big data and cognitive computing

MDPI

Article

The "Unreasonable" Effectiveness of the Wasserstein Distance in Analyzing Key Performance Indicators of a Network of Stores

Andrea Ponti [1,2,*], Ilaria Giordani [1,3], Matteo Mistri [1], Antonio Candelieri [2] and Francesco Archetti [3]

[1] Oaks S.R.L., 20125 Milan, Italy
[2] Department of Economics, Management and Statistics, University of Milano-Bicocca, 20126 Bicocca, Italy
[3] Department of Computer Science, Systems and Communication, University of Milano-Bicocca, 20126 Bicocca, Italy
* Correspondence: andrea.ponti@unimib.it

Abstract: Large retail companies routinely gather huge amounts of customer data, which are to be analyzed at a low granularity. To enable this analysis, several Key Performance Indicators (KPIs), acquired for each customer through different channels are associated to the main drivers of the customer experience. Analyzing the samples of customer behavior only through parameters such as average and variance does not cope with the growing heterogeneity of customers. In this paper, we propose a different approach in which the samples from customer surveys are represented as discrete probability distributions whose similarities can be assessed by different models. The focus is on the Wasserstein distance, which is generally well defined, even when other distributional distances are not, and it provides an interpretable distance metric between distributions. The support of the distributions can be both one- and multi-dimensional, allowing for the joint consideration of several KPIs for each store, leading to a multi-variate histogram. Moreover, the Wasserstein barycenter offers a useful synthesis of a set of distributions and can be used as a reference distribution to characterize and classify behavioral patterns. Experimental results of real data show the effectiveness of the Wasserstein distance in providing global performance measures.

Keywords: Wasserstein distance; customer experience; key performance indicators

Citation: Ponti, A.; Giordani, I.; Mistri, M.; Candelieri, A.; Archetti, F. The "Unreasonable" Effectiveness of the Wasserstein Distance in Analyzing Key Performance Indicators of a Network of Stores. *Big Data Cogn. Comput.* **2022**, *6*, 138. https://doi.org/10.3390/bdcc6040138

Academic Editors: Domenico Talia, Fabrizio Marozzo and Min Chen

Received: 7 October 2022
Accepted: 11 November 2022
Published: 15 November 2022

1. Introduction

1.1. Motivations

Among the many facets of omni-channel retailing, this paper refers to a set of analytics and decision processes that support the seamless focus of a brand across many channels (in-store, online, mobile, call center or social). Retailers have come to recognize the importance of integrating information and services from multiple available channels to reduce data mismatch in order to create a seamless Customer eXperience (CX) and to obtain data-supported insight into the management of a network of stores. However, it is important to identify, promote and provide customers with various experiential benefits to enhance both shopping intentions and satisfaction. Although price and convenience are still primary considerations, customers are putting more emphasis on competence in specific categories and the overall customer experience. This aspect is particularly strong for categories that are highly fragmented or in which advice to customer plays a large role in sales, such as furniture, do-it-yourself products, apparel and consumer electronics. Personalization, meaning the quality of individual attention and tailored service, is largely regarded as the top criterion in evaluating CX. The analysis of customer data, from questionnaires and the analyses of online behavior, is instrumental in providing personalized services such as customized purchase recommendations, sending promotion information based on individual preferences and providing location-based services. The focus of this paper is on

the analysis of CX while considering a multinational retail company operating through a network of stores. To enable this analysis, a number of key performance indicators (KPI), acquired for each customer through different channels, are associated to the main drivers of the customer experience. It is important to remark that this analysis must be performed from a granular perspective on what a consumer really wants, today and in the future, in order to understand which services/products to offer on which channel. Developing this detailed understanding of consumers requires harnessing consumer data, which should be combined with consumer behavior insight from interviews and observations. It also requires analytics, which can work at the required granular level, gain a clear understanding of consumer expectations and derive a global picture of the strengths and weaknesses of each store. Capturing the full potential of omni-channel retailing requires a cross-channel perspective and transparency to measure and manage channel interplay, obtaining at the same time measures for the entire network of stores and improvement actions. More recently, the use of machine learning methods has been gaining more importance to leverage the wealth of customer data into a richer representation of the CX. It is the opinion of the authors of this paper that, given the growing number of channels and heterogeneity of customers, the standard statistical approach, which analyzes samples of the customer behavior only on parameters such as average and variance, might capture only a part of the hidden value of the data.

This paper proposes a different approach in which the samples from customer surveys are represented as discrete probability distributions, in particular as histograms or cloud points. In this distributional context, the variation in performance between two stores, considering one KPI, is the distance between two univariate histograms. The method can be naturally extended to jointly consider several KPIs, leading, for each store, to a multi-variate histogram. The statistical and, more recently, the machine learning communities have developed many alternative models to measure the distance between distributions. A general class of distances, known as f-divergences, is based on the expected value of a convex function of the ratio of two distributions. Some examples are Kullback–Leibler (and its symmetrized version Jensen–Shannon), Hellinger, Total Variation and χ-square divergence. In this paper, the focus is on the Wasserstein (WST) distance. Although other distances measure pointwise differences in densities (or weights), the WST distance (also known as the optimal transport distance) is a cross-binning distance; this distinction can be summed up by saying that the optimal transport distance is horizontal, whereas other distances are based on vertical displacement. Two important elements of the WST theory are the barycenter and WST clustering. The WST barycenter offers a useful synthesis of a set of distributions. A standard clustering method such as k-means can be generalized to WST spaces, enabling the WST barycenters and k-mean WST clustering, which is used to characterize and classify behavioral patterns. In general, WST enables the synthetization of a comparison between two multi-dimensional distributions through a single metric by using all information in the distributions. Moreover, the WST distance is generally well defined and provides an interpretable distance metric between distributions.

This study was motivated by the emerging need for a multination retailer to revise the performance measurement system—currently based on NPS—which has been adopted to rank the 50 stores of its commercial network. The limitations of NPS and the desire to design a new performance measurement system able to deal with multiple KPIs coming from omni-channel customer surveys lead us to propose a completely new analytical framework based on multi-variate discrete distributions and the Wasserstein distance. Indeed, using a more comprehensive system to evaluate the relative performance of each store with respect to the others is a critical decision for the company as a basis for the distribution of a performance-related bonus (on a quarterly basis), which is subject to negotiation with trade unions. Although multi-channel surveys are available, this study focuses on only one specific channel to better evaluate the benefits and limitations of the new framework.

1.2. Related Works

The cornerstone of the implementation of a CX strategy is the metric used to measure the performance of a company. A widely used such metric is the Net Promoter Score [1], which is associated with customer loyalty and is considered a reliable indicator of the future of a company's performance.

The author of [2] offered a view about a complete system of performance measurements for an enterprise based on over twenty years of research and development activities. The system was designed to provide key persons at different units/levels with useful quantitative information, such as board members to exercise due diligence, leaders to decide where to focus attention next and people to carry out their work well. Later, the author of [3] provided a review of various methods for tackling performance measurement problems. Although technical statistical issues are buried somewhat below the surface, statistical thinking is very much part of the main line of the argument, meaning that performance measurements should be an area attracting serious attention from statisticians. More recently, the authors of [4] re-visited the use of NPS (Net Promoter Score) as a predictor of sales growth by analyzing data from seven brands operating in the U.S. sportswear industry measured over five years. Interestingly, the results confirmed that, although the original premises are reasonable, methodological concerns arise when NPS is used as a metric for tracking overall brand health. Only the more recently developed brand health measure of NPS (using an all-potential customer samples) is effective at predicting future sales growth.

An interesting approach leveraging machine learning to analyze Customer Experience (CX) was proposed in [5,6]. The authors of these works considered beyond the NPS and the Customer SATisfaction score (CSAT) to measure the CX, and they performed a wide comparative evaluation of several machine learning approaches, analyzing the specific case of a telecommunication company and applying a wide set of classification methods to categorize the survey results.

In this paper we propose a distributional approach to performance evaluation; the performance is measured through KPIs represented as discrete probability distributions whose similarities are computed through the Wasserstein distance. The Wasserstein distance can be traced back to the works of Gaspard Monge [7] and Lev Kantorovich [8]. Recently, also under the name of the Earth Mover Distance (EMD), it has been gaining increasing importance in several fields, such as Imaging [9], Natural Language Processing [10] and a generation of adversarial networks [11]. Important references include [12], which gave a complete mathematical characterization, and [13], which also gave an up-to-date survey of numerical methods. The authors of [14] provided an overview of the Wasserstein space. A specific analysis of its geometry and geodesic Principal Components Analysis was given in [15]. Specific computational results related to barycenters and clustering were given in [16]. A novel Wasserstein distance and fast clustering method were proposed in [17]. One should note that the computational cost of the WST distance is amplified in computations of the barycenters of multi-variate distributions for computational as well as theoretical reasons [13].

The Wasserstein distance has also been receiving attention in economic theory, where the key reference is [18], in which it was shown that a number of seemingly unrelated problems can be modelled and solved as optimal transport problems. For the term "unreasonable effectiveness" in the title of this paper, we are indebted to [19]. Some key problems in finance have been also dealt with using optimal transport as the pricing of financial derivatives [18] and the analysis of robustness in risk management [20]. Other contributions to finance are [21], which provided a Wasserstein-based analysis of stability in finance, and [22], which proposed Wasserstein k-means clustering to classify market regimes. An important application domain of the Wasserstein distance is the analysis of distributional robustness. In [23], the authors analyzed Wasserstein-based distributionally robust optimization and its application in machine learning using the Wasserstein metric [24,25]. Two contributions, along the line of stochastic programming, were given

in [26], which proposed an approximation of data-driven chance-constrained programs over Wasserstein balls, and in [27], which proposed a distributionally robust two-stage Wasserstein model with recourse. We are not aware of significant applications of the Wasserstein distance in management science. A management topic where the Wasserstein distance enables significant contributions is the design of recommender systems using metric learning [28,29], which has shown to enable the measurement of uncertainty and the embedding of user/item representations in a low-dimensional space.

1.3. Contributions

The main contribution of this paper is the representation of performance metrics as measured through KPIs as discrete probability distributions. Embedding these distributions in the "Wasserstein space" enables the comparison and ranking of different stores. In addition, through the definition of Wasserstein barycenters, it is possible to perform clustering in the Wasserstein space with the aim of finding groups of similar stores. Moreover, since some KPIs are correlated with each other, in this paper, a subset of the most "informative" ones are chosen using feature selection and information gain. To further motivate the usage of the Wasserstein distance, a barycenter-based measure of how KPI data are not Euclidean is proposed; the computational results show that the discrepancy between the analysis in the Euclidean space and the WST space grows with the size of the subset.

2. Key Performance Indicators and the Formulation of the Problem

The focus of this paper is on a multinational retailer company which operates through a network of stores. The performance of each store is characterized in terms of service to the customer and is evaluated by the customers themselves through a number of Key Performance Indicators. Each store receives its evaluation through a survey composed of a number of questions. For each question, a customer can answer with a number on a scale from -100 to 100, which represents the satisfaction of a specific service. Each KPI_i, with $i = 1, \ldots, K$, is computed as the average of a set of questions and captures one feature of the customer experience. Figure 1 shows an example considering the experience of a customer inside a store. This aspect of the CX can be evaluated through seven different KPIs, each of which is obtained from the answers to a set of different questions.

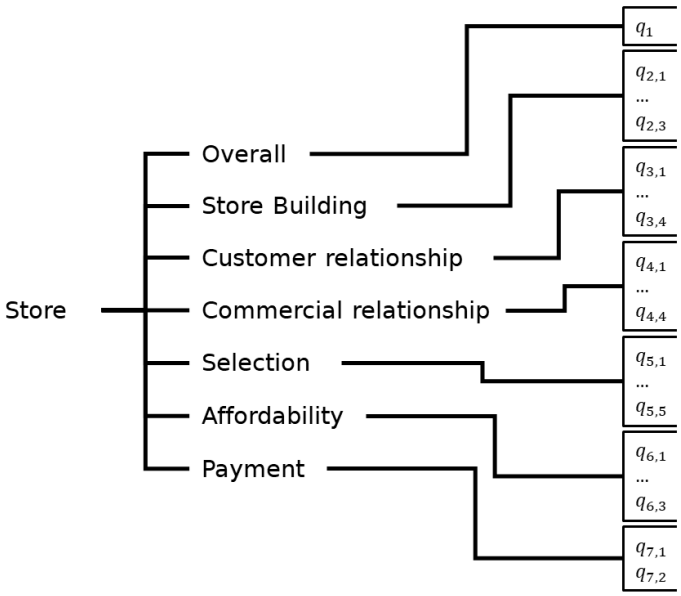

Figure 1. An example of the KPI tree related to the experience inside a store.

The objective of this study is to propose a system to assess stores' performances while simultaneously considering different KPIs. As a case study, a network of 50 stores owned

and operated by a multinational retailer is considered. In this paper, the seven KPIs related to the customer experience inside the store are considered. The following list of KPIs provides an idea of the scope of this study:

- Overall: Measures the overall sentiment of the customer for the whole process.
- Store Building: Features of the store, such as parking spaces and cleanliness.
- Customer Relationship: Measures sentiment about the vendors.
- Commercial Relationship: Aggregates scores given by customers in the customer relation before conversion.
- Selection: Aggregates scores from features such as the availability of products and clarity of presentation.
- Affordability: Aggregates scores from customers related to prices and discounts.
- Payment: Aggregates scores such as the length of the queue and easy payment.

Usually, the mean of each KPI for each store is analyzed to build a ranking or to evaluate different aspects of the stores. A very effective way to visualize these means is by using the parallel coordinates plot, as shown in Figure 2. This chart enables the easy and clear visualization of a set of points (stores) in a multi-dimensional space (KPIs).

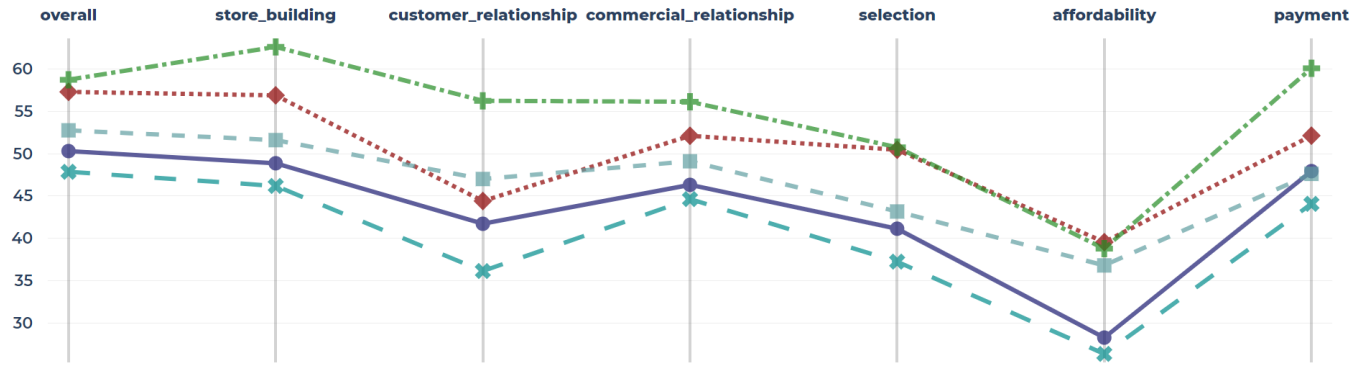

Figure 2. Parallel coordinates plot showing the seven KPIs of five stores (each line represents a store).

3. Space of Data and Distributional Representation

3.1. Distributional Representation

All the data of a store s_i can be stored in a matrix $L^{(s_i)} \in \mathbb{R}^{m \times K}$, in which the columns represent the K KPIs, and the rows represent the m users that completed the survey for a specific store s_i (Table 1). Then, each cell contains the value of a KPI for a customer.

Table 1. Matrix representing store s_i. Each column refers to a KPI, and each row refers to a customer.

$L^{(s_i)}$	KPI_1	KPI_2	...	KPI_K
1				
2				
...				
m				

Each column of $L^{(s_i)}$ can be considered to be a sample of the data related to a KPI. A column k can then be represented as a one-dimensional histogram $h_k^{(s_i)}$, whose support space $[z_k, u_k]$ can be divided into η bins. The weight of each bin is given by the number of customers of the sample, whose score for the specific KPI falls into that bin. Figure 3 shows an example of the histograms associated with three different stores regarding a KPI.

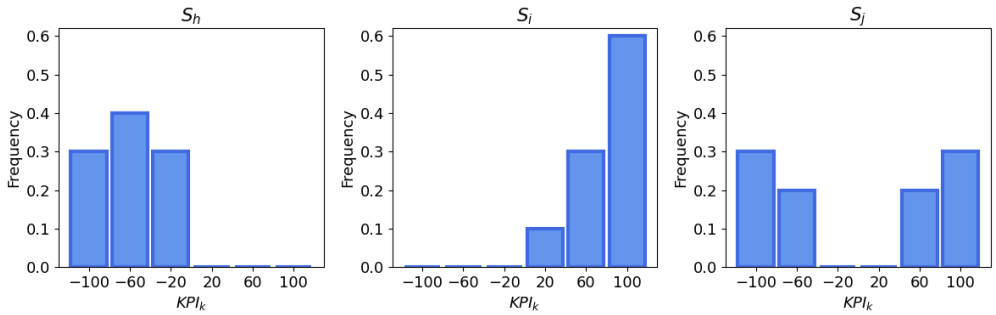

Figure 3. Three different stores represented as univariate histograms. KPI values are on the x-axis, and their relative frequencies are on the y-axis.

As each histogram represents a single KPI, it is possible to compute the distance between two stores as the distance between the two histograms given by the same KPI. This representation naturally extends to multi-dimensional histograms. Characterizing a store using all KPIs, each store is represented as a K-dimensional histogram. For instance, considering two KPIs, the supports of the two-dimensional bins are squares, and the weights of the bins are the number of customers whose KPI_i and KPI_j scores fall into that bin. The natural representation is a heatmap, as shown in Figure 4.

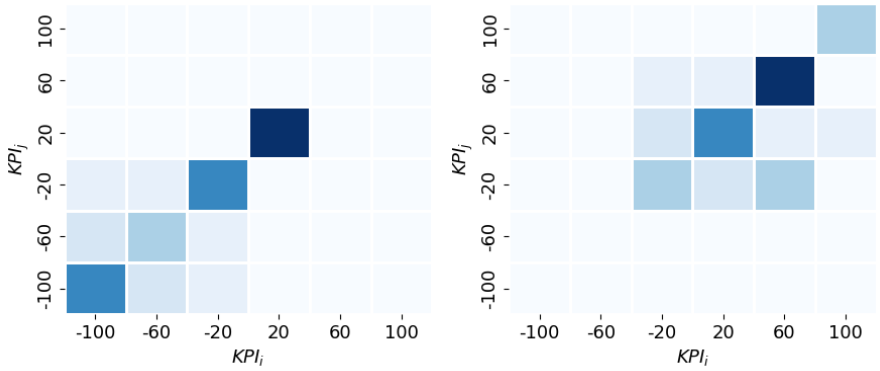

Figure 4. Two different stores represented as bivariate histograms. KPI values related to KPI_i and KPI_j are on the x-axis and y-axis, respectively, and each bin is colored by their relative frequencies.

Since histograms are instances of discrete probability distributions, the stores become elements in a probabilistic space. Another characterization of stores in this probabilistic space can be obtained by representing the matrices $L^{(s_i)}$ as point clouds. Figure 5 displays an example of point cloud representation. On the left, one KPI for two stores is shown, and on the right, a plot of the same two stores for two KPIs is shown.

Figure 5. Point cloud representations of two stores. The **left** plot considers one KPI: KPI values are on the x-axis, and the absolute frequency is on the y-axis. The **right** plot considers two KPIs: KPI values are on the x-axis and y-axis, and each point represents a user.

The set of all KPIs is denoted as S. The power set of S is the set of all subsets, including the empty one and S itself. If S has cardinality K, then the number of subsets is 2^K. All subsets but the empty one can be regarded as a description of a store. Therefore, the analysis can be performed on each element (except the empty one) of the power set of S.

A subset of cardinality $k = 1, \ldots, K$ is associated with store k's KPIs, which can be analyzed as k one-dimensional histograms or one k-dimensional histogram. The informational value of the two approaches is different, and the computational cost is also very different, as it increases with k. To mitigate this cost, one can choose the most significant KPIs using feature selection methods, as outlined in Section 5.1.

The histogram is a convenient representation of the $m \times K$ matrix $L^{(s_i)}$ in a space $\mathbb{R}^d$, where $d = \eta^K$, with η representing the number of bins. It is important to remark that d does not depend on the number of users m and can be reduced by considering an element of the power set S of cardinality $k < K$ or a smaller number of bins.

3.2. Graph Representation

An effective way to visualize all the stores and their similarities is by building a graph $G = (V, E)$, where the vertices V represent the stores that are connected with an edge if their similarities are above a given threshold. As previously mentioned, each store can be represented as a k-dimensional histogram $H^{(s_i)}$. Therefore, the set of edges can be given by $E = \left\{ (s_i, s_j) : D\left(H^{(s_i)}, H^{(s_j)}\right) < \tau \right\}$. Any distance between histograms can be used, and in this case, the Wasserstein distance (whose basic definition and properties are provided in Section 3) is considered. Figure 6 shows an example of the graph resulting from 4 KPIs and 50 stores. In this case, only the stores whose distances are below the first decile are connected.

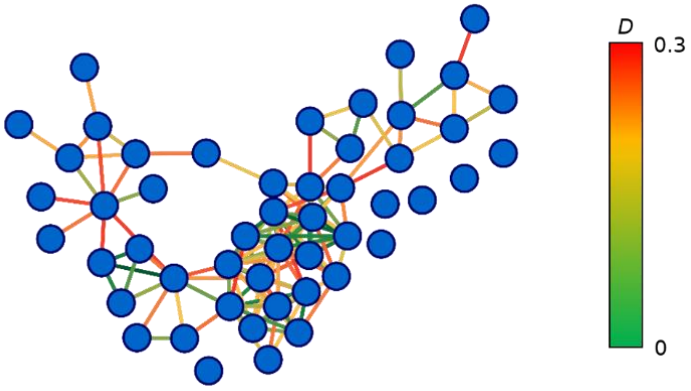

Figure 6. Graph representation of 50 stores. Edges are colored from green to red based on the distance from each other.

4. Wasserstein Distance

4.1. Basic Definitions

Consider the case of a discrete distribution P specified by a set of support points x_i with $i = 1, \ldots, m$ and their associated probabilities w_i, such that $\sum_{i=1}^{m} w_i = 1$ with $w_i \geq 0$ and $x_i \in M$ for $i = 1, \ldots, m$. Usually, $M = \mathbb{R}^d$ is the d-dimensional Euclidean space where x_i are the support vectors. M can also be a symbolic set provided with a symbol-to-symbol similarity. Therefore, P can be written as follows in Equation (1):

$$P(x) = \sum_{i=1}^{m} w_i \delta(x - x_i) \tag{1}$$

where $\delta(\cdot)$ is the Kronecker delta.

The WST distance between two distributions $P^{(1)} = \left\{w_i^{(1)}, x_i^{(1)}\right\}$ with $i = 1, \ldots, m_1$ and $P^{(2)} = \left\{w_i^{(2)}, x_i^{(2)}\right\}$ with $i = 1, \ldots, m_2$ is obtained by solving the following linear program (2):

$$W\left(P^{(1)}, P^{(2)}\right) = \min_{\gamma_{ij} \in \mathbb{R}^+} \sum_{i \in I_1,\, j \in I_2} \gamma_{ij}\, d\left(x_i^{(1)}, x_j^{(2)}\right) \tag{2}$$

The cost of transport between $x_i^{(1)}$ and $x_j^{(2)}$, $d\left(x_i^{(1)}, x_j^{(2)}\right)$ is defined by the p-th power of the norm $\|x_i^{(1)},\ x_j^{(2)}\|$, which is usually the Euclidean distance.

Two index sets can be defined as $I_1 = \{1, \ldots, m_1\}$ and I_2 likewise, such that

$$\sum_{i \in I_1} \gamma_{ij} = w_j^{(2)},\ \forall j \in I_2 \tag{3}$$

$$\sum_{j \in I_2} \gamma_{ij} = w_i^{(1)},\ \forall i \in I_1 \tag{4}$$

Equations (3) and (4) represent the in-flow and out-flow constraints, respectively. The terms γ_{ij} are called matching weights between support points $x_i^{(1)}$ and $x_j^{(2)}$ or the optimal coupling for $P^{(1)}$ and $P^{(2)}$. The basic computation of OT between two discrete distributions involves solving a network flow problem whose computation typically scales cubically in the sizes of the measure. In the case of a one-dimensional histograms, the computation of the Wasserstein distance can be performed by a simple sorting algorithm and with the application of Equation (5).

$$W_p\left(P^{(1)}, P^{(2)}\right) = \left(\frac{1}{n}\sum_{i}^{n}\left|x_i^{(1)*} - x_i^{(2)*}\right|^p\right)^{\frac{1}{p}} \tag{5}$$

where $x_i^{(1)*}$ and $x_i^{(2)*}$ are the sorted samples. The discrete version of the WST distance is usually called the Earth Mover Distance (EMD). For instance, when measuring the distance between grey scale images, the histogram weights are given by the pixel values and the coordinates by the pixel positions.

Consider now the three univariate histograms in Figure 3, which represent three different stores. Support x_i is the range of values of the KPI, and the weights w_i are the number of users whose KPI score falls into that interval. Table 2 shows the differences between the Wasserstein distance and the Manhattan and Euclidean distances.

Table 2. The difference between Manhattan, Euclidean and Wasserstein distances.

Distance	Order	$D(S_h,S_i)$	$D(S_h,S_j)$	$D(S_i,S_j)$
Manhattan	1	2.000	1.000	1.000
Euclidean	2	0.894	0.510	0.490
Wasserstein	1	0.583	0.250	0.333
	2	0.677	0.324	0.374

The Wasserstein distance agrees with the intuition that S_h is closer to S_j than S_i. Instead, the Manhattan distance does not discriminate because it assigns the same value to the pairs (S_h, S_j) and (S_i, S_j). In [30], it was remarked that the information reflected in histograms lies more in the relative value of their coordinates rather than on their absolute value.

The computational cost of optimal transport can quickly become prohibitive. The method of entropic regularization [13] enables scalable computations, but large values of the regularization parameter can induce an undesirable smoothing effect, whereas low values not only reduce the scalability but might induce several numerical instabilities.

4.2. Barycenter and Clustering

Under the optimal transport metric, it is possible to compute the mean of a set of empirical probability measures. This mean is known as the Wasserstein barycenter and is the measure that minimizes the sum of its Wasserstein distances to each element in that set. Consider a set of N discrete distributions, $\mathbf{P} = \left\{P^{(1)}, \ldots, P^{(N)}\right\}$, with $P^{(k)} = \left\{\left(w_i^{(k)}, x_i^{(k)}\right) : i = 1, \ldots, m_k\right\}$ and $k = 1, \ldots, N$. Therefore, the associated barycenter, denoted with $\overline{P} = \{(\overline{w}_1, x_1), \ldots, (\overline{w}_m, x_m)\}$, is computed as follows in Equation (6):

$$\overline{P} = \underset{P}{\operatorname{argmin}} \frac{1}{N} \sum_{k=1}^{N} \lambda_k W\left(P, P^{(k)}\right) \tag{6}$$

where the values λ_k are used to weigh the different contributions of each distribution in the computation. Without the loss of generality, they can be set to $\lambda_k = \frac{1}{N} \; \forall \, k = 1, \ldots, N$.

The concept of the barycenter enables clustering among distributions in a space whose metric is the Wasserstein distance. More simply, the barycenter in a space of distributions is the analog of the centroid in a Euclidean space. The most common and well-known algorithm for clustering data in the Euclidean space is *k*-means. Since it is an iterative distance-based (also known as representative-based) algorithm, it is easy to propose variants of *k*-means by simply changing the distance adopted to create clusters, such as the Manhattan distance (leading to *k*-medoids) or any kernel allowing for non-spherical clusters (i.e., kernel *k*-means). The crucial point is that only the distance is changed, and the overall iterative two-step algorithm is maintained. This is also valid in the case of the Wasserstein k-means, where the Euclidean distance is replaced by the Wasserstein distance and where centroids are replaced by barycenters.

5. Results

5.1. Feature Selection

The computational complexity of the Wasserstein distance can quickly become intractable in the case of multi-variate histograms, as already mentioned. The computation of the barycenter and performing the clustering procedure using the WST distance add substantially to the computational cost. It is therefore important to reduce the number of variables to consider, and for this reason, a feature selection strategy based on the Information Gain (IG) is used to select the most relevant KPIs. In turn, each KPI is considered as a target variable in a classification problem, and the IGs of all the others KPIs are computed. Since seven KPIs are considered, for each of them, six different values of IG are obtained, each of which represents the importance for the specific KPI in predicting the other six. Therefore, for each KPI, the average of these six represents its IG. Table 3 reports these results. In the following analysis, the four most relevant KPIs are considered.

Table 3. Information Gain of the seven KPIs.

KPI	Information Gain
Selection	0.37
Customer Relationship	0.34
Commercial Relationship	0.33
Store Building	0.32
Affordability	0.30
Overall	0.29
Payment	0.26

5.2. Wasserstein Analysis

The distributional representation of the stores enables the definition of an ideal store that can be used to build a Wasserstein-based ranking. The histogram associated with the ideal store has the entire mass concentrated on the bin of the most favorable assessment.

For instance, Figure 7 shows the ideal univariate and bivariate histograms. This definition can be naturally extended to multi-dimensional histograms.

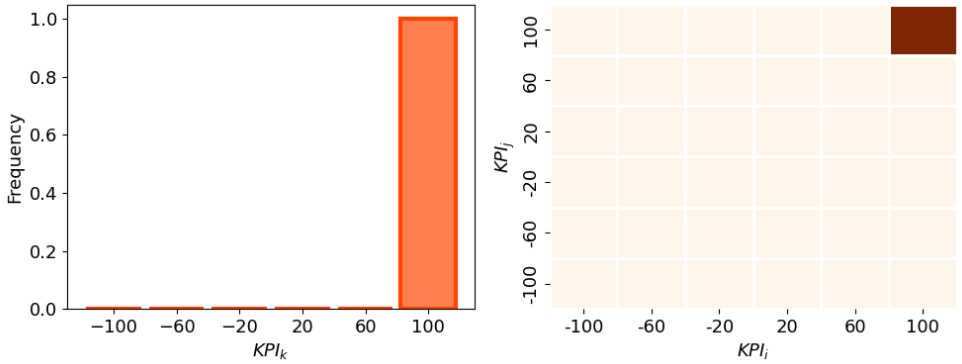

Figure 7. Univariate histogram associated with the ideal store (**left**): KPI values are on the x-axis, and their relative frequencies are on the y-axis. Bivariate histogram associated with the ideal store (**right**): KPI values related to KPI_i and KPI_j are on the x-axis and y-axis, respectively, and each bin is colored by their relative frequencies.

Given such a histogram, it is now possible to build a ranking computing the Wasserstein distance between the histograms associated with each store and the ideal one. In this way, the ranking is built upon the entire distributions of the KPIs' values and not only considering their means. To highlight the advantages of this framework, Figure 8 shows the correlation between the distances of the stores for the ideal and different statistics of the KPIs' distributions while considering histograms in one to four dimensions.

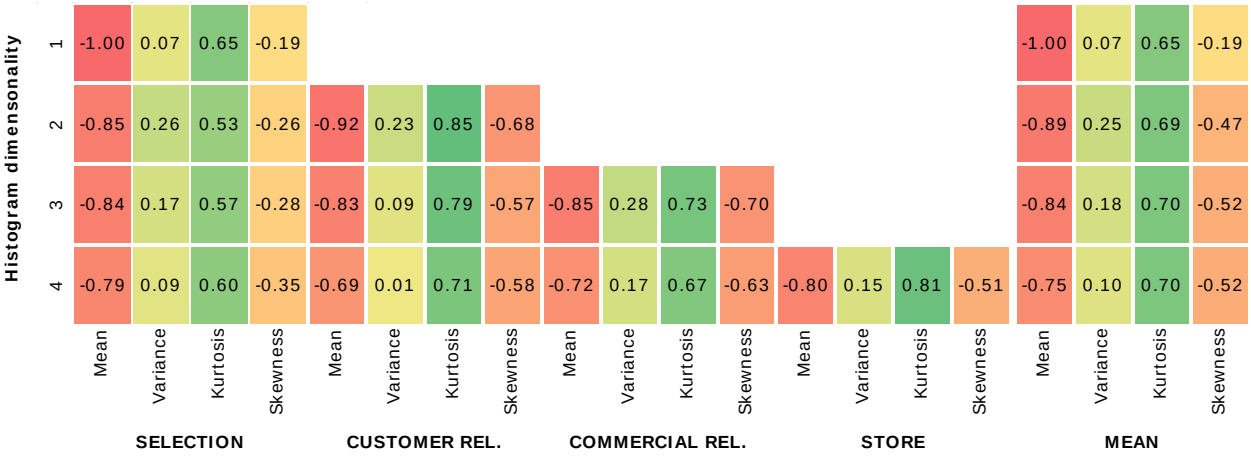

Figure 8. Pearson correlation between the distances of stores from the ideal and four statistics (mean, variance, kurtosis and skewness).

In the case of univariate histograms related to the KPI "selection", the correlation between the distance from the ideal and the mean is at a maximum; therefore, the rankings built upon the mean and the distance from being ideal are the same. By increasing the dimensionality of the data, the correlation of the Wasserstein distance with the mean tends to decrease, but the correlation with other statistics, such as kurtosis and skewness, tends to increase. The Wasserstein distance is able to capture other aspects or statistics of a distribution rather than just the mean, resulting in a more robust ranking.

5.3. Clustering Analysis

The barycenters enable the clustering of a set of distributions using the Wasserstein distance in a framework similar to *k*-means, as explained in Section 3.2. Two different clustering approaches are considered and compared with the standard *k*-means.

Consider a network of n stores and a set of k KPIs. Each store can be represented as k univariate histograms (one for each KPI) or one k-dimensional histogram. Clustering can be performed to divide the stores into two different groups. In the first case, each clustering iteration requires the computation of $2k$ different univariate barycenter and $2kS$ Wasserstein distances between univariate histograms. In the second case, each clustering iteration requires the computation of two different k-dimensional barycenter and $2S$ Wasserstein distances between k-dimensional histograms. The first approach considers just the marginals of the entire distribution of KPIs, losing the correlations between them and resulting in a more efficient but less effective algorithm. The second approach can instead quickly become too computationally expensive as the number of KPIs k grows.

These two approaches are compared with the k-means algorithm performed on the mean of KPIs. Each store is represented as a k-dimensional vector, where each component contains the mean of a KPI. To enable the visualization of the clustering, the results of the three algorithms are mapped on the network representation of the stores, as shown in Figure 9.

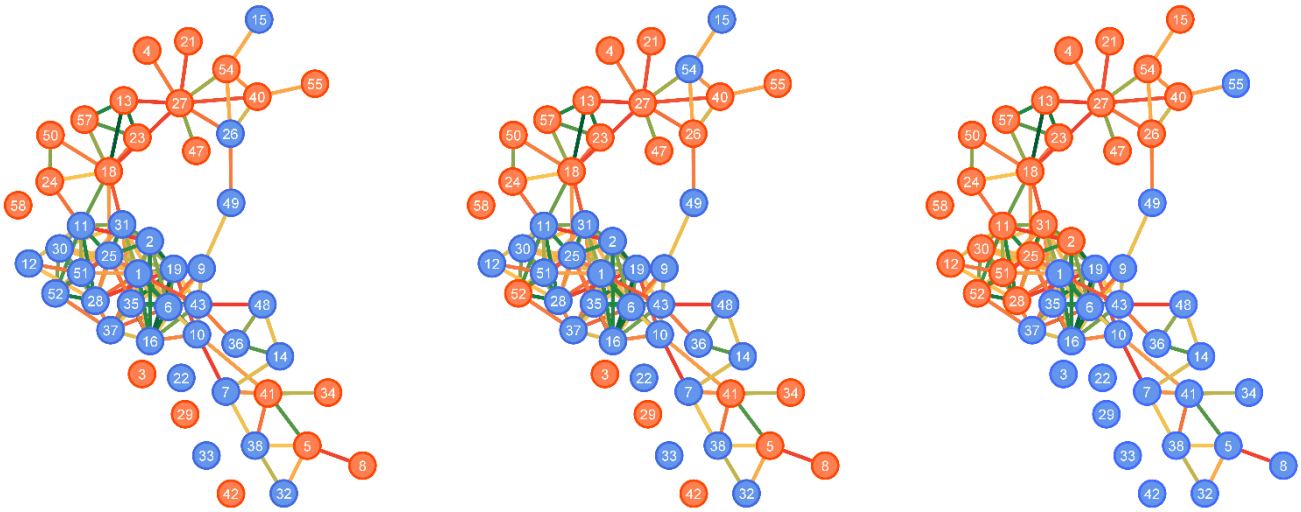

Figure 9. Clusters resulting from the three different approaches: k-means (**left**), clustering of the marginals (**center**) and clustering of the multi-dimensional histograms (**right**).

The resulting clusters using the standard k-means approach and the approach that considers just the marginals are visually similar, while the approach that consider the whole distributions of KPIs bring to different groups. Therefore, using the multi-dimensional histogram representations of the stores allows one to capture the entire distribution of the KPIs and their correlations, thus bringing different insight.

5.4. Nonlinear Structures in Data

A key assumption in this paper is that large datasets can exhibit a nonlinear structure, which is not easily captured by a Euclidean space. A key conjecture of this paper is that the WST space of histograms is a non-linear manifold. As a consequence, one can expect that embedding the problem in a Wasserstein space and using barycenters can provide a better synthesis of the dataset than the Euclidean mean.

To test this conjecture, the difference between the Euclidean mean and the barycenter is analyzed. First, a single KPI is considered, and the Euclidean mean and the barycenter of the histograms associated with the 50 stores are computed. The same process is also repeated in the cases of two, three and four KPIs to consider multi-dimensional histograms.

The computational results support the initial hypothesis. Figure 10 shows the Wasserstein distances between the Euclidean means of the histograms and the barycenters. This distance monotonically increases with the dimension of the support space of the histograms.

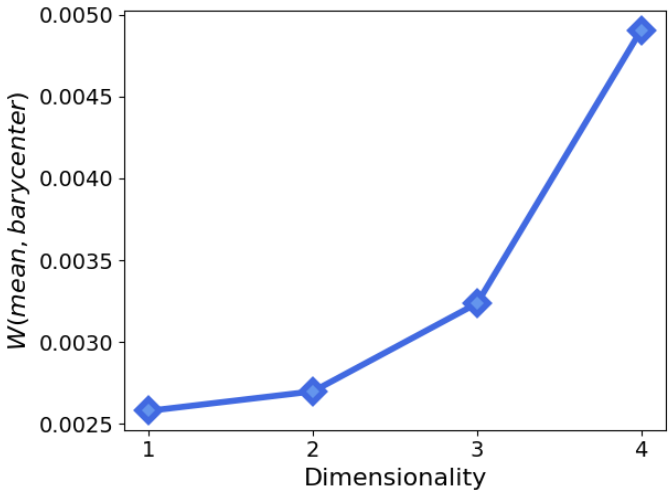

Figure 10. Distance between the Euclidean mean and the Wasserstein barycenter as the histograms' dimensionality increases. The dimensionality of the histograms is on the x-axis, and the Wasserstein distances between the Euclidean mean of the histograms and their Wasserstein barycenters are on the y-axis.

6. Conclusions, Limitations and Perspectives

The analytics proposed in this paper, based on the Wasserstein distance and barycenters, enables one to capture the quality of the customer experiences and to provide performance measures for the entire network of stores. It is the authors' opinion that the growing diversity and heterogeneity of customers makes a distributional approach more effective for analyzing samples of customer behavior than relying only on parameters such as average and variance. The Wasserstein distance (also known as the optimal transport distance) is shown to uncover nonlinear dependencies in the dataset without requiring the alignment of the distributions' support. This is demonstrated by the growing gap between the Euclidean average and the barycenter as the dimensionality of the support increases. The histograms can also be clustered in the Wasserstein space.

These features are demonstrated in a challenging business problem: the performance evaluation of the Italian store network (50 stores) of a multination retailer. Assessing the relative performance of each store with respect to the others is a critical decision for a company as a basis for the distribution of a performance-related bonus. The results enable the company to move towards a different evaluation platform. The analytics proposed in this paper, based on the Wasserstein distance and barycenters, is suitable to obtain a credible ranking system for the stores.

In terms of limitations, it is fair to remark that, although univariate distributions can be easily handled using the quantile-based closed formula, computational problems may hinder the application of the WST distance to large-scale multivariate problems. This problem is amplified in the computation of the barycenter and in the clustering of histograms in the Wasserstein space.

In terms of perspectives, it should be remarked that a byproduct of the computation of the WST distance between two stores is an optimal transport plan that indicates how much of the "probability mass" is to be moved between each couple of bins in the multivariate histograms representing the two stores. This result can be read as the impact of an improvement of each KPI on the overall score of a store.

Author Contributions: All authors contributed equally to this paper. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: The data are available upon request to Ilaria Giordani (giordani@oaks.cloud). The data are built on a real word project and were randomized during the study.

Acknowledgments: The authors greatly acknowledge the Data Science Lab, Department of Economics Management and Statistics (DEMS), for supporting this work by providing computational resources.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Reichheld, F.F. The One Number You Need to Grow. *Harv. Bus. Rev.* **2003**, *81*, 46–55.
2. Fisher, N.I. *Analytics for Leaders. A Performance Measurement System for Business Success*, 1st ed.; Cambridge University Press: Cambridge, UK, 2013.
3. Fisher, N.I. A Comprehensive Approach to Problems of Performance Measurement. *J. R. Stat. Soc. Ser. A Stat. Soc.* **2019**, *182*, 755–803. [CrossRef]
4. Baehre, S.; O'Dwyer, M.; O'Malley, L.; Lee, N. The Use of Net Promoter Score (NPS) to Predict Sales Growth: Insights from an Empirical Investigation. *J. Acad. Mark. Sci.* **2022**, *50*, 67–84. [CrossRef]
5. Markoulidakis, I.; Rallis, I.; Georgoulas, I.; Kopsiaftis, G.; Doulamis, A.; Doulamis, N. A Machine Learning Based Classification Method for Customer Experience Survey Analysis. *Technologies* **2020**, *8*, 76. [CrossRef]
6. Markoulidakis, I.; Rallis, I.; Georgoulas, I.; Kopsiaftis, G.; Doulamis, A.; Doulamis, N. Multiclass Confusion Matrix Reduction Method and Its Application on Net Promoter Score Classification Problem. *Technologies* **2021**, *9*, 81. [CrossRef]
7. Monge, G. Mémoire Sur La Théorie Des Déblais et Des Remblais. In *Histoire de l'Académie Royale des Sciences de Paris*; Nabu Press: Charleston, NC, USA, 1781; pp. 666–704.
8. Kantorovitch, L. On the Translocation of Masses. *Manag. Sci.* **1958**, *5*, 1–4. [CrossRef]
9. Bonneel, N.; Peyré, G.; Cuturi, M. Wasserstein Barycentric Coordinates: Histogram Regression Using Optimal Transport. *ACM Trans. Graph.* **2016**, *35*, 71-1. [CrossRef]
10. Huang, G.; Quo, C.; Kusner, M.J.; Sun, Y.; Weinberger, K.Q.; Sha, F. Supervised Word Mover's Distance. In Proceedings of the 30th International Conference on Neural Information Processing Systems, Barcelona, Spain, 5–10 December 2016; pp. 4869–4877.
11. Arjovsky, M.; Chintala, S.; Bottou, L. Wasserstein Generative Adversarial Networks. In Proceedings of the 34th International Conference on Machine Learning, Sydney, Australia, 6–11 August 2017; pp. 214–223.
12. Villani, C. *Optimal Transport: Old and New*; Springer: Berlin, Germany, 2008.
13. Peyré, G.; Cuturi, M. Computational Optimal Transport. *Found. Trends Mach. Learn.* **2019**, *11*, 355–607. [CrossRef]
14. Panaretos, V.M.; Zemel, Y. *An Invitation to Statistics in Wasserstein Space*; Springer: Berlin, Germany, 2020.
15. Bigot, J. Statistical Data Analysis in the Wasserstein Space. *ESAIM Proc. Surv.* **2020**, *68*, 1–19. [CrossRef]
16. Cohen, S.; Arbel, M.; Deisenroth, M.P. Estimating Barycenters of Measures in High Dimensions. *arXiv* **2020**, arXiv:2007.07105.
17. Verdinelli, I.; Wasserman, L. Hybrid Wasserstein Distance and Fast Distribution Clustering. *Electron. J. Stat.* **2019**, *13*, 5088–5119. [CrossRef]
18. Galichon, A. *Optimal Transport Methods in Economics*; Princeton University Press: Princeton, NJ, USA, 2018.
19. Galichon, A. The Unreasonable Effectiveness of Optimal Transport in Economics. *arXiv* **2021**, arXiv:2107.04700.
20. Kiesel, R.; Rühlicke, R.; Stahl, G.; Zheng, J. The Wasserstein Metric and Robustness in Risk Management. *Risks* **2016**, *4*, 32. [CrossRef]
21. Backhoff-Veraguas, J.; Bartl, D.; Beiglböck, M.; Eder, M. Adapted Wasserstein Distances and Stability in Mathematical Finance. *Financ. Stoch.* **2020**, *24*, 601–632. [CrossRef]
22. Horvath, B.; Issa, Z.; Muguruza, A. Clustering Market Regimes Using the Wasserstein Distance. *arXiv* **2021**, arXiv:2110.11848. [CrossRef]
23. Kuhn, D.; Esfahani, P.M.; Nguyen, V.A.; Shafieezadeh-Abadeh, S. Wasserstein Distributionally Robust Optimization: Theory and Applications in Machine Learning. *arXiv* **2019**, arXiv:1908.08729. [CrossRef]
24. Mohajerin Esfahani, P.; Kuhn, D. Data-Driven Distributionally Robust Optimization Using the Wasserstein Metric: Performance Guarantees and Tractable Reformulations. *Math. Program.* **2018**, *171*, 115–166. [CrossRef]
25. Lau, T.T.-K.; Liu, H. Wasserstein Distributionally Robust Optimization via Wasserstein Barycenters. *arXiv* **2022**, arXiv:2203.12136.
26. Chen, Z.; Kuhn, D.; Wiesemann, W. Data-Driven Chance Constrained Programs over Wasserstein Balls. *Oper. Res.* **2022**. [CrossRef]
27. Xie, W. Tractable Reformulations of Distributionally Robust Two-Stage Stochastic Programs With∞- Wasserstein Distance. *arXiv* **2019**, arXiv:1908.08454.
28. Ma, C.; Ma, L.; Zhang, Y.; Tang, R.; Liu, X.; Coates, M. Probabilistic Metric Learning with Adaptive Margin for Top-K Recommendation. In Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, Virtual Event, 6–10 July 2020; pp. 1036–1044.
29. Rakotomamonjy, A.; Traoré, A.; Berar, M.; Flamary, R.; Courty, N. Distance Measure Machines. *arXiv* **2018**, arXiv:1803.00250.
30. Le, T.; Cuturi, M. Adaptive Euclidean Maps for Histograms: Generalized Aitchison Embeddings. *Mach. Learn.* **2015**, *99*, 169–187. [CrossRef]