alpha}) \succ \mathcal{H}(\mathbf{l}, \boldsymbol{\alpha'})$ *or* $\mathcal{H}(\mathbf{l}, \boldsymbol{\alpha}) \prec \mathcal{H}(\mathbf{l}, \boldsymbol{\alpha'})$ *iif* $\boldsymbol{\alpha} \succ_{FSD} \boldsymbol{\alpha'}$ *or* $\boldsymbol{\alpha'} \prec_{FSD} \boldsymbol{\alpha}$, *respectively.*

A. PROPERTIES

We enumerate here the properties of the defined order, based on the underlying order $\succ_{FSD}$ between frequencies:

1) The order $\succ$ between histograms is a partial order, i.e. it is reflexive, antisymmetric and transitive, and is a total order only for the binary case. This is because histogram order inherits the order properties of $\succ_{FSD}$.
2) Any clustering of neighbour bins, i.e., merging of intervals, done on two ordered histograms will keep the order. Observe that any clustering of neighbour values (i.e. adding the corresponding weights) will keep Eq. 2 true, thus, if $\boldsymbol{\alpha} \succ_{FSD} \boldsymbol{\alpha'}$, and $\boldsymbol{\alpha}^c, \boldsymbol{\alpha'}^c$ are the clustered distributions then $\boldsymbol{\alpha}^c \succ_{FSD} \boldsymbol{\alpha'}^c$, and thus the clustered histograms will be ordered.
3) Repeated clustering of bins, i.e., merging of intervals, done on two non-ordered histograms, will eventually result in ordered histograms, as binary histograms are always ordered because for any binary distribution Eq. 2 is always true.
4) If two histograms are ordered, all quasi-arithmetic means of the same representative values $\mathbf{b}$ for both histograms will keep the same order. This happens because from Definition 4, if two histograms are ordered the respective frequency distributions are ordered, and then we can apply Corollary 2.
5) If two histograms are ordered, any change in representative values from sequence $\mathbf{b}$ to a new sequence $\mathbf{b}'$ will not affect the order of the means. Observe first that $\mathbf{b}'$ is also an increasing sequence. As the respective frequency distributions of the histograms are ordered, we have can apply Theorem 1, condition b) to the sequence $\mathbf{b}'$.
6) If we refine two ordered intervals, or change its limits, in identical ways for both histograms, we can not ensure in general that the resulting histograms are still ordered. However, if the two histograms have underlying continuing distributions, changing the limits of the bins, in particular refining the histogram, should not affect the order.
7) Two histograms $\mathcal{H}(\mathbf{l}, \boldsymbol{\alpha}), \mathcal{H}(\mathbf{l}', \boldsymbol{\alpha'})$ are ordered $\iff$ $\mathbf{l} \equiv \mathbf{l}'$ and the cumulative distribution functions of $\boldsymbol{\alpha}$ and $\boldsymbol{\alpha'}$ do not cross each other, see the two CDFs in Fig. 1 bottom right.
8) Histogram order is invariant to scaling. This is, if two histograms $\mathcal{H}(\mathbf{l}, \boldsymbol{\alpha}), \mathcal{H}(\mathbf{l}, \boldsymbol{\alpha'})$ are ordered and $f(x)$ is a strictly increasing function then the histograms $\mathcal{H}(\mathbf{f}(\mathbf{l}), \boldsymbol{\alpha}), \mathcal{H}(\mathbf{f}(\mathbf{l}), \boldsymbol{\alpha'})$ keep the order, where $\mathbf{f}(\mathbf{l}) = \{f(l_k)\}_{k=0}^n$. Observe that representative values $\mathbf{b}$ are mapped to $\mathbf{f}(\mathbf{b}) = \{f(b_k)\}_{k=1}^n$. One example of this property is that representing luminances in the interval $[0, 1]$ or in $[0, 255]$ should not change the

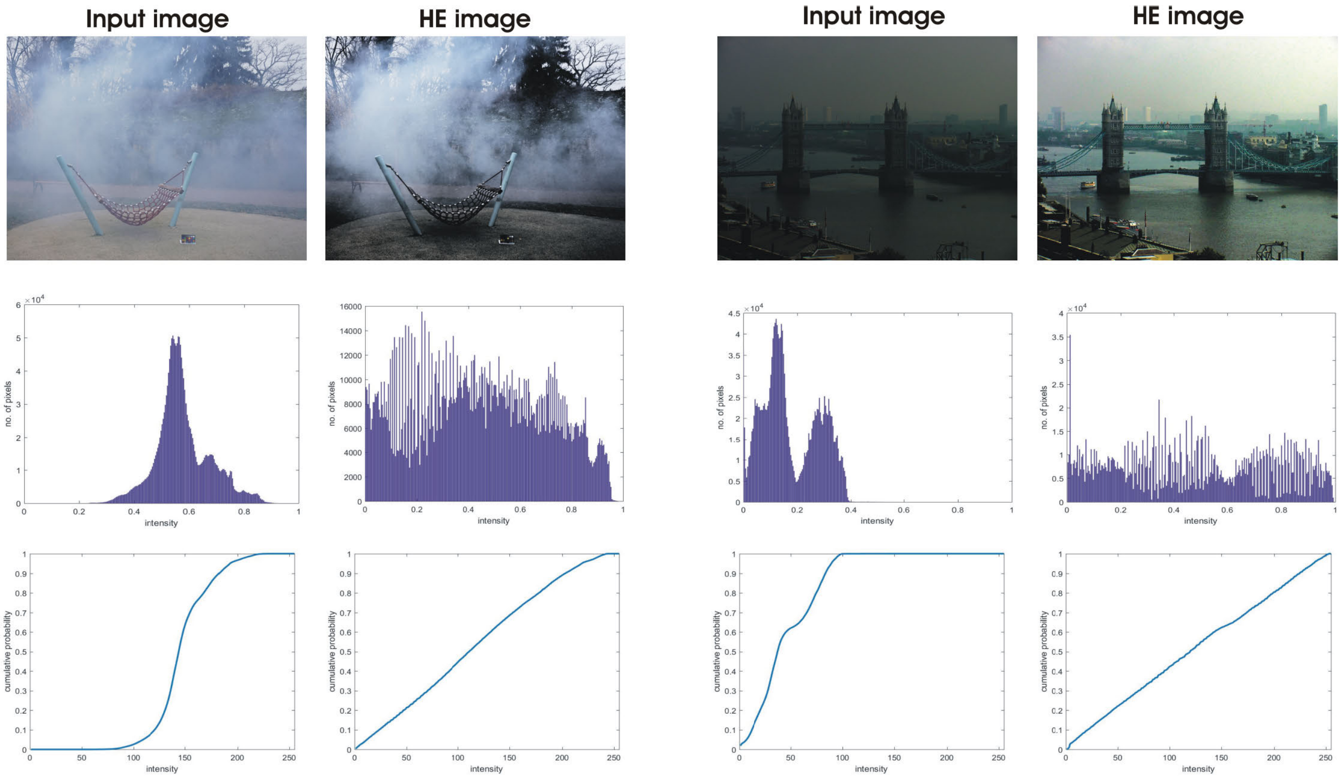


FIGURE 1. Left: Histogram equalization (HE) is not effective due to the shape of the CDF (bottom row) that presents some variations around the main diagonal. Right: HE is effective, CDF of original image is over the equalized one, the histograms are ordered.

order of the luminance frequency histograms (see Section IV-B), neither should changing to a logarithmic scale, for instance by tone-mapping, using function $f(x) = \log(x+1)$. Or a change is speed units should not alter the order of speed frequency histograms (see Section IV-A), a change in currency should not change the order of income distribution histograms (see Section IV-C), or a change from Celsius to Fahrenheit should not alter the order of temperature histograms.

9) If two histograms $\mathcal{H}(\mathbf{l}, \boldsymbol{\alpha})$, $\mathcal{H}(\mathbf{l}, \boldsymbol{\alpha'})$ are ordered and $f(x)$ is a strictly decreasing function then the histograms $\mathcal{H}(\mathbf{f}(\mathbf{l}^\star), \boldsymbol{\alpha}^\star)$, $\mathcal{H}(\mathbf{f}(\mathbf{l}^\star), \boldsymbol{\alpha'}^\star)$ are ordered inversely, where $\boldsymbol{\alpha}^\star = \{\alpha_{n-k+1}\}_{k=1}^n$, $\boldsymbol{\alpha'}^\star = \{\alpha'_{n-k+1}\}_{k=1}^n$, $\mathbf{f}(\mathbf{l}^\star) = \{f(l_{n-k+1})\}_{k=0}^n$. This is because if $\boldsymbol{\alpha} \succ_{FSD} \boldsymbol{\alpha'}$ then $\boldsymbol{\alpha}^\star \prec_{FSD} \boldsymbol{\alpha'}^\star$ and vice versa. Observe that representative values $\mathbf{b}$ are mapped to $\mathbf{f}(\mathbf{b}) = \{f(b_{n-k+1})\}_{k=1}^n$. An example of this property is that if two luminance histograms (see Section IV-B) are ordered, the histograms of the negative images, obtained for luminances in $[0, 1]$ with the function $f(x) = 1 - x$, are ordered in the inverted sense.

IV. EXAMPLES

A. TIME SPEED AND SPACE SPEED IN TRAFFIC FLOW

In traffic flow, two mean speeds are used to measure a mean speed between two spots. Time speed v_t is defined as the arithmetic mean of speeds, while space speed, v_s, is defined as the harmonic mean of speeds [7], [13]. When considering the possible speeds divided in bins, both speeds are weighted means. If $\{\alpha_i\}$ are the relative frequencies of the speed intervals, $\sum_{i=1}^n \alpha_i = 1$, and the $\{v_i\}$ are the representative speeds, then

$$v_t = \sum_{i=1}^n \alpha_i v_i, \tag{4}$$

and

$$v_s = \frac{1}{\sum_{i=1}^n \alpha_i \frac{1}{v_i}}. \tag{5}$$

The set of relative frequencies $\{\alpha_i\}$ of speeds $\{v_i\}$ define the traffic conditions between the given spots, with mean speeds v_s and v_t. Changes in these frequencies, with a new distribution $\{\alpha'_i\}$, lead to new mean speeds v'_t, v'_s. The speeds can change in any order, but if the relative frequencies $\{\alpha_i\}$ and $\{\alpha'_i\}$ hold Eq. 2, this is, $\{\alpha_k\} \succ_{FSD} \{\alpha'_k\}$, then both speeds will change in the same direction. Observe also that if Eq. 2 hold, then according to Theorem 1, invariance also happens for any other quasi-arithmetic mean, defined as Eq. 1. For instance, we could consider using the geometric mean of speeds, say v_g,

$$v_g = \exp\Big(\sum_{i=1}^n \alpha_i \log v_i\Big), \tag{6}$$

which we know is always between harmonic and arithmetic mean, $v_s \leq v_g \leq v_t$, and we would have too $v'_g \leq v_g$.

1) SPEED HISTOGRAMS

We show in Figs. 2–7 six examples corresponding to the distributions of speeds in two different instants of time. In all the cases, the first distribution (in blue color) is the same, and what changes is the second distribution (in orange). In Figs. 2&3, the direction of change of v_t and v_s is reversed. In Figs. 4–7, the direction of change is the same. The relationship between frequency distributions in Figs. 2&3 and Figs. 6&7 does not hold Eq. 2, thus changes in the mean speed can be in any order, while in Figs. 4&5, Eq. 2 is held and the histograms are ordered, thus changes in mean speed have to follow the same order. The mean speed values for Figs. 2–7 are the following, where v_s, v_t correspond to the first distribution, and v'_t, v'_s to the second one:

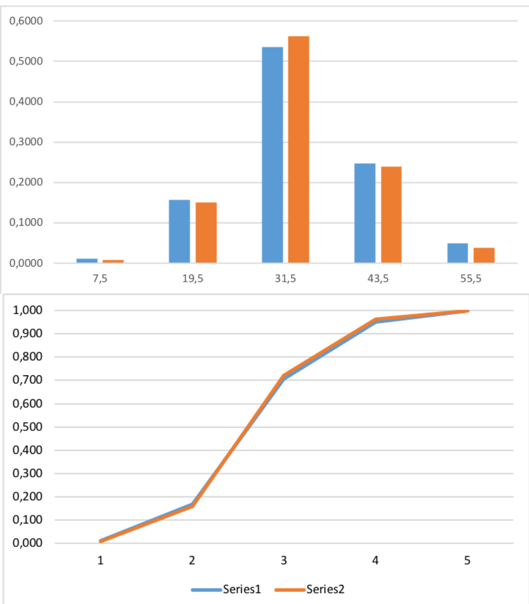


FIGURE 2. **Top: Histograms of frequency distributions of speeds at two times. v_t decreases, but v_s increases. Bottom: the CDFs cross each other.**

For Fig. 2:

$$v'_t - v_t = -0,2133, \quad v'_s - v_s = 0,1327$$

For Fig. 3:

$$v'_t - v_t = 0,2133, \quad v'_s - v_s = -0,1985$$

For Fig. 4:

$$v'_t - v_t = 0,5333, \quad v'_s - v_s = 0,5783$$

For Fig. 5:

$$v'_t - v_t = -0,3874, \quad v'_s - v_s = -0,6012$$

For Fig. 6:

$$v'_t - v_t = 0,64, \quad v'_s - v_s = 0,0379$$

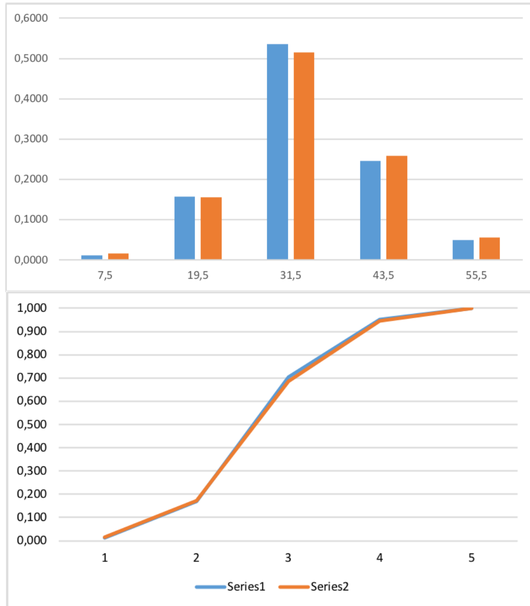


FIGURE 3. **Top: Histograms of frequency distributions of speeds at two times. v_t increases, but v_s decreases. Bottom: the CDFs cross each other.**

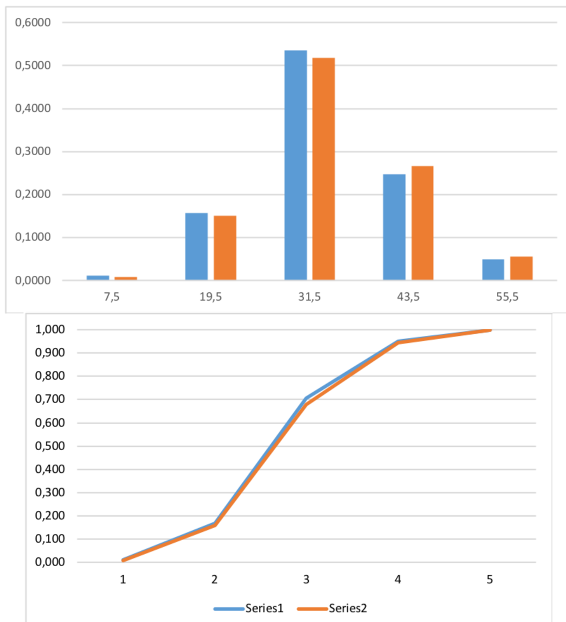


FIGURE 4. **Top: Histograms of frequency distributions of speeds at two times. Both v_t and v_s increase. Bottom: the CDFs do not cross each other.**

For Fig. 7:

$$v'_t - v_t = -0,0266, \quad v'_s - v_s = -0,1877$$

B. LUMINANCE HISTOGRAM EQUALIZATION

Histogram equalization (HE) is a well-known technique to adjust the contrast of an image [14]. This is achieved by modifying the original intensity distribution in order to obtain a more uniform distribution. It consists in changing the luminance of the original image so that the CDF of the

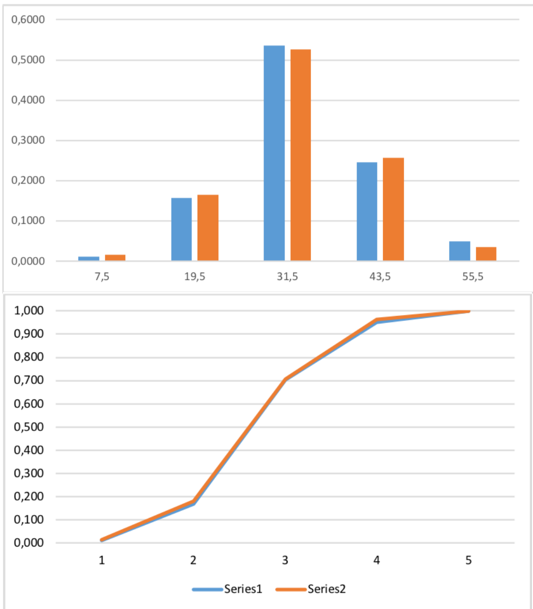


FIGURE 5. Top: Histograms of frequency distributions of speeds at two times. Both v_t and v_s decrease. Bottom: the CDFs do not cross each other.

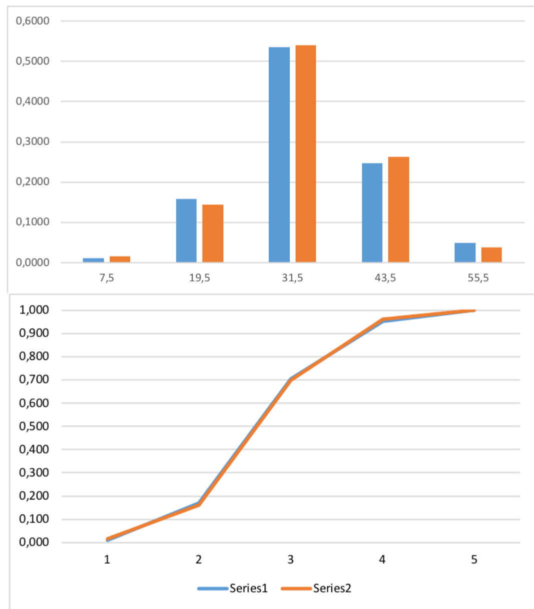


FIGURE 7. Top: Histograms of frequency distributions of speeds at two times. Both v_t and v_s decrease. Bottom: the CDFs cross each other.

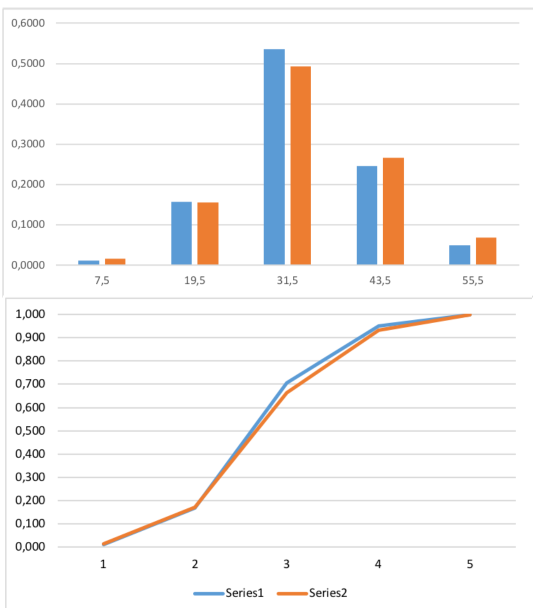


FIGURE 6. Top: Histograms of frequency distributions of speeds at two times. Both v_t and v_s increase. Bottom: the CDFs cross each other.

transformed image is a linearized function, i.e., for a luminance value i, where i ranges between 0 and 255, $CDF(i) = iK$ so CDF is a linearized function. For a color RGB image, it is first converted to HSV space and equalization is done on the luminance component.

Arici *et al.* [18] and Toet and Wu [19] discuss some cases where HE does not work well. If the luminance is concentrated in a small range or there is a spike when the image has a large low intensity background, some regions are overenhanced and appearing unnatural, and noise can be enhanced too. To overcome this problem they propose preprocessing the luminance histogram so that the modified histogram is closer to a uniformly distributed histogram. On the other hand, we do not introduce in this paper a new preprocessing method for HE, but study the relationship of HE with histogram ordering.

An interesting observation is related to the shape of the CDF of the original images. In general, the CDF of the original image looks like a logarithmic function. However, a special case is represented by the images with the shape of the CDF that present some variation around the main diagonal (see Figure 8 and Fig. 1 left). Observe that this means the histograms corresponding to the original image and the equalized image are not ordered. In such cases, applying HE is shown not to be effective and the resulted images are not enhanced properly (the contrast is not improved significantly or the images are distorted). As can be seen in Figure 8 the histograms of such images show a significant gap in the frequencies. This gap partially remains also in the histograms of the images processed by histogram equalization (HE).

Examples with usual images where the histogram equalization works quite well are shown in Figure 9 and Fig. 1 right. The first image (from left to right) shown in Figure 9 is an underwater scene. As can be seen the walls of the underwater cave are almost completely dark but after applying the HE the details are better revealed and the local contrast in general is better restored. The third image in Figure 9 is an outdoor landscape. After applying the HE the contrast of the picture is slightly improved and therefore the shadows of the clouds and the forest details look more natural compared with the original picture. The middle image in Figure 9 represents an over-exposed building scene. This could be observed also by

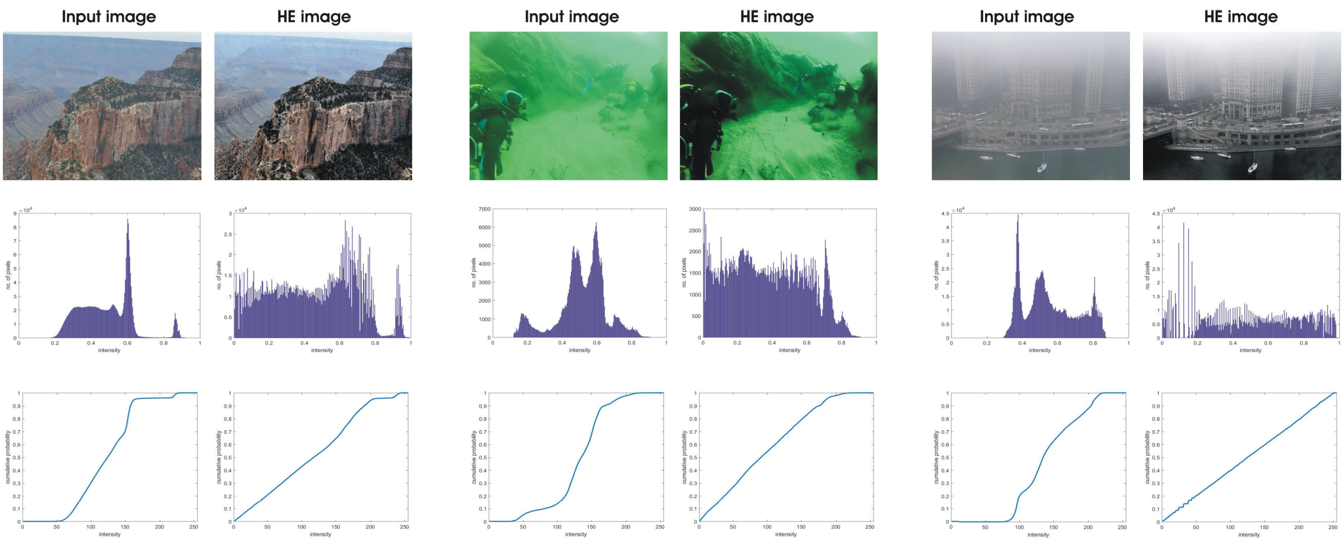


FIGURE 8. Three examples where the HE (top row) is not effective due to the shape of the CDF (bottom row) that presents some variations around the main diagonal. The middle row shows the corresponding histograms of the input images and the images processed by the HE. Observe that the histograms of original images and of corresponding equalized images are not ordered (the CDFs cross each other).

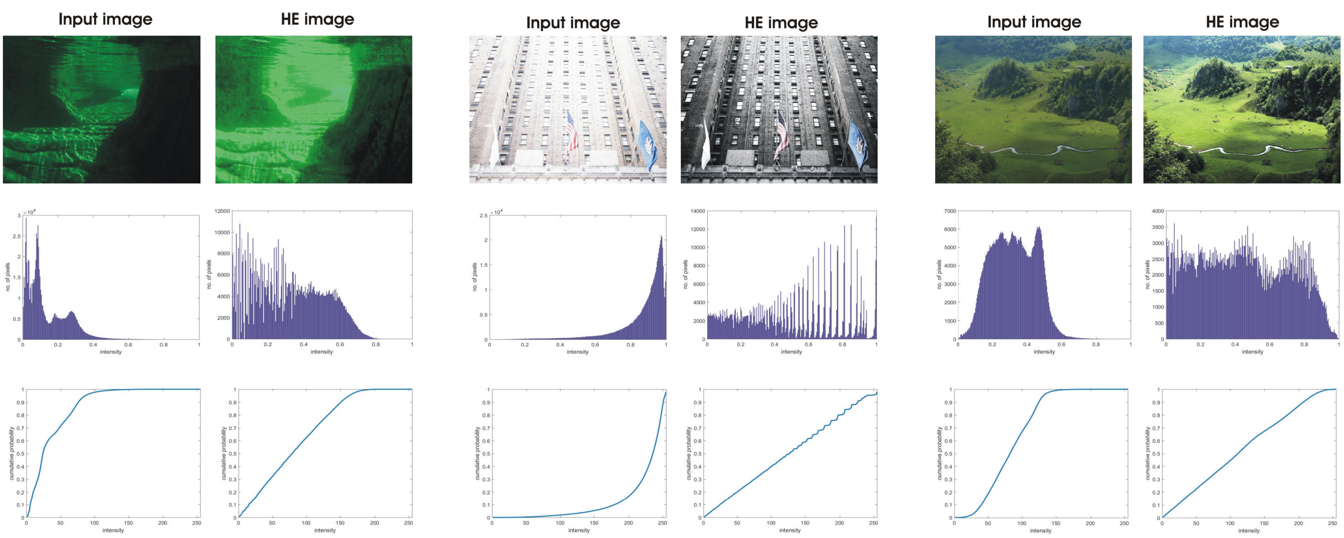


FIGURE 9. In contrast to the examples shown in Figure 8, in this figure the HE (top row) is effective. The middle row shows the corresponding histograms of the input images and the images processed by the HE, and the bottom row the CDFs. Observe that the CDFs do not cross each other, histograms of original images and of corresponding equalized images are ordered.

analyzing the histogram and the CDF in the second and third row, respectively. Similarly, applying HE the yielded result presents more local details and in general the contrast of the resulted image is better enhanced. Observe that the histogram of each of the three images is ordered respectively to the histogram of the corresponding equalized image.

An explanation about why histogram order is important for histogram equalization would be the following. First, remember that if histograms are ordered, according to Corollary 2 all weighted means will keep the order. Consider now the particular case of weighted power means. Each mean represents mostly a rank of the intensities. The higher the exponent of the power mean, the more the higher values of intensity are over-represented on average, while for negative exponents the bigger the absolute value of the exponent the more the smaller values are over-represented on average. Consider that the contrast is perceived as the differences between the averages for the different ranks of intensities. If the histograms are not ordered it would mean that contrast for the equalized intensity would increase for some ranks but would decrease for another ones, while for ordered histograms the contrast would always increase (or decrease).

In Fig. 10 we present more examples with images from [20]. Observe that in general the results are good, and in general the histograms are ordered, this is, the original image CDF keeps over the equalized image CDF. A notable

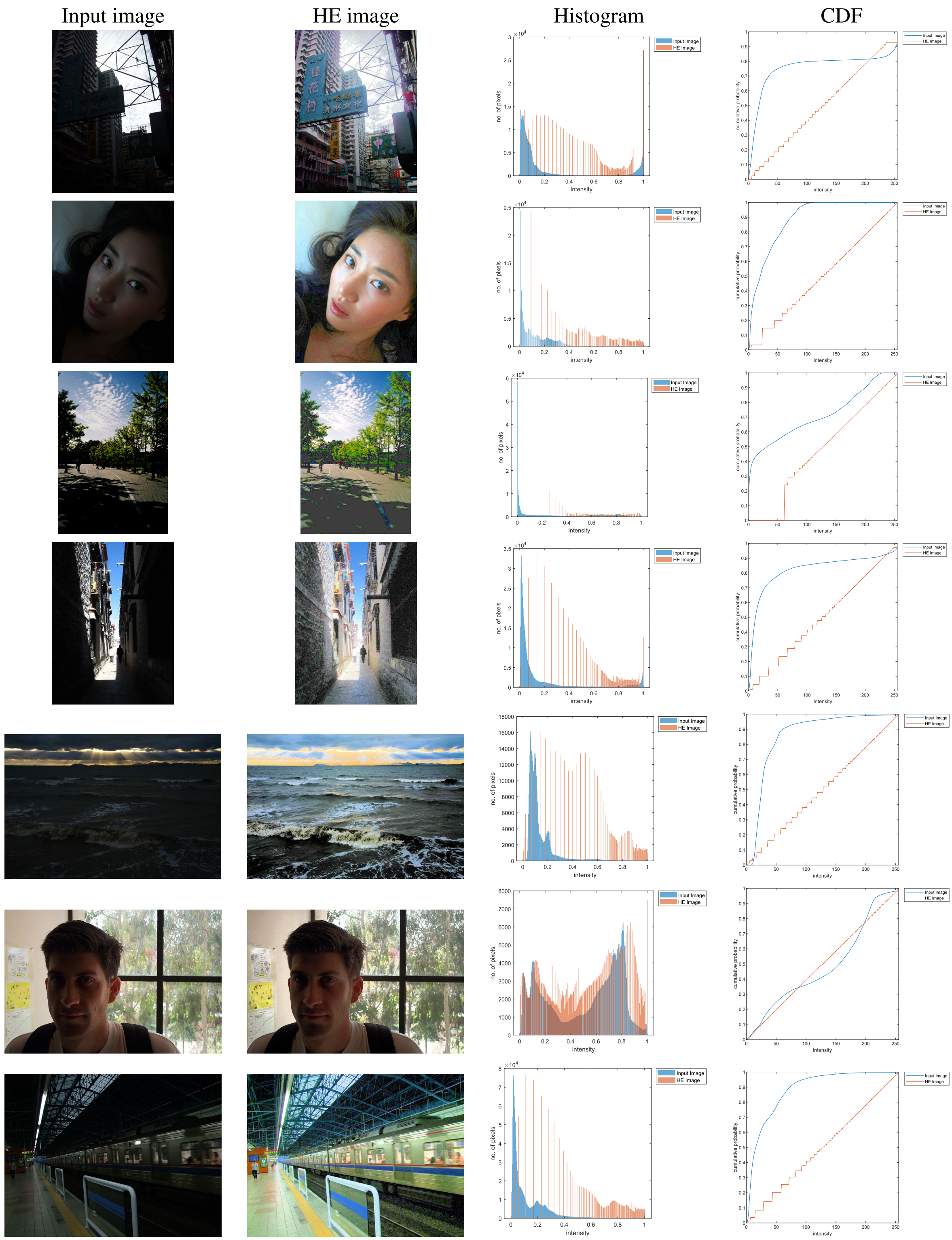


FIGURE 10. Examples of HE. From top to bottom, advertisement, girl, street01, street07, dark_ocean, guy_5 and subway images. The original images are in the first column, in the second the equalized image, in the third the intensity histograms of the intensity of the original image (in blue) and of the equalized image (in red), and in the fourth column the corresponding CDFs.

exception is for *guy5* image, where looking at the CDF's we see that the original CDF of the intensity histogram crosses several times the one of the equalized image. Observe also the sky regions and the tallest part of the white building of the *advertisement* equalized image, which appear a bit noisy and which correspond to the higher intensities, by looking at the corresponding CDFs we see that they cross at the higher intensities.

C. INCOME DISTRIBUTION

Measuring inequality in income distribution has been usually done by comparing frequency histograms, cumulative distribution functions, or Lorentz curves [15]. As Lorentz curves usually are very near one another (see Sec. IV-C1), scalar measures are also used, but they are sometimes contradictory each other [15]. We show in this Section how applying the ordering concept to income distribution frequency histograms can be useful to evaluate the income distribution inequality.

In Figure 11 we see an example of a three year series (2010-11, 2011-12, 2012-13) of income per capita per week. Data corresponds to the distribution of total weekly income in Northern Ireland by household income range and have been taken from [21], very slightly modified to add to 1, which did not in the original data due to rounding. Figure 11 top shows the three year histograms, while Figure 11 bottom shows the three CDFs. From Figure 11 bottom we can see that series 3 dominates series 1 and 2, but neither series 1 or 2 dominates the other one. In terms of weighted means, all means constructed from series 3 using any representative value in each bin will be greater than the corresponding means for series 1 and 2, while some means for series 1 can be less than for series 2 or vice versa. As an example, taking as representative the middle value of each bin and the value 2000 for the last bin, the arithmetic means are {655, 50; 651, 50; 692, 50}, the harmonic means are {339.03; 338.94; 353, 09}, the geometric means are {474.51; 472.48; 500, 11}, and the quadratic means are {866.59; 863.00; 910, 56} for series 1,2, and 3, respectively. Observe that harmonic, geometric, arithmetic and quadratic means are particular cases of the power mean function, with parameter r equal to -1, 0, 1, 2, respectively. The power mean is increasing on parameter r, and the mean is biased towards higher or lower values of the series according to the values of r. The obtained means for series 3 are always greater than for series 1 and 2. This is in concordance with Figure 11 bottom, where the CDF of series 3 is below the CDFS of series 1 and 2. Also, the four means for series 1 are greater than the corresponding ones for series 2. However, from Figure 11 bottom, we see that the CDFs for series 1 and 2 cross, and thus it is not guaranteed that all possible means can be ordered in the same way. In fact, if we compute the power mean for $r = -2$, we obtain {236.74; 236.99; 241.76}, and the mean for series 2 is slightly greater than the one for series 1.

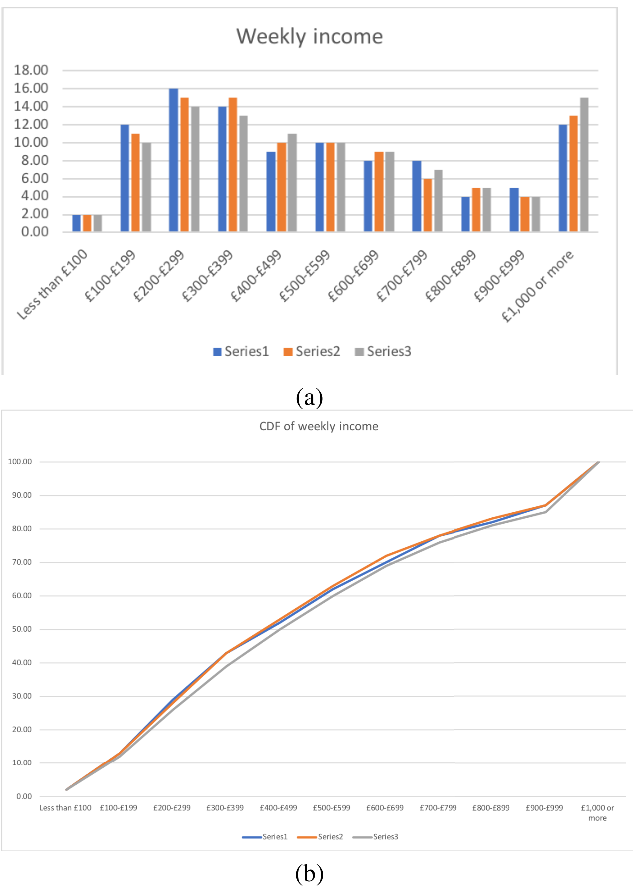


FIGURE 11. Histograms of income per capita per week (a) and cumulative distribution frequencies (b) for three year series.

The decrease from series 1 to series 2 means that all segment rents, except the lowest income ones, fared worse, which can be seen also directly from Figure 11 top, while the preceding order of series 3 with 1 and 2 means that all segments of income fared better in series 3.

1) FIRST STOCHASTIC ORDER AND LORENTZ ORDER

Observe that the first stochastic order is different than Lorentz order [12], [17], used to show inequalities in wealth distribution. Lorentz curve is always convex, with x and y values ranging from 0 to 1, and representing that the poorest $100\%x$ of population possess the $100\%y$ of wealth. Thus the ideal welfare distribution (and limiting curve) is the line $y = x$. In Fig. 12 we show the Lorentz curves for the three series in Figure 11 top. To compute Lorentz curve we compute the accumulated values and accumulated respective distribution of population, and represent accumulated wealth on y axis and accumulated population in x values. The nearest to the line $y = x$ the better distributed the wealth. Observe that the three interpolated lines corresponding to the three series cross each other, and it is very difficult to analyze the change in distribution of wealth. Visually, it looks like series 1 is a bit over the other two, which would mean wealth was a bit better distributed in series 1 than on the other two. This result is not contradictory with series 3 preceding in histogram order to series 1 and 2, which means that all segments of income in series 3 fared better than in series 1 and 2. Because if higher income segments fare better than middle or lower income ones, it could mean a worse wealth distribution. In fact, looking at the difference in the means between series 3 and series 1 and 2, we observe in series 3 a relative higher increase

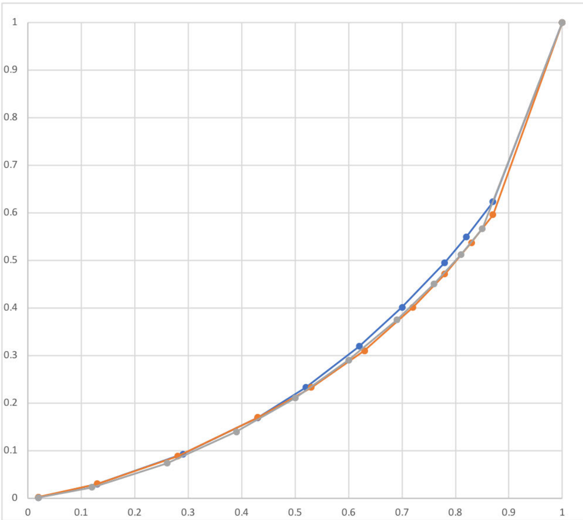


FIGURE 12. Lorentz curves of income distribution for series 1 to 3, shown in Figure 11(a).

in mean for $r = 1, 2$, that correspond to higher incomes, than for $r = -1, -2$, which correspond to lower incomes.

D. GINI INDEX

The Gini index G [16] was introduced to be able to compare wealth distributions given by the Lorentz curve. It is equal to the area between the diagonal $y = x$ and the Lorentz curve divided by the total area below the diagonal. As the axes scale from 0 to 1, it can be computed as two times the area contained between the diagonal and the Lorentz curve. The most equal society will be one in which every person receives the same income (G = 0), i.e. when the Lorentz curve is the diagonal line; the most unequal society will be one in which a single person receives 100% of the total income and $G \approx 1$, i.e. when the Lorentz curve degenerates into a step curve. Thus the lower the index, the better the wealth distribution. If we compute the Gini index, for the three year series Lorentz curves (see Fig. 12), we will obtain the values 0.284, 0.321, 0.378, this is, wealth distribution worsened from year to year, and we could conclude that welfare also worsened. However, we have seen in Fig. 11 that the cdf of the third year is below the cdf's of the first two years, that is, the histogram of third year precedes both first and second year, which means that all averages increased and all segments of income improved. Observe that as discussed in Section IV-C1 there is no contradiction here, both results complement each other. All segments of income improved, although the wealth was less equally distributed.

V. CONCLUSION AND FUTURE WORK

We have introduced in this paper a partial order relationship between frequency histograms. To our knowledge, this is the first work where an order between histograms is considered. We have defined formally an histogram as consisting of a set of intervals and the frequencies associated with each interval. Then we consider that two histograms, defined on the same set of intervals, are ordered if the corresponding frequencies are ordered under the stochastic order. We have presented many properties of the histogram order, maybe the more important being that if two histograms are ordered, all the quasi-arithmetic means taken over representative values of the histograms bins and weighted by the corresponding frequencies follow the same order. Two ordered histograms can be easily visualized because the corresponding CDFs do not cross each other.

We have presented three examples, in traffic flow, histogram equalization, and income distribution. In traffic flow, the order is able to explain the discrepancies in the time speed and space speed when the distribution of speeds changes. In histogram equalization, we are able to predict when equalization will work well, or why equalization has not worked well. In income distribution we have seen that the order implies a general improvement (or a general worsening) for all levels of income, and can give a more clear, albeit complementary, information to Lorentz ordering and Gini index.

In future work we will consider other possible orders, based mainly on the existing work on the stochastic orders that can be established between distributions. The order considered in this paper is based on first stochastic order, that allows for the nice order invariance property for the means, but it is restricted to distributions with CDFs not crossing each other. More relaxed (or more strict) conditions could be considered, or even conditions for the representative values can be imposed. We believe these additional orders may introduce interesting properties and explain further characteristics when comparing histograms. We will consider extending the order to 2d histograms, based on 2d stochastic orders [22], [23], as two dimensional frequency histograms are used for image registration [24]. Finally, we will evaluate whether the defined order can be extended to category histograms, this is, histograms where variable X takes symbolic, discrete values, and which properties are kept.

APPENDIX

Proof of Theorem 1:

Proof: Let us first consider the case where $h(x)$ is monotonous increasing. Then, condition b)

$$h^{-1}\Big(\sum_{k=1}^{n}\alpha'_k h(b_k)\Big) \le h^{-1}\Big(\sum_{k=1}^{n}\alpha_k h(b_k)\Big)$$

is equivalent to

$$\sum_{k=1}^{n}\alpha'_k h(b_k) \le \sum_{k=1}^{n}\alpha_k h(b_k).$$

To prove $b) \Rightarrow a)$ we proceed in the following way. Consider the increasing sequence $\{h(b_1), \ldots, h(b_1), h(b_n), \ldots, h(b_n)\}$, with $b_1 < b_n$ (and thus $h(b_1) < h(b_n)$ by the strict monotonicity of $h(x)$), and where $h(b_1)$ is written l times, denote $\mathbf{L} = a_1 + \ldots + a_l$, $\mathbf{L}' = a'_1 + \ldots + a'_l$. Since $a_{l+1} + \ldots + a_n = 1 - \mathbf{L}$,

$a'_{l+1} + \ldots + a'_n = 1 - \mathbf{L}'$ then applying condition *b*) gives

$$\mathbf{L}'h(b_1) + (1 - \mathbf{L}')h(b_n) - \mathbf{L}h(b_1) - (1 - \mathbf{L})h(b_n) \leq 0,$$

i.e.,

$$(\mathbf{L}' - \mathbf{L})(h(b_1) - h(b_n)) \leq 0 \Rightarrow \mathbf{L}' \geq \mathbf{L},$$

and varying l from 1 to $n-1$ we obtain the first $n-1$ inequalities in *a*), and thus *b*) $\Rightarrow$ *a*).

Let us see now that *a*) implies *b*). Define for $1 \leq k \leq n$, $A_k = \sum_{j=1}^{k} \alpha_k$, $A'_k = \sum_{j=1}^{k} \alpha'_k$, and $A_0 = A'_0 = 0$. Then

$$\begin{aligned}
\sum_{k=1}^{n} \alpha_k h(b_k) - \sum_{k=1}^{n} \alpha'_k h(b_k) &= \sum_{k=1}^{n} (\alpha_k - \alpha'_k) h(b_k) \\
&= \sum_{k=1}^{n} (A_k - A_{k-1} - A'_k + A'_{k-1}) h(b_k) \\
&= \sum_{k=1}^{n} (A_k - A'_k) h(b_k) - \sum_{k=1}^{n} (A_{k-1} - A'_{k-1}) h(b_k) \\
&= \sum_{k=1}^{n-1} (A_k - A'_k) h(b_k) - \sum_{k=0}^{n-1} (A_k - A'_k) h(b_{k+1}) \\
&= \sum_{k=1}^{n-1} (A_k - A'_k) h(b_k) - \sum_{k=1}^{n-1} (A_k - A'_k) h(b_{k+1}) \\
&= \sum_{k=1}^{n-1} (A_k - A'_k)(h(b_k) - h(b_{k+1})) \geq 0,
\end{aligned}$$

as *a*) implies that for all k, $A'_k - A_k \geq 0$, and $\{h(b_k)\}$ is an increasing sequence. Thus *a*) $\Rightarrow$ *b*).

Consider now the case where $h(x)$ is monotonous decreasing. Then, as $h^{-1}(x)$ is decreasing too, condition b)

$$h^{-1}\Big(\sum_{k=1}^{n} \alpha'_k h(b_k)\Big) \leq h^{-1}\Big(\sum_{k=1}^{n} \alpha_k h(b_k)\Big)$$

is equivalent to

$$\sum_{k=1}^{n} \alpha_k h(b_k) \leq \sum_{k=1}^{n} \alpha'_k h(b_k),$$

which is equivalent to

$$\sum_{k=1}^{n} \alpha'_k (-h(b_k)) \leq \sum_{k=1}^{n} \alpha_k (-h(b_k)).$$

But as $h(x)$ is decreasing, $-h(x)$ is increasing, and we have the same case than above. $\square$

Proof of Corollary 1:

Proof: The first inequality in condition a) in Theorem 1 gives us $\alpha'_1 \geq \alpha_1$. Subtracting 1 from the last inequality $\alpha'_1 + \ldots + \alpha'_{n-1} \geq \alpha_1 + \ldots + \alpha_{n-1}$, reversing it and taking into account that $\sum_{i=1}^{n} \alpha'_i = \sum_{i=1}^{n} \alpha_i = 1$ we obtain $\alpha'_n \leq \alpha_n$. $\square$

REFERENCES

[1] K. Pearson, "Contributions to the mathematical theory of evolution. II. Skew variation in homogeneous material," *Philos. Trans. Roy. Soc. London A, Math. Phys. Sci.*, vol. 186, pp. 343–414, Jan. 1895.

[2] P. Bullen, *Handbook of Means and Their Inequalities*. Dordrecht, The Netherlands: Springer, 2003.

[3] Wikipedia contributors. *Quasi-Arithmetic Mean*. Accessed: Oct. 28, 2020. [Online]. Available: https://en.wikipedia.org/wiki/Quasi-arithmetic_mean

[4] A. Rényi, "On measures of entropy and information," in *Proc. 4th Berkeley Symp. Math. Statist. Probab.*, vol. 1. Berkeley (CA), USA: Univ. of California Press, 1961, pp. 547–561.

[5] F. Coggeshall, "The arithmetic, geometric, and harmonic means," *Quart. J. Econ.*, vol. 1, no. 1, pp. 83–86, 1886. [Online]. Available: http://www.jstor.org/stable/1883111

[6] J. Fernández-Villaverde, "The econometrics of DSGE models," NBER, Cambridge, MA, USA, Tech. Rep. 14677, 2009. [Online]. Available: http://www.nber.org/papers/w14677

[7] Wikibooks Contributors. (2017).*Fundamentals of Transportation/Traffic Flow*. Accessed: Jan. 18, 2018. [Online]. Available: https://en.wikibooks.org/wiki/Fundamentals_of_Transportation

[8] T. Padois, O. Doutres, F. Sgard, and A. Berry, "On the use of geometric and harmonic means with the generalized cross-correlation in the time domain to improve noise source maps," *J. Acoust. Soc. Amer.*, vol. 140, no. 1, p. EL56, 2016, doi: 10.1121/1.4955007.

[9] M. Sbert, V. Havran, L. Szirmay-Kalos, and V. Elvira, "Multiple importance sampling characterization by weighted mean invariance," *Vis. Comput.*, vol. 34, nos. 6–8, pp. 843–852, May 2018, doi: 10.1007/s00371-018-1522-x.

[10] N. D. Singpurwalla and B. Lai, "What does the 'mean' really mean?" 2020, *arXiv:2003.01973*. [Online]. Available: https://arxiv.org/abs/2003.01973

[11] M. Sbert and J. Poch, "A necessary and sufficient condition for the inequality of generalized weighted means," *J. Inequalities Appl.*, vol. 2016, no. 1, p. 292, Dec. 2016.

[12] M. Shaked and G. Shanthikumar, *Stochastic Orders*. New York, NY, USA: Springer-Verlag, 2007.

[13] J. G. Wardrop, "Road paper. Some theoretical aspects of road traffic research," *Proc. Inst. Civil Eng.*, vol. 1, no. 3, pp. 325–362, May 1952, doi: 10.1680/ipeds.1952.11259.

[14] A. C. Bovik, *The Essential Guide to Image Processing*. New York, NY, USA: Academic, 2009.

[15] A. Atkinson and F. Bourguignon, "Income distribution," in *International Encyclopedia of the Social & Behavioral Sciences*. Elsevier, 2001, pp. 7265–7271, doi: 10.1016/B0-08-043076-7/02236-1.

[16] Wikipedia Contributors. *Gini Coefficient*. Accessed: Oct. 28, 2020. [Online]. Available: https://en.wikipedia.org/wiki/Gini_coefficient

[17] F. Belzunce, C. Martinez-Riquelme, and J. Mulero, *An Introduction to Stochastic Orders*. New York, NY, USA: Academic, 2016.

[18] T. Arici, S. Dikbas, and Y. Altunbasak, "A histogram modification framework and its application for image contrast enhancement," *IEEE Trans. Image Process.*, vol. 18, no. 9, pp. 1921–1935, Sep. 2009.

[19] A. Toet and T. Wu, "Efficient contrast enhancement through log-power histogram modification," *J. Electron. Imag.*, vol. 23, no. 6, Dec. 2014, Art. no. 063017.

[20] Image Processing Lab, Kyung Hee University, South Korea. *Channel Division Code*. Accessed: Nov. 1, 2020. [Online]. Available: http://vision.khu.ac.kr/?mid=Downloads

[21] NER Institute. (2019). *Quarterly Economic Facts Winter 2014*. [Online]. Available: https://www.nerinstitute.net/sites/default/files/research/2019/qef_winter14_complete_reducedsize.pdf

[22] Y. Yoshida, "Weighted quasi-arithmetic means on two-dimensional regions: An independent case," in *Modeling Decisions for Artificial Intelligence—MDAI 2016* (Lecture Notes in Artificial Intelligence), vol. 9880, V. Torra and Y. Narukawa, Eds. Cham, Switzerland: Springer, 2016, pp. 82–93.

[23] M. Sbert and Y. Yoshida, "Stochastics orders on two-dimensional space: Application to cross entropy," in *Modeling Decisions for Artificial Intelligence—MDAI 2020* (Lecture Notes in Artificial Intelligence), vol. 12256, V. Torra, Y. Narukawa, J. Nin, and N. Agell, Eds. Cham, Switzerland: Springer, 2020.

[24] F. Maes, A. Collignon, D. Vandermeulen, G. Marchal, and P. Suetens, "Multimodality image registration by maximization of mutual information," *IEEE Trans. Med. Imag.*, vol. 16, no. 2, pp. 187–198, Apr. 1997.

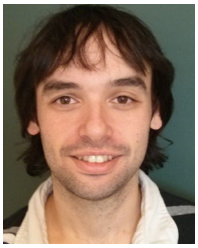
MATEU SBERT (Member, IEEE) received the M.Sc. degree in theoretical physics from the University of Valencia, in 1977, the M.Sc. degree in mathematics (statistics and operations research) from UNED University, Madrid, in 1983, and the Ph.D. degree in computer science from the Universitat Politècnica de Catalunya (UPC), in 1997. He is currently a Full Professor of Computer Science with the University of Girona, Spain. He has authored or coauthored more than 200 articles and four books in his areas of research, and participated in several Eurographics, SIGGRAPH Asia, and VisWeek tutorials. His research interests include the application of Monte Carlo, integral geometry, and information theory techniques to radiosity, global illumination, visualization, and image processing. He was a recipient of the Best Ph.D. Award.

COSMIN ANCUTI received the Ph.D. degree from Hasselt University, Belgium, in 2009. He was a Postdoctoral Fellow with the IMINDS and Intel Exascience Laboratory (IMEC), Leuven, Belgium, from 2010 to 2012, and a Research Fellow with the University Catholique of Louvain, Belgium, from 2015 to 2017. He is currently a Senior Researcher and a Lecturer with Politehnica University Timisoara, Romania, and the Institute of Informatics and Applications, University of Girona, Spain. He is the author of more than 50 papers published in international conference proceedings and journals. His research interests include image and video enhancement techniques, computational photography, and low-level computer vision. He is a Co-Organizer of the IEEE CVPR NTIRE Workshop.

CODRUTA O. ANCUTI (Member, IEEE) received the Ph.D. degree from Hasselt University, Belgium, in 2011. From 2015 to 2017, she was a Research Fellow with the ViCOROB Group, University of Girona, Spain. She is currently an Associate Professor with the Faculty of Electrical and Telecommunication Engineering, Politehnica University Timisoara. Her main research interests include image understanding and visual perception. She was the first researcher to introduce several single image-based enhancing techniques built on the multiscale fusion, such as color-to-grayscale, image dehazing, underwater image, and video restoration. Her work received the Best Paper Award at CVPR NTIRE 2017 Workshop.

JORDI POCH is currently an Associate Professor with the Department of Computer Science, Applied Mathematics, and Statistics, University of Girona. He has been the principal investigator of several projects of e-learning. His main research interests include mathematical modeling, and the development and application of new tools to support teaching.

SHUNING CHEN received the Ph.D. degree from the Natural Resource Economics Division, Kyoto University, Japan, in 2018. She is currently an Assistant Professor with the Urban Institute and the Department of Civil Engineering, School of Engineering, Kyushu University, Japan. Her main research interest includes mathematical modeling and analyzing the environment and sustainable development.



MARIUS VILA received the Computer Engineering degree, the M.Sc. degree in industrial and automated computer engineering, and the Ph.D. degree from the Universitat de Girona, in 2008, 2009, and 2015, respectively. He is currently a Researcher with the Department of Computer Science, Applied Mathematics, and Statistics, Universitat de Girona, and with the Graphics and Imaging Laboratory. He has worked on document classification and video processing. His main research interests include image processing and information theory.