ISSN: 2281-1346

UNIVERSITÀ DI PAVIA
Department of Economics and Management

DEM Working Paper Series

Aggregating ESG scores: a Wasserstein distance-based method

Arianna Agosto
(Università degli Studi di Pavia)

Antonio Balzanella
(Università della Campania "Luigi Vanvitelli")

Paola Cerchiello
(Università degli Studi di Pavia)

# 228 (05-25)

Via San Felice, 5
I-27100 Pavia

https://economiaemanagement.dip.unipv.it/it

Aggregating ESG scores: a Wasserstein distance-based method

A. Agosto, A. Balzanella, P. Cerchiello

May 23, 2025

Abstract

The evaluation of the Environmental, Social and Governance (ESG) profile of companies is gaining more and more importance in the credit and financial system and is made more challenging by the availability of alternative - and often divergent - ESG ratings. In addition, the contribution of the three dimensions (E, S and G) to the final evaluation is not disclosed by the raters. This paper proposes an approach for aggregating the three dimensions constituting ESG ratings by means of optimal transport from the perspective of the Wasserstein distance. An empirical exercise, conducted on a dataset related to Small and Medium Enterprises (SMEs), shows that the proposed aggregated indicator represents a statistically sound and explainable tool for the users of ESG ratings, especially when non-homogenous evaluations are provided. Our proposal is also compared to Principal Component Analysis (PCA), a state of the art machine learning algorithm widely employed in the literature concerning the building of synthetic indicators.

Keywords: Sustainability risk, ESG, SMEs, Summary indicators, Wasserstein distance

1 Introduction

The environmental impact of corporate activities has become crucial over the last decades. The sustainability of companies concerns how corporate activities affect external stakeholders and is often proxied by Environmental, Social and Governance (ESG) factors (Pollman, 2022).

While Environment factors relate to the impact on environment deriving from the production of goods or services, such as carbon emissions, Social factors refer to how the company affects society, including issues such as employee satisfaction, diversity, inequality, gender gap. Lastly, Governance factors are accounted for to evaluate the "good" governance of companies. In the financial context, European regulators request financial intermediaries to include ESG aspects for lending and investment activities (EBA, 2020; ESMA, 2020a,b) and call for the inclusion of ESG aspects into credit worthiness evaluations provided by credit rating agencies (European Action Plan for Financing Sustainable Growth (European Commission, 2018), to direct funds to the best-performing companies in terms of environmental and social impact.

To inform the investors concerning corporate ESG performance, specialised companies (including rating agencies) provide measures and proxies for ESG behaviour, publishing ESG ratings or scores that should express the level of sustainability and the degree of accountability of companies on ESG aspects (Scalet and Kelly, 2010; Avetisyan and Ferrary, 2013).

As each rating provider collects information from different sources (company reports, news, stock exchange information, etc.) and applies proprietary methodologies to combine available inputs and produce a synthetic measure of ESG behaviour, ESG evaluations assigned by different providers often produce divergent results (Berg et al., 2022; Dorfleitner et al., 2015; Abhayawansa and Tyagi, 2021; Dimson et al., 2020; Billio et al., 2021).

Further differences among ratings arise when considering the Environmental, Social and Governance dimensions separately. Such discrepancies are connected with the nature and measurement of the three factors. Indeed, while the Environmental impact can be relatively easy to measure, the Social and Governance impacts are related to qualitative aspects, which are more difficult to assess (Muñoz-Torres et al., 2019).

The aggregation of ESG metrics is of primary importance for both investors, who can use improved ESG evaluations to choose sustainable investment opportunities, and for evaluated companies, who can take appropriate countermeasures to improve their environmental, social and governance profile.

Several recent works used data-driven techniques to build aggregate ESG measures from a number of indicators. Agosto et al. (2023a) and Agosto et al. (2023b) relied on a Bayesian approach - based on the methodological results of Giudici et al. (2003) and Cerchiello and Giudici (2014) - to obtain an aggregated indicator for the ESG performance of companies by integrating ratings assigned by different providers. In both works the aggregated ESG indicator is computed starting from single ESG scores, by attributing a weight to each. The proposed weighting procedure is data-driven, as it relies on the relationship between the ESG performance and the creditworthiness measured by credit ratings issued by recognized agencies.

A recent paper by Gucciardi et al. (2024) employs Principal Component Analysis to evaluate the ”common factors” driving environmental scores.

The present work contributes to the literature concerning the aggregation of ESG metrics by developing a method for obtaining an ESG compound indicator using optimal transport from the perspective of the Wasserstein distance, as proposed by Agosto et al. (2025) for generic statistical indicators. Such a consensus indicator represents a statistically sound synthesis when the raters express non-homogeneous ratings, and it can be applied to ratings having different distribution and support. In addition, the possibility of computing confidence intervals for the aggregated indicator allows the rating user to choose between a more prudential evaluation - provided by the lower bound - and a more tolerant one - given by the upper bound - based on the aim of the analysis and on the

level of environmental risk aversion.
The work is structured as follows. Section 2 summarises the proposed methodology; Section 3 describes the dataset used in our empirical study; Section 4 presents the results obtained by applying the proposed aggregation methodology to real data; Section 6 concludes.

2 Methods

2.1 Preliminaries on Wasserstein distance

Wasserstein distances are metrics on probability distributions which measure the minimal effort required to reconfigure the probability mass of a distribution in order to recover another distribution (Villani, 2003).

In general, the *p-Wasserstein distance* between two probability distributions μ and ν on a metric space (M, d) is defined as:

$$W_p(\mu, \nu) = \left(\inf_{\gamma \in \Gamma(\mu, \nu)} \int_{M \times M} x - y^p \, d\gamma(x, y) \right)^{\frac{1}{p}}, \tag{1}$$

where:

- x and y are events on the metric space M
- $d = x - y^p$ is the metric on M.
- $\Gamma(\mu, \nu)$ denotes the set of all couplings of μ and ν.
- $p \geq 1$ is a parameter that determines the type of Wasserstein distance.

Eq. 1 provides an analytic definition of the Monge–Kantorovich optimization problem (Panaretos and Zemel, 2019). Here, the Wasserstein distance $W_p(\mu, \nu)$ quantifies the "cost" of transforming one distribution into another. This cost is computed using the metric $\|x - y\|^p$ on M. The couplings $\Gamma(\mu, \nu)$ play a critical role in this computation, as they represent all possible ways to "move mass" from the distribution μ to match the distribution ν.

In other words, each coupling γ in $\Gamma(\mu, \nu)$ describes a way of redistributing the mass of μ so that it matches ν. The p-Wasserstein distance is then the minimum "cost" over all such couplings.

The Wasserstein distances W_p are proper distances in that they are nonnegative, symmetric, and satisfy the triangle inequality (Santambrogio, 2015).

In optimal transport theory, the goal is to find the optimal coupling γ^* that minimizes the cost function:

$$\int_{M \times M} x - y^p \, d\gamma(x, y). \tag{2}$$

This optimal coupling γ^* represents the most efficient way to reallocate the mass from μ to match ν under the given cost function defined by the metric $x - y^p$.

The computation of the Wasserstein distance in Eq.1 has, in general, no closed form, but is an infinite-dimensional linear program over a space of measures; however, there are some specific cases for which a closed form exists.

Among these, there is the probability distribution on the real line $\mathbb{R}$ with the metric $d(x, y) = |x - y|$.

As shown in Panaretos and Zemel (2019), given two cumulative distribution functions (CDFs) F_μ and F_ν corresponding to μ and ν, respectively, the p-Wasserstein distance is:

$$W_p(\mu, \nu) = \left(\int_0^1 \left| F_\mu^{-1}(s) - F_\nu^{-1}(s) \right|^p \, ds \right)^{\frac{1}{p}} \quad \text{with } s = 0, \ldots, 1, \tag{3}$$

where F_μ^{-1} and F_ν^{-1} are the quantile functions (inverse CDFs) of μ and ν, respectively. In this case, the Wasserstein distance is isometric to a linear space equipped with the norm L^p so that the Wasserstein distance for the measures in $\mathbb{R}$ is a Hilbertian metric (Panaretos and Zemel (2019)). This makes the geometry of 1-D optimal transport different from its geometry in higher dimensions, which is not Hilbertian.

Formula 3 leverages the fact that on the real line, the optimal coupling can be derived from the quantile functions of the distributions.

As proved in Cuesta-Albertos et al. (1993), in the case of absolutely continuous distributions on the real line, a deterministic optimal map $T : \mathbb{R} \to \mathbb{R}$ that transports μ to ν can be derived directly from the quantile functions. The optimal transport map T is given by:

$$T(x) = F_\nu^{-1}(F_\mu(x)), \tag{4}$$

where F_μ is the CDF of μ and F_ν^{-1} is the inverse CDF (quantile function) of ν.

Such map T ensures that if X is a random variable with distribution μ, then $T(X)$ will have distribution ν. This is optimal in the sense that it minimizes the expected cost $\mathbb{E}[d(X, T(X))^p]$.

The optimal transport map is a key-point in our strategy because it allows us to map the initial ratings on a common probability measure. To develop this aspect, we need to recall the concept of Wasserstein barycenters (Agueh and Carlier, 2011).

Following the concept of Fréchet mean (Fréchet, 1960) as a generalization of centroids to metric spaces, the Wasserstein barycenter can be expressed as the minimizer of the following functional:

$$F(b) = \mathbb{E}W_p^2(\mu_i, b) \tag{5}$$

By means of convex analysis, in Agueh and Carlier (2011) the existence, uniqueness and a characterization of empirical Fréchet means are proved in $W_2(R^d)$. In particular, for probability measures in $\mathbb{R}$, the Fréchet means are unique with the sole restriction that the functional F is finite. In such a case, there is also a closed-form solution to the minimization of the functional F which is still based on quantile functions.

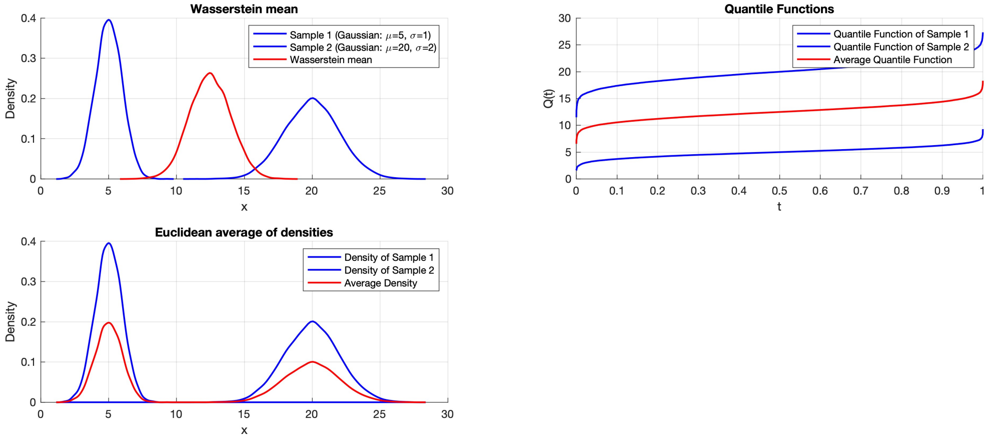

Figure 1: The figure on the top-left shows the Wasserstein barycenter (in red) of two Gaussian distributions. The figure on the right shows the computation of the mean quantile function from the quantile functions of two Gaussian distributions. The bottom-left figure shows the average distribution computed as average of the pdfs

Given a set of probability measures $\{\mu_i\}_{i=1}^n$ on the real line $\mathbb{R}$ with corresponding cumulative distribution functions (CDFs) $\{F_{\mu_i}\}_{i=1}^n$, the *Wasserstein mean* μ_W is defined as the probability measure whose quantile function $F_{\mu_W}^{-1}$ is the barycenter of the quantile functions $\{F_{\mu_i}^{-1}\}_{i=1}^n$.

The quantile function of the Wasserstein mean is given by:

$$F_{\mu_W}^{-1}(s) = \frac{1}{n}\sum_{i=1}^{n} F_{\mu_i}^{-1}(s), \quad s \in [0,1]. \tag{6}$$

The Wasserstein mean, also known as the Wasserstein barycenter, represents the ”average” distribution in the Wasserstein sense, and it has some interesting features, which make it suitable in our context.

The Wasserstein mean for probability measures on the real line provides an easy-to-interpret view of the concept of barycenter distribution. In fact, as shown in Fig. 1, if we consider two Gaussian probability measures with different means and standard deviations, the Wasserstein mean, computed through the quantile functions of the two measures, is a measure with the mean as the average of the means and the standard deviation as the average of the standard deviations. In Fig. 1 we also plot the Euclidean mean of the same Gaussian densities, which appears as a blurred version of the initial distribution.

Wasserstein mean captures a common-sense view of the concept of mean. It is also important to note that it allows processing distributions with different supports. This is essential when ratings are expressed on different scales.

2.2 Proposal

Following Agosto et al. (2025), we consider a data matrix $X = \{X_1, \dots, X_j, \dots, X_p\}$ with $X_j = \{x_{i,j}\}_{i=1,\dots,n}$. Each X_j records the ratings provided by a subject/rater to n units, so that $x_{i,j}$ is the value of the j-th rating for the i-th unit.

Our primary goal is to develop a consensus rating $I = \{I_i\}_{i=1,\dots,n}$ that accounts for the ratings in X provided by the individual subjects.

We assume that X_j is a sample with an unknown continuous distribution and that the empirical measure μ_j associated with the sample is identified by the empirical cumulative distribution function $F_j(t) = n^{-1} \sum_{i=1}^{n} \mathbf{1}(x_{i,j} < t)$ and the corresponding quantile function $Q_j(s) = F_j^{-1}(s)$.

The idea underlying the proposed approach is that there is a hidden rating law for the involved units; however, we only have access to warped/deformed realizations of this law provided by individual rating subjects. Thus, the aim is to recover such a hidden law and provide consensus ratings I for the n units. We assume that the behavior of a rater is fully described by the empirical measure associated with the provided ratings. For instance, a positively skewed distribution could be associated with a conservative rater, while a negatively skewed distribution could indicate more optimistic ratings/raters.

To achieve our aims, we approach the problem using optimal transport from the perspective of the Wasserstein distance (Villani, 2009).

We propose to discover the consensus ratings I by starting from a Fréchet mean measure, obtained by minimizing an appropriate functional of the empirical measures μ_j. The detection of the Fréchet mean of the empirical measures provides a description of the barycentric rater.

The ratings X_j provided by each rater are mapped, using optimal transport maps, onto the mean measure. That is, we transform the initial ratings into new ratings that align with the perspective of the barycentric rater.

The impact of this transformation is two-fold: 1) it provides an easy way to understand how the initial ratings must be warped/aligned in order to recover the consensus ratings; 2) the transformed ratings have a common distribution, which is the barycentric rater's one. This last point allows for consistent comparisons among the ratings of different raters, as every new rating shares the same distribution. In this sense, our proposal performs a standardization from the perspective of the data distribution. This allows us to obtain the consensus rating I for each unit by averaging the initial transformed scores.

By examining how each rating X_j maps to the barycenter, we can also quantify its contribution to the consensus. Areas where multiple ratings map to the same region of the barycenter indicate agreement.

Since the ratings are mapped onto the Wasserstein mean, and since the latter is a functional of the distribution of the initial ratings, we also provide α-confidence bounds for the consensus distribution. Additionally, we offer suitable confidence intervals for the consensus rating I by means of an appropriate resampling strategy.

2.3 Consensus indicator based on the Wasserstein distance

In this section, we develop our proposal for obtaining a consensus indicator starting from the basic ratings recorded in a given X.

As shown in before, to each rating X_j corresponds an empirical cumulative distribution function $F_j(t)$ (with t being the rating domain) and a quantile function $Q_j(s) = F_j^{-1}(s)$ (with $s \in [0, 1]$).

Adapting the Eq. 6, we can compute the Wasserstein barycenter of ratings by:

$$Q_m(s) = \frac{1}{p} \sum_{i=1}^{p} Q_j(s), \quad s \in [0, 1]. \tag{7}$$

Consistent with the definition of the Fréchet mean, $Q_m(s)$ allows for the identification of a distribution that is barycentric in the Wasserstein sense, thus making it the natural choice for identifying the distribution of the consensus indicator.

It is interesting to note that from the definition of the Wasserstein barycenter of the rating distributions it is also possible to define an empirical Wasserstein-based measure of heterogeneity by:

$$\sigma_w^2 = \frac{1}{p} \sum_{j=1}^{p} \int_0^1 (Q_j(s) - Q_m(s))^2 ds. \tag{8}$$

It is a variance-like measure that allows for the assessment of the heterogeneity of different raters and helps to understand how closely they align with the average rater.

Notably, since σ_w^2 is based on measuring the Wasserstein distance between each rater and the Wasserstein barycenter, it provides a measure of the average effort required to reconfigure the probability masses of a distribution in order to recover the barycenter distribution. In other words, σ_w^2 can be interpreted as a measure of the **average** amount of deformation undergone by an individual's rating distribution to be transformed into the mean distribution.

In order to provide the consensus indicator, we initially map the ratings X_j (for each $j = 1, \ldots, n$) to the Wasserstein barycenter by:

$$T_j^*(t) = Q_m(F_j(t)) \tag{9}$$

The map $T_j^*(t)$ adjusts the distribution $F_j(t)$ of the rating X_j to match the barycenter distribution identified through its quantile function Q_m.

As shown in the example in Fig. 2, the support t of the distribution F_j associated with a rater X_j is mapped onto the support of the mean distribution. We can see the map as a tool for converting each rating $x_{i,j}$ provided by a rater X_j into a new rating $x_{i,j}^*$ obtained by $x_{i,j}^* = Q_m(F_j(x_{i,j}))$. In this sense, the map allows to evaluate the warping of the initial ratings of a rater X_j in order to match the mean rater.

By applying the map T_j^* to each X_j, we get a new data table X^* with the same dimensions as X, where the distribution of each X_j^* matches that of the mean distribution.

From X^* it is possible to compute the final consensus indicator $I = \{I_i\}_{i=1}^n$ by:

$$I_i = \frac{1}{n} \sum_{j=1}^{p} x_{i,j}^* \quad \text{for all } i = 1, \ldots, n \tag{10}$$

It is computed as average of the mapped ratings so that it is able to account for individual ratings having different distribution and support. The a-priori mapping of the initial ratings X_j on the mean distribution allows to get consistence in the obtained consensus rating. Similarly to indexes computed starting from classic variable standardization, where basic indicators are made consistent by shifting and stretching their distribution in order to have zero-average and unitary standard deviation, our proposal performs a normalization in a distribution sense. Thus, unlike to classic standardization, it is naturally effective when initial indicators have different distributions and support. It is still interesting to note that Wasserstein optimal mapping supports an easy interpretation of how the consensus index is built. In fact, we can evaluate the contribution of each initial rating to the consensus rating on the basis of the Wasserstein distance by:

$$C_{X_j} = \frac{\int_0^1 (Q_j(s) - Q_m(s))^2 ds}{\sum_{l=1}^{p} \int_0^1 (Q_l(s) - Q_m(s))^2 ds}. \tag{11}$$

which is the ratio between the Wasserstein distance computed on the distribution of the initial rating X_i, the distribution of the consensus rating I barycenter and the Wasserstein based deviance.

Since the consensus indicator I gives us the ratings under the perspective of the average rater, we can use the inverse mapping, from the consensus space to the space of individual raters, in order to transport the consensus ratings I_i to the space of the rater X_j:

$$T_j^{inv}(t) = Q_j(F_m(t)) \tag{12}$$

Here, the optimal transport map $T_j^{inv}(t)$ maps each point of the consensus indicator I_i, through its ECDF F_m, to the quantile function Q_j of the rater X_j, such that the cumulative distribution function (CDF) of I is equal to the CDF F_j of the rater X_j.

This can help in interpreting what a particular rating in the consensus space means in terms of original ratings.

2.4 Confidence bounds for the consensus indicator

The consensus indicator in this paper is constructed through a two-step process: first, the optimal mapping of initial ratings to the Fréchet mean distribution

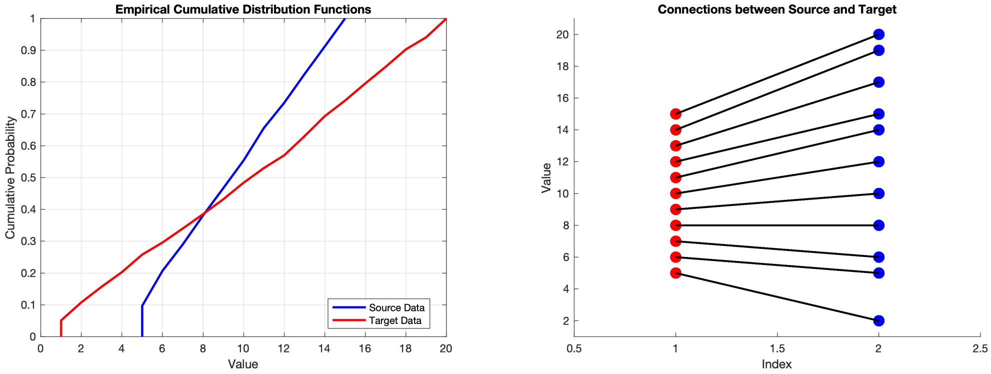

Figure 2: The figure on the left shows the ECDF of the source distribution (randomly generated data from a uniform distribution in $5-15$) and the ECDF of the target distribution (randomly generated data from a uniform distribution in $1-20$). The figure on the right shows the optimal mapping of source points to the target distribution

through the map T_j^* in Eq.9 is applied, followed by mapping the average of the mapped data onto the Wasserstein mean. To assess the reliability and variability of this indicator, we develop appropriate confidence bounds.

The confidence bounds provide a range of plausible values for the true consensus, accounting for uncertainties in the underlying data and in the aggregation process. The bounds are calculated using a bootstrap resampling method, which allows us to estimate the sampling distribution of the consensus indicator without making strong assumptions about the underlying data distribution. As detailed in Agosto et al. (2025), the procedure for computing the confidence bounds is based on a bootstrap algorithm. In particular, we consider B bootstrap samples in which each initial rating X_j is resampled with replacement. For each resampling iteration $b = 1, \ldots, B$, we get a new data matrix $Y = \{Y_1, \ldots, Y_j, \ldots, Y_p\}$ with $Y_j = \{y_{i,j}\}_{i=1,\ldots,n}$ made of resampled data.

Consistently, with the basic procedure for obtaining the consensus rating, we compute the Wasserstein mean of the resampled data through Eq. 7 on the quantile functions Q_{Y_j}. The quantile function of the Wasserstein mean will be denoted by Q_m^*.

The next step is to map the resampled ratings Y_j on the mean distribution by $Y_j^* = Q_m^*(F_{Y_j}(Y_j))$. Finally, we can compute the consensus indicator I^b, for a bootstrap replicate b, by averaging the mapped indicators Y_j^*.

Once B iterations have been performed, we can recover an $\alpha/2$ lower bound and an $1-\alpha/2$ upper bound of the proposed indicator by selecting percentiles from the distribution of I_i^b for each i.

The resulting interval [lower bound, upper bound] can be interpreted as containing the true value of the consensus indicator with $(1-\alpha) \times 100\%$ confidence, given the observed data and the assumptions of the bootstrap method.

The width of the confidence interval provides insight into the precision of the consensus indicator. Narrower intervals suggest greater precision, while wider intervals indicate more uncertainty. Factors influencing the width of the interval include the number and variability of the basic indicators, the sample size, and the strength of agreement among the indicators.

These confidence bounds are particularly useful in our context, as they account for both the uncertainty in the individual basic indicators and the potential variability introduced by the quantile mapping and averaging processes. They provide a more comprehensive view of the reliability of the consensus rating, allowing for more informed interpretation and decision-making based on the aggregated data.

3 Data

The data used in this work have been provided by Modefinance, a company specialized in the assessment of companies' and banks' creditworthiness [1]. Besides credit rating, Modefinance provides ESG rating measures to assess the long-term sustainability and future social and environmental impact of corporate activity.

The dataset provided by Modefinance includes anonymized data for 1559 Italian companies (SMEs). Specifically, data includes the following variables:

- the ESG rating class: an ordinal variable made up of 7 levels (where S1 denotes the best ESG evaluation and S7 indicates the worst one). This metric is the result of the aggregation of the three dimensions related to corporate sustainability: Environmental, Social and Governance;
- the Environmental rating class: an ordinal variable made up of 7 levels (where S1 denotes the best Environmental evaluation and S7 indicates the worst one);
- the Social rating class: an ordinal variable made up of 7 levels (where S1 denotes the best Social evaluation and S7 indicates the worst one);
- the Governance rating class: an ordinal variable made up of 7 levels (where S1 denotes the best Governance evaluation and S7 indicates the worst one).

The ESG ratings are referred to year 2022. Before performing our analyses, we restrict the sample to the companies for which all the three ESG ratings (Environmental, Social, Governance) are available, ending up with a sample of 1245 institutions. Note that, as the ESG ratings are referred to a single financial year, the data are cross-sectional, i.e. each institution enters the sample only once. Figure 3 shows the distribution of companies among the ESG rating classes for the aggregated ESG evaluation and for each of the three dimensions (E, S and G). It can be seen that the Environmental indicator is skewed towards

[1] Modefinance was registered as a Credit Rating Agency for the evaluation of companies in 2015, and was officially listed as ECAI (External Credit Assessment Institution) bu the European Banking Authority in 2018.

	Mean	Median	Standard deviation	Min	Max	Skewness
Environmental	2.24	2.00	1.27	1	7	1.70
Social	3.94	4.00	1.66	1	7	0.09
Governance	3.15	3.00	1.02	1	6	0.35

Table 1: Descriptive statistics for the analyzed E, S and G ratings.

the highest rating classes, while the Social and the Governance one show a more symmetric distribution. The aggregated ESG indicator shows a quite symmetric distribution, with a high concentration in the central classes. It can be seen from Table 1, which reports descriptive statistics for each of the three ratings, that the lower mean - corresponding to an average more optimistic evaluation - is calculated for the E rating, which is also positively skewed, while the others are nearly symmetric. In addition, it can be seen from the standard deviation values that the three indicators are quite similar in terms of variability.

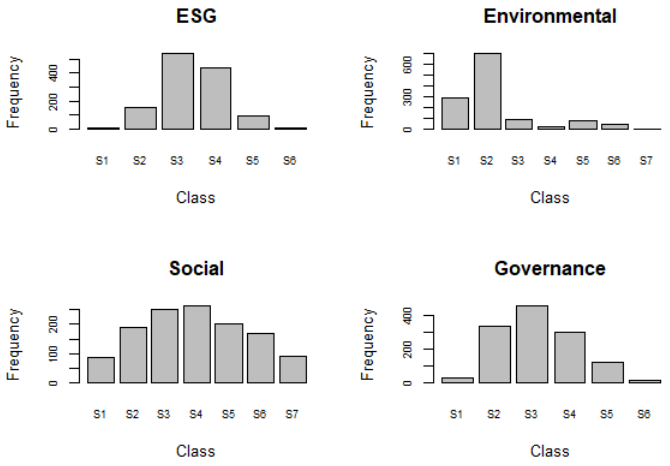

Figure 3: Distribution of sample companies among the ESG rating classes.

A crucial issue when dealing with ESG scores is to assign a weight reflecting the contribution of the E, S and G components to the aggregated ESG evaluations. As Modefinance uses a proprietary algorithm to aggregate the three components, it is not possible to use the provider's aggregation methods as a benchmark. However, some data description can help shedding light on this aspect. In particular, in the analyzed data over 28% of the companies belong to an ESG class that is different from the (rounded) simple mean of the E, S and G parts, suggesting that some expert or data-driven criteria are used to merge the three dimensions. Furthermore, the overall ESG evaluation coincides with the environmental one in only 11% of cases, revealing that the social and governance factors have a decisive impact on the sustainability rating.

	Environmental	Social	Governance
Environmental	1.0000	0.0439 (0.1218)	0.1635 (< 0.001)
Social	0.0439 (0.1218)	1.0000	0.157 (< 0.001)
Governance	0.1635 (0.001)	0.157 (0.001)	1.0000

Table 2: Spearman correlation coefficients between the Environmental, Social and Governance ESG ratings (associated p-values in brackets)

	Environmental	Social	Governance
Environmental	1.0000	0.0368 (0.1150)	0.1409 (< 0.001)
Social	0.0368 (0.1150)	1.0000	0.1271 (< 0.001)
Governance	0.1409 (0.001)	0.1271 (0.001)	1.0000

Table 3: Kendall's Tau correlations between the Environmental, Social and Governance ESG ratings (associated p-values in brackets)

Another relevant aspect to our analysis is related to the strength of the relationships between the environmental, social and governance components. Looking at correlation measures, it can be seen from Tables 2 and 3, which report the Spearman and the Kendall's Tau correlation values (and associated p-values) between the ESG dimensions, the three ratings show very low correlations. In particular, while there is a significant positive correlation – though not high in magnitude - between the Environmental and the Governance evaluations and between the Social and the Governance one, no significant correlation is found between the Environmental and the Social ratings. This is probably connected with the fact that the social factors are less objective, more difficult to measure, and are often retrieved by companies' self-assessment results.

4 Results

This section shows the results obtained by applying the methodology by Agosto et al. (2025) and summarised in Section 2 to the data described in Section 3. In particular, we build a new ESG measure by using and aggregating the evaluations assigned by the provider to each of the three corporate sustainability dimensions (E, S and G).

The first step to aggregate the three (E, S and G) indicators is to compute the empirical cumulative distribution and the quantile function for each rating. The latter is shown in Figure 4. Note that, as we are dealing with discrete variables, the related quantile function is a step one. Confirming what shown in

Section 3, the evaluations of the Social dimension are stricter than the others - as higher classes correspond to worst evaluations - for most quantiles, while the most optimistic evaluation is the one related to the Environmental dimension.

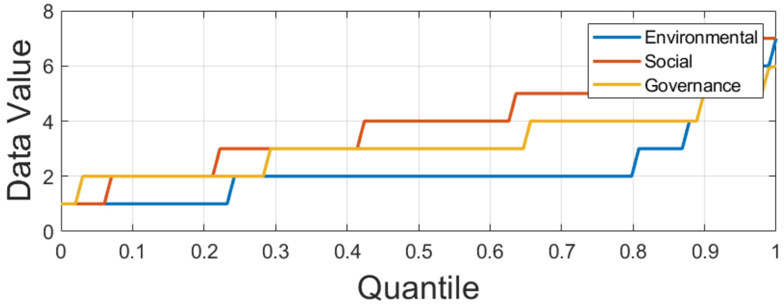

Figure 4: Quantile function for the E, S and G ratings.

After computing the quantile functions, we can calculate the Wasserstein barycenter, which, in this case, represents the average distribution of the three sustainability dimensions. This allows to assign each company three additional scores, representing the transformation of the ratings related to the individual dimensions into new ratings that map the initial ones to the Wasserstein mean of the three distributions. The mapped indicators are shown in Figure 5, which shows the correspondence between each value of the initial indicator (on the x-axis) and the corresponding value on the Wasserstein mean (blue points), together with the lower and upper bounds. Points above the black dashed line indicate that the mapping leads to higher values, i.e. to worst evaluations. This happens for all values of the E and G ratings. On the converse, points below the black dashed line indicate that the values are transformed into lower ones, corresponding to better evaluations. It can be seen from Figure 5 that all values of the E and G ratings are shifted towards worst classes, while an opposite shift, towards better values, is applied to the S rating, except for the left tail of its distribution. The histograms shown in Figure 6 confirm these results: while the E and G ratings are shifted to the right, the S rating distribution is less impacted and is shifted to the left. For all three indicators, the mapped version is less dispersed than the original one.

Finally, the aggregated rating, for which descriptive statistics are reported in Table 4, is computed by averaging the mapped individual indicators. By comparing Table 4 with Table 1, it can be seen that, while an averaging effect is observed for the mean, the value of standard deviation of the aggregated rating is lower than the one of the individual ratings. Furthermore, differently from the original ones, the new indicator is slightly left-skewed.

In this case, the new indicator can be considered as an overall evaluation of the ESG company profile, taking the three sustainability dimensions into account. In this sense, it represents an alternative indicator with respect to

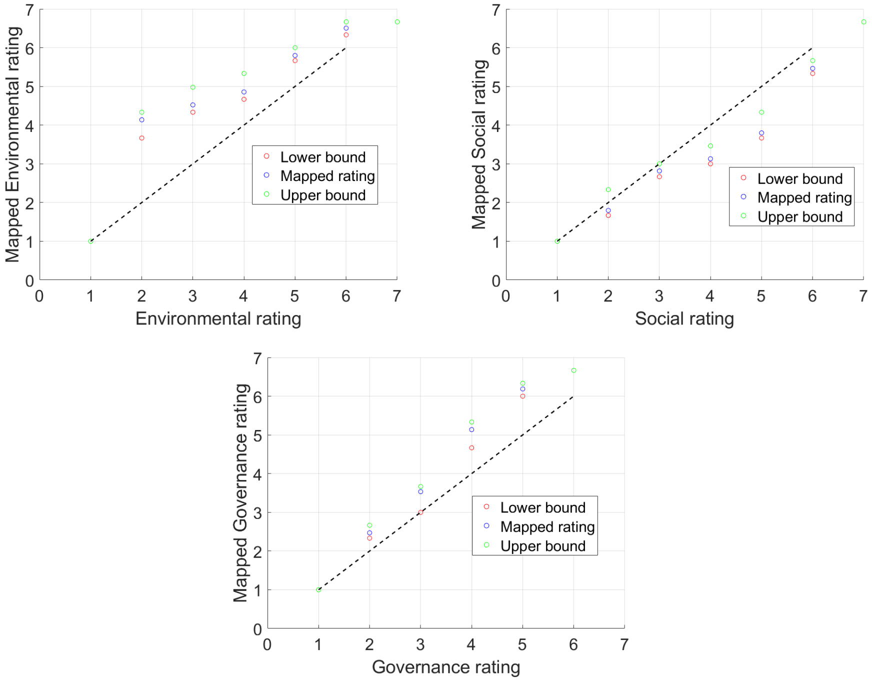

Figure 5: Actual and Wasserstein-mapped indicators for the three ESG dimensions. From top-left to bottom: E, S and G.

	Mean	Median	Standard deviation	Min	Max	Skewness
Aggregated indicator	3.64	3.60	0.93	1.00	6.22	-0.112

Table 4: Descriptive statistics for the aggregated rating calculated based on the E, S and G ratings.

the ESG overall rating provided by agencies, for which the weight of the single dimensions and the aggregation procedure are unknown. Thus, the Wasserstein-based indicator constitutes a statistically robust and explainable additional evaluation.

The quantile function of the aggregated rating and the associated confidence bounds are shown in Figure 7. As in the previous application to ESG scores, lower and upper bounds, further than providing a measure of uncertainty, can be interpreted by the final user as a more optimistic and a more prudential evaluation, respectively. It is worth remarking that, while the initial individual ratings are expressed by ordinal discrete variables, the mean indicator is continuous. The possible attribution of companies to aggregated rating classes based on the new measure depends on the cut-off choice made by the rating user, similar to what is done in credit scoring models.

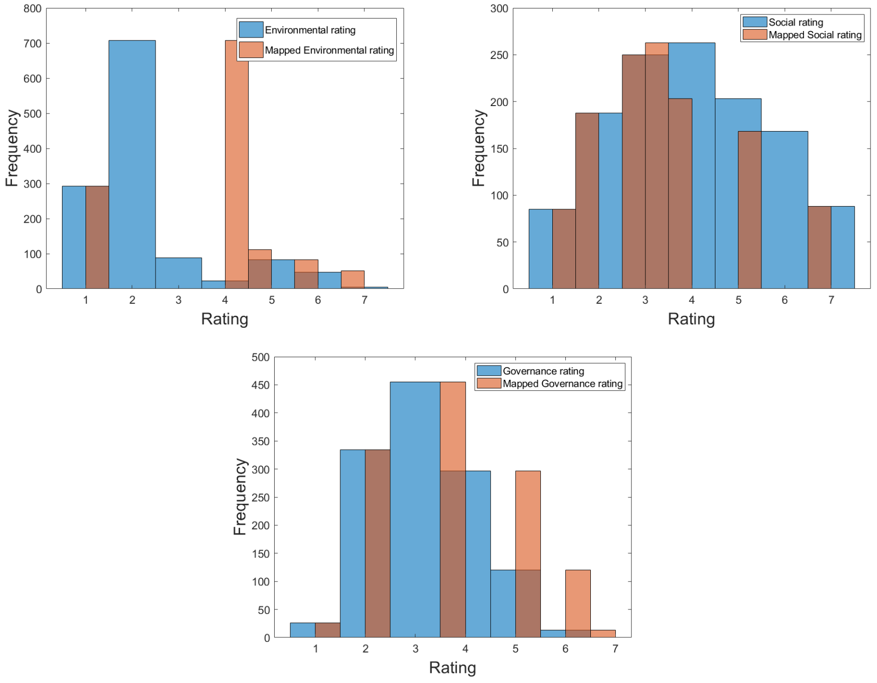

Figure 6: Distribution of actual and Wasserstein-mapped ratings for the three ESG dimensions. From top-left to bottom: E, S and G.

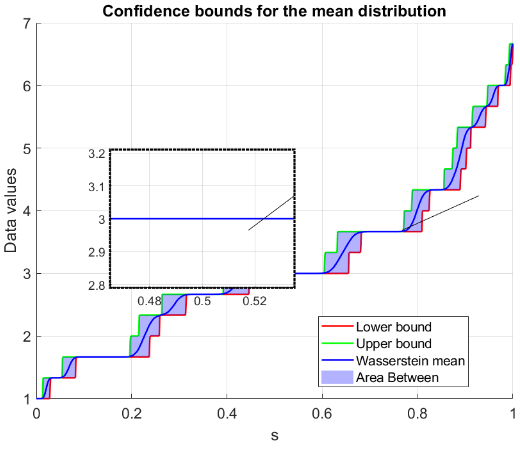

Figure 7: Quantile function of the aggregated ESG rating.

As previously mentioned, the proposed metric represents an additional tool with respect to ESG ratings provided by specialized agencies, such as Modefinance. To analyze the extent to which, in our data, our indicator differs from the ESG rating issued by the provider, Figure 8 shows the scatterplot of the two measures. It can be seen that, while there is a substantial agreement in the tails of the rating distribution, and especially in the left one - corresponding to the best-rated companies, the Wasserstein-based rating provides a more granular evaluation in the central classes.

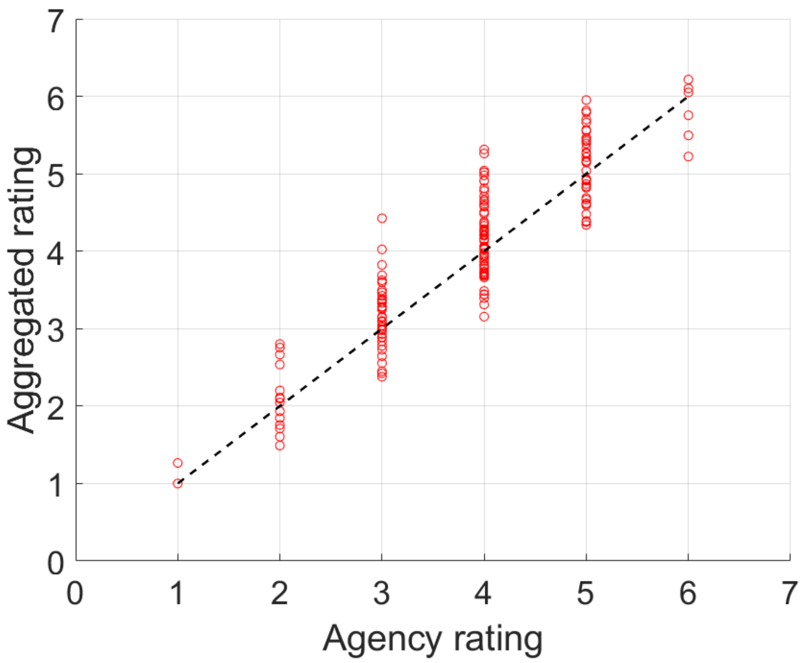

Figure 8: Modefinance ESG rating and Wasserstein-based aggregated ESG rating.

5 ML Models comparison

The approach presented in this paper exploits by design the distributional characteristics of the data at hand. In particular, as described in Section 2, we build the new consensus indicator by leveraging upon the empirical cumulative distribution function and the corresponding quantile functions of each rating which are mapped onto to the Wasserstein barycenter afterward. This, on one hand prevents us from the assumption of a specific distributional form for each rating variable, on the other hand informs us about the amount of wrapping needed to convert the original rating onto the Wasserstein barycenter. For the sake of completeness and comparability, we explore how well a popular machine learning approach adapts to the considered data. More in detail, as mentioned in Section 1, Principal Component Analysis (hereafter PCA) is a particularly well-suited methodology to synthesize data and to reveal the possible presence and characteristics of latent traits. Moreover, PCA has been widely used in the literature to build new aggregated indicators able to produce natural rankings of data units. PCA has also the advantage to be considered an explainable approach, as the role played by each and every input variable is highlighted by a system of weights, referred to as loadings. In this regard, we have fitted PCA to the data described in Section 3, considering the extension to discrete data.

Results are reported in Table 5.

	PC1	PC2	PC3
E	-0.913	−0.037	0.406
S	0.232	0.721	0.653
G	-0.367	0.733	-0.572
Expl Var	37.220	32.460	30.220

Table 5: Loadings resulting from the categorical PCA analysis applied to the E, S and G scores and quota of explained variability per component.

From Table 5 emerges that the three scores tend to distribute quite homogeneously along with the three principal components. In particular, apart from the first dimension that is more related to the E score, the second appears as a combination of S and G, whereas the third as a weighted mean of the three scores. Thus, PCA would result not to be useful for building up a synthetic indicator, as evidently there is no clear latent trait to be revealed, especially with regards to the first component, which should explain the largest part of the total variability. Indeed, again from Table 5 we can see that each component seems to explain similar quota of total variability (around 30%) and this is a further confirmation of the absence of a latent ESG trait that PCA can exploit. This could possibly depend on the type of approach adopted by PCA, which, differently from our approach, does not consider by design the whole distribution of the input variables. PCA rather focuses on the creation of alternative variables representing new directions in the latent space by maximizing the amount of explained variability contained in the original data.

6 Conclusions

This paper proposes an approach for synthesizing ratings or, more in general, indicators in the ESG context. Indeed, traditionally used position indexes do not properly represent the complexity of the data and, in particular, cannot properly deal with the differences in distribution between the available ratings. The well-known mean, for example, beyond not being robust if not enriched with information regarding the variability, is not able to distinguish between two or more distributions, if characterized by the same location parameter.

In this paper, we propose a new ESG metric based on the methodology developed by Agosto et al. (2025), who showed how to use optimal transport from the Wasserstein measure perspective to create aggregated indicators. By exploiting the distribution characteristics of each initial rating (E, S and G), we produce an aggregated ESG indicator merging the information retrieved from the different factors, and the associated confidence interval. This also allows to consistently compare the ratings regarding the different dimensions (E,S and G) and assess their contribution to the aggregated indicator, thus complying with the explanability principle, which is crucial for the understanding from the final users' perspective.

To provide a comprehensive understanding of the features of the proposed approach, we present an empirical study based on the investigation of a dataset composed by E, S and G ratings assigned to SMEs. Our results show that the proposed aggregated indicator can consistently summarize the information contained in the original ratings by capturing the natural notion of an "average distribution", since it preserves the shape of the initial distributions. Moreover, through our optimal mapping, we produce an aggregated rating that has the same distribution of the Wasserstein mean. The new consensus indicator can also provide meaningful and intuitive results in terms of the contribution of each initial rating. Furthermore, we compared the new approach to a well-known machine learning algorithm represented by Principal Component Analysis, widely employed to create synthetic indicators. Results have clearly shown the difficulties of PCA in producing a representative synthesis of the three dimensions.

Further research would consider the same method to build a consensus rating by aggregating evaluations provided by different raters, who may rely on different input variables and/or expert-based considerations, and potentially use different algorithms and weighting systems. Moreover, we will investigate a generalization of the consensus indicator in a multivariate context, when input variables used by providers to compute ESG scores can be added to better produce the final consensus rating.

Acknowledgements

This research received funding from European Union- Next Generation EU, Component M4C2, Investment 1.1, Progetti di Ricerca di Rilevante Interesse Nazionale (PRIN) – Notice 1409, 14/09/2022- BANDO PRIN 2022 PNRR. Project title: "Climate risk and uncertainty: environmental sustainability and asset pricing". Project code "P20225MJW8", CUP: F53D23009320001 and from PNRR - Grins – Growing Resilient, Inclusive and Sustainable (Grins)", ID: Pe00000018, Cup: C93c22005270001.

References

Abhayawansa, S. and Tyagi, S. (2021). Sustainable investing: the black box of environmental, social, and governance (esg) ratings. *The Journal of Wealth Management*, 24(1):49–54.

Agosto, A., Balzanella, A., and Cerchiello, P. (2025). A new compound indicator based on optimal transport. *Statistics*, pages 1–22.

Agosto, A., Cerchiello, P., and Giudici, P. (2023a). Bayesian learning models to measure the relative impact of esg factors on credit ratings. *International Journal of Data Science and Analytics*, pages 1–12.

Agosto, A., Giudici, P., and Tanda, A. (2023b). How to combine esg scores? a proposal based on credit rating prediction. *Corporate Social Responsibility and Environmental Management*, 30(6):3222–3230.

Agueh, M. and Carlier, G. (2011). Barycenters in the wasserstein space. *SIAM Journal on Mathematical Analysis*, 43(2):904–924.

Avetisyan, E. and Ferrary, M. (2013). Dynamics of stakeholders' implications in the institutionalization of the csr field in france and in the united states. *Journal of Business Ethics*, 115(1):115–133.

Berg, F., Koelbel, J. F., and Rigobon, R. (2022). Aggregate confusion: The divergence of esg ratings. *Review of Finance*, 26(6):1315–1344.

Billio, M., Costola, M., Hristova, I., Latino, C., and Pelizzon, L. (2021). Inside the esg ratings:(dis) agreement and performance. *Corporate Social Responsibility and Environmental Management*, 28(5):1426–1445.

Cerchiello, P. and Giudici, P. (2014). Bayesian credit ratings. *Communications in Statistics-Theory and Methods*, 43(4):867–878.

Cuesta-Albertos, J., Ruschendorf, L., and Tuero-Diaz, A. (1993). Optimal coupling of multivariate distributions and stochastic processes. *Journal of Multivariate Analysis*, 46(2):335–361.

Dimson, E., Marsh, P., and Staunton, M. (2020). Divergent esg ratings. *The Journal of Portfolio Management*, 47(1):75–87.

Dorfleitner, G., Halbritter, G., and Nguyen, M. (2015). Measuring the level and risk of corporate responsibility–an empirical comparison of different esg rating approaches. *Journal of Asset Management*, 16(7):450–466.

EBA (2020). Discussion paper on management and supervision of esg risks for credit institutions and investment firms. *EBA/DP/2020/03*.

ESMA (2020a). No action letter on sustainability-related disclosures for benchmarks. https://www.esma.europa.eu/sites/default/files/library/esma41-137-1300_esmar_article_9a3_opinion_-_bmr_nca.pdf.

ESMA (2020b). Strategy on sustainable finance. ESMA 22-105-1052. https://www.esma.europa.eu/press-news/esma-news/esma-sets-out-its-strategy-sustainable-finance.

Fréchet, M. (1960). Sur les tableaux dont les marges et des bornes sont données. *Revue de l'Institut International de Statistique / Review of the International Statistical Institute*, 28(1/2):10–32.

Giudici, P., Mezzetti, M., and Muliere, P. (2003). Mixtures of products of dirichlet processes for variable selection in survival analysis. *Journal of statistical planning and inference*, 111(1-2):101–115.

Gucciardi, G., Ossola, E., Parisio, L., and Pelagatti, M. M. (2024). Common factors behind companies' environmental ratings. *University of Milan Bicocca Department of Economics, Management and Statistics Working Paper Forthcoming.*

Muñoz-Torres, M. J., Fernández-Izquierdo, M. Á., Rivera-Lirio, J. M., and Escrig-Olmedo, E. (2019). Can environmental, social, and governance rating agencies favor business models that promote a more sustainable development? *Corporate Social Responsibility and Environmental Management*, 26(2):439–452.

Panaretos, V. M. and Zemel, Y. (2019). Statistical aspects of wasserstein distances. *Annual Review of Statistics and Its Application*, 6(Volume 6, 2019):405–431.

Pollman, E. (2022). The making and meaning of esg. *U of Penn, Inst for Law & Econ Research Paper*, (22-23).

Santambrogio, F. (2015). *Optimal Transport for Applied Mathematicians: Calculus of Variations, PDEs, and Modeling.* Progress in Nonlinear Differential Equations and Their Applications. Springer International Publishing.

Scalet, S. and Kelly, T. F. (2010). Csr rating agencies: What is their global impact? *Journal of Business Ethics*, 94(1):69–88.

Villani, C. (2003). *Topics in Optimal Transportation.* Graduate studies in mathematics. American Mathematical Society.

Villani, C. (2009). *The Wasserstein distances*, pages 93–111. Springer Berlin Heidelberg, Berlin, Heidelberg.