Article

Comprehensible Visualization of Multidimensional Data: Sum of Ranking Differences-Based Parallel Coordinates

Ádám Ipkovich ¹ , Károly Héberger ² and János Abonyi ^{1,*} ¹ MTA-PE “Lendület” Complex Systems Monitoring Research Group, University of Pannonia, Egyetem u. 10, H-8200 Veszprem, Hungary; ipkovich.adam@mk.uni-pannon.hu² ELKH Research Centre for Natural Sciences, Institute of Excellence of the Hungarian Academy of Sciences, Magyar Tudósok Krt. 2, H-1117 Budapest, Hungary; heberger.karoly@ttk.mta.hu

* Correspondence: janos@abonyilab.com

Abstract: A novel visualization technique is proposed for the sum of ranking differences method (SRD) based on parallel coordinates. An axis is defined for each variable, on which the data are depicted row-wise. By connecting data, the lines may intersect. The fewer intersections between the variables, the more similar they are and the clearer the figure becomes. Therefore, the visualization depends on what techniques are used to order the variables. The key idea is to employ the SRD method to measure the degree of similarity of the variables, establishing a distance-based order. The distances between the axes are not uniformly distributed in the proposed visualization; their closeness reflects similarity, according to their SRD value. The proposed algorithm identifies false similarities through an iterative approach, where the angles between the SRD values determine which side a variable is plotted. Visualization of the algorithm is provided by MATLAB/Octave source codes. The proposed tool is applied to study how the sources of greenhouse gas emissions can be grouped based on the statistical data of the countries. A comparison to multidimensional scaling (MDS)-based ordering is also given. The use case demonstrates the applicability of the method and the synergies of the incorporation of the SRD method into parallel coordinates.

Keywords: sum of the ranking differences; parallel coordinates; high-dimensionality data; visualization; greenhouse gas emissions; MATLAB/Octave toolbox

Citation: Ipkovich, Á.; Héberger, K.; Abonyi, J. Comprehensible Visualization of Multidimensional Data: Sum of Ranking Differences-Based Parallel Coordinates. *Mathematics* **2021**, *9*, 3203. <https://doi.org/10.3390/math9243203>

Academic Editors: Santoso Wibowo and Michael Li

Received: 31 October 2021

Accepted: 7 December 2021

Published: 11 December 2021

Publisher’s Note: MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Copyright: © 2021 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

1. Introduction

We present a method for the visualization of multidimensional data by extending the sum of ranking differences (SRD) method by a goal-oriented modification of the widely applied parallel coordinates method.

Parallel coordinates is a popular visualization technique thanks to its straightforward use and comprehensibility [1–3]. Axes are drafted for each variable, where the data points of each variable are depicted and connected to the respective neighboring axis. The method visualizes the correlation between the variables, as the intersection of the lines connecting the axes may indicate how different the variables are. The rank correlation coefficient (Kendall’s tau, τ) measures the correlation if intersections are present [4]. If no intersections can be observed between the lines, the ordering of the two variables is identical, and their rank correlation is equal to one. To provide comprehensible visualization, the number of intersections between the axes should be reduced. The intersections of the lines can be optimized through the careful placement of the variables on the x-axes, which is achieved by measuring the similarity between the axes [5]. Multidimensional scaling curtails the number of intersections by reducing the variables into 2D coordinates and ordering based on their proximity to each other [6]. Employing hierarchical clustering and occlusion algorithms enables the number of lines to be reduced, improving the comprehensibility of its visualization [7]. These examples illustrate that the technique of parallel coordinates requires the effective ordering of the variables based on their similarity.

The aim of this work is to improve the parallel coordinate visualization by ordering the variables based on the SRD method and also to provide a more detailed visualization of how the variables are grouped in the SRD method.

SRD provides an ordering based on how similar the variables are to a common reference, which can be an ideal one or aggregated gold standard [8–10]. As this straightforward method is visualized through the 1D ordering of the SRD values, there are inevitably limitations to the captured information in one figure. The correlation between the variables is not shown, nor is capable of determining mutually similar variables. The motivation behind this study was to bring more information to the figure of the SRD results. The aim of this work is to overcome the deficiencies of the present visualization tools for the SRD technique by a goal-oriented modification of the widely applied parallel coordinates method.

The key idea of this work is to incorporate SRD into the parallel coordinates method. As the parallel coordinates method excels at describing correlations, it seems reasonable to use the techniques for multi-criteria analysis methods, such as the SRD, which is used to determine the relationship of the variables to a gold standard. The SRD sorts the axes of parallel coordinates, efficiently depicting the ranking and showing the correlation between neighboring variables.

The SRD values promote either the close proximity of a variable to the ideal ranking or the random or reverse nature of the ranking to provide an unambiguous ordering, which may further enhance the visualization technique based on the parallel coordinates method.

For the visualization of mutually similar variables, two directions, namely on the left and right-hand sides on the x-axis are defined. As the deviation angle between the SRD values of the variables is another indicator of the consistency of the visualization, we developed an iterative algorithm that determines the angles between the last variables placed in two directions; the selected variable is placed to the left- or right-hand side of the reference based on the lesser deviation angle. Additionally, the method is compared to simple, multidimensional scaling and unidirectional SRD-based orderings.

The contribution of this work is as follows:

- We utilize SRD as the axes of parallel coordinates, providing a thorough visualization, and define two directions for dissimilar variables.
- We introduce a new visualization technique for the sum of ranking differences instead of the classical unidirectional ordering.
- We extend the parallel coordinates to two direction on the x-axis (to the left and right from the middle) to provide information on mutually dissimilar variables.
- We provide a toolbox for MATLAB and Octave to enhance the visualization of SRD-based parallel coordinates.

In the following, we first provide a review of the relevant studies and identify the major gaps in the state of the art and state how these gaps will be addressed by this study. In Section 3, the details of the method are presented. Section 4 demonstrates the applicability of the visualization tool in the analysis of the sources of greenhouse emissions. We compare the orderings based on the alphabet, MDS, SRD and the deviation angle-based SRD approaches to benchmark the orderings for the purpose of visualizing the parallel coordinates. According to the benchmarking results concerning the visualization methods, the angle-based method yields the best arrangement of the climate-change data.

2. Review of Related Methods—Motivation

The visualization of multivariate data focuses primarily on the visualization of data points, i.e., projecting data points into a lower-dimensional subspace. The principal component analysis (PCA) utilizes the correlation of data. Other techniques, such as Sammon mapping, preserve the distances or neighborhoods by searching for two- or three-dimensional mappings [11].

It is a crucial but distinct task to explore and visualize the relationships between the variables themselves. Exploratory data analysis is used to visualize the relationship between variables by using correlation and distance matrices (scatter matrix) widely and

unambiguously. Although it also represents the variables themselves, a scatter matrix also shows the correlation that a comparison between variables can determine. Heatmap-based representations of correlation matrices are probably the most common technique in this area. The seriation of the matrices also supports the representation of objects, where related groups of variables can be defined and arranged [12]. A further development is when the dominant relations can be interpreted as a network, and the internal structure of the variables is explored by displaying this network [13]. Classical multivariate statistical tools are also enriched by the visualization of the variables. The biplot visualization is developed in connection with PCA. The biplot represents the variables by vectors, and the angles of these vectors exert the similarity of the variables in a quantitative manner [14].

The application of the above methods is resource intensive, the evaluation of the results requires significant background knowledge, and the importance of correlations is not revealed. The SRD technique was created to fulfill the need for simplicity and interpretability: the SRD algorithm also serves an additional purpose besides exploring the relationship between the variables by aggregating the variables (aspects) and characterizing the other variables (aspects) in relation to them. Therefore SRD visualization not only sorts variables based on distance, but also illustrates the likelihood that rank ordering(s) can occur.

The advantage of the method is also confirmed by its broad applicability. It has been used for the following:

- To determine the similarity between models and facilitate the selection of models without considering weight allocation problems [15];
- For tea grade identification [16];
- In QSAR modeling to determine training-test set splits correctly [17];
- For comparison of performance parameters (merits) in QSAR/QSPR model validation [18];
- For column selection in chromatography [19,20];
- For comparison of lipophilicity parameters [21,22];
- For outlier detections [23];
- In political sciences to determine the optimal constituency size [24] and to rank universities and institutions [25];
- Ranking in sports [26].

The 1D visualization of SRD is very special. Colored bars are assigned to each variable: The x-axis and one of the y-axes contain the same quantity, namely, the scaled SRD values between 0 and 100, i.e., the 2D plot realizes, in fact, a 1D ordering. Hence, the lengths of the bars do not carry any information; their tops lie on the 45 grad line. The essential information is the distances of the bars from the zero (gold standard) and the random distribution. (The SRD distribution from random numbers is calculated from the exact theoretical distribution only if the number of objects (lines) is small (<14 for untied observations and <9 if ties are also present in the input matrix.) In all other cases, the random SRD distribution is well approximated with a Gaussian curve based on Monte Carlo simulations (preferably in the form of a cumulated distribution function). Figure 1 shows the traditional way to visualize the SRD ordering [8,9]. The large distance from zero would suggest that the present variables are not optimal (or the gold standard has little to do with them). The cumulated probability curve indicates that the variables located here are indistinguishable from random ranking. XX1, fairly 5%, shows the predefined error limit (the random order cannot be excluded; it happens less than 20 cases from 100). On the right to the median, the reverse rankings are located. If a line is placed to the right of the XX19 dotted line, the reverse ranking is considered to be significant at the 5% level. The grouping of lines also holds some information: not the distance from each other, but considering it on the condition of the gold standard. The ordering of SRD values is one-dimensional from left to right.

Figure 1. Explanation of the randomization test and 1D bar chart visualization. Blue curve is the cumulated Gaussian approximation of the random distribution, the probabilities are in the first y axis.

There is another visualization possibility for SRD. Namely, the uncertainty of SRD bars can be estimated by leave-one-out or leave-many-out cross validation or bootstrap. Preferably, such uncertainties can be plotted on a box and whisker plot, showing a kind of distributional information. It was invented in 2011 [8].

The ultimate visualization of SRD is a heatmap. It solves the problematic selection of the gold standard in a way that all variables serve as the gold standard once and only once; it is called comparisons with one variable at a time (COVAT) [22]. This distance matrix is symmetric and ordered according to the sum of the distances, column-wise. However, this matrix of pairwise distances with color codes shows only binary couplings, similarly to a correlation matrix, just in a reverse order (the smaller an SRD value is, the greater the similarity).

All ways of visualizations provide information for the variables and their ordering (grouping). The ranking of objects (enumerated in the rows of the input matrix) and their specific rankings have not been shown in any of the figures yet, although it would be imperative to do so, as well. This train of thought led to eliminating the hiatus of visualization of object rankings. As it will be presented in the following section, we propose a method that arranges the parallel coordinates in bidirectional from the middle to the left and the right based on SRD ordering.

3. SRD-Based Ordering of Parallel Coordinates

The proposed algorithm orders the variables according to their respective SRD values, before the angle of deviation is calculated between each variable and the last axes from each direction so that the intersection are minimized. Therefore, the SRD method is described first, before the performance metrics and ordering algorithms are introduced.

3.1. Sum of Ranking Differences Method

The steps of the classical SRD method are described in Ref. [27]. Here, a small synopsis is given. The main building blocks of the algorithm are depicted in Figure 2. SRD is designed to analyze data stored in an X data matrix with $k = 1, \dots, N$ number of rows (objects) and $j = 1, \dots, n$ columns (variables), respectively.

Figure 2. Flowchart of the SRD algorithm.

The SRD algorithm eventually includes the transformation of the original $x_{k,j}$ variables, e.g., the normalization of the variables in the $[0, 1]$ range:

$$x'_{k,j} = \frac{x_{k,j} - \min(\mathbf{x}_j)}{\max(\mathbf{x}_j) - \min(\mathbf{x}_j)} \quad (1)$$

where the transformed data are denoted by $x'_{k,j}$, $\min(\mathbf{x}_j)$ stands for the minimum and $\max(\mathbf{x}_j)$ for the maximum of the j -th variable (represented by the j -th column of the $\mathbf{X}$ matrix).

Transformation is followed by the aggregation of variables to generate a gold standard in the absence of an ideal reference. The gold standard is represented by ρ_k that can be obtained by aggregating the rows of the $\mathbf{X}$ matrix

$$\rho_k = \text{aggregate}(\mathbf{x}_k) \quad (2)$$

Arithmetic mean, median, minimum and maximum functions are frequently employed for aggregation. With the arithmetic mean, the similarity to the reference is analyzed, whereas the median functions the same, though it is not sensitive to outliers. The employment of the arithmetic mean is based on the maximum likelihood principle, i.e., we can reasonably assume that the errors cancel each other. The selection of the maximum as the reference corresponds to defining the hypothetical best variable, joining the preferable features of all individual variables. It is generally used for correlation coefficients, explained variances, receiver operator characteristic curves, *etc.*, where the higher, the better it holds. The minimum is considered the logical conjunction, and the “inverse” of the maximum, preferably for residual errors, standard deviations, predicted error sum of squares, their cross-validated counterparts, *etc.*

As it is shown in Figure 2, the variables and the gold standard are ranked before their distances are calculated. We can calculate the sum of ranking differences (SRDs) by calculating the city block (Manhattan) distance between the rank of the gold standard and rank value of the data

$$d_{k,j} = |\text{rank}(x'_{k,j}) - \text{rank}(\rho_k)| \quad (3)$$

After summing each variable, the SRD values are obtained, which are only equal to Spearman's footrule, if no ties are present in the input data:

$$\text{SRD}_{0,j} = \sum_{k=1}^N |\text{rank}(x'_{k,j}) - \text{rank}(\rho_k)| \quad (4)$$

where $\text{SRD}_{0,j}$ denotes the SRD value of a variable in terms of the gold standard.

Lastly, the SRD values are normalized between 0 and 100 and the SRD values are plotted to determine the relationship between the variables and the gold standard.

3.2. SRD-Based Ordering for Reduction of Intersections

As the respective ranked values of an object are connected in a plot of parallel coordinates, the number of intersections presents the correlation of the rankings. It is better to visualize fewer intersections for better representations of the interrelation between variables. The rankings are incorporated to the axes, and the connection between two neighboring ones measures the correlation, or the similarity, between the rankings. Kendall's τ determines the number of moves required to make the two rankings the same [28]:

$$\tau_{j,m} = 1 - \frac{2w_{j,m}}{N(N-1)} \quad (5)$$

where $2w_{j,m}$ is the number of moves required to obtain the same j -th ranking from the m -th ranking, and $N(N-1)/2$ is the number of maximum steps when the change in the

rankings happen in pairs (one step causes one intersection twice, as both rankings retain the intersection) [28].

For parallel coordinates with N elements (ranks), the number of intersections can be calculated using Kendall's τ , or rank correlation by transforming Equation (5), as the number of moves can be considered the number of intersections:

$$c_{j,m} = \frac{(1 - \tau)(N(N - 1))}{4} \quad (6)$$

where $c_{j,m}$ denotes the number of intersections between rankings.

The visualization ordering can be performed in several ways, as we define the negative values of the horizontal axis as being mutually dissimilar to the positive axis, where the differences between the neighboring variables can be visualized. Kendall's τ [4] can determine how significantly they differ from each other, as it is a non-parametric measure of the association between two rankings.

The summation of intersections can perfectly depict how proficient a method of ordering is, which can be divided by the theoretical maximum number of intersections:

$$s = \frac{\sum_{l=1}^{n-2} c_l}{(n-2) \frac{N(N-1)}{2}} \quad (7)$$

where s denotes a percentage of intersections with regard to the maximum number of intersections, c_l represents the intersections in a section and $n - 2$ is the number of sections between the axes.

In the novel SRD-based parallel coordinates visualization technique, the positions of the axes are determined by the SRD value of each variable. Therefore, the distances between the axes are ununiform, providing information on how similar the variables from the viewpoint of the gold standard can be considered to be.

To provide a more detailed information, not only are the SRD values against the gold standard calculated, but also the SRDs of all variable pairs:

$$SRD_{i,j} = \sum_{k=1}^N |rank(x'_{k,i}) - rank(x'_{k,j})| \quad (8)$$

It has to be noted that the resultant $SRD_{i,j}$ values do not depend on the golden standard, as these distance values can be directly calculated independently of which gold standard is used, and there is no need for their recalculation, if the user defines another gold standard.

The proposed method evaluates the consistency of the gold-standard-based distances based on these $SRD_{i,j}$ values. If the two respective gold-standard-based SRD and the SRD of the two variables (one is defined as a reference, such as in Equation (8)) are constructed as a triangle, the angle of deviation between two SRD values can be calculated as:

$$\cos(\gamma_{j,m}) = \frac{SRD_{0,j}^2 + SRD_{0,m}^2 - SRD_{j,m}^2}{2SRD_{0,j}SRD_{0,m}}; j \neq m \quad (9)$$

where $\gamma_{j,m}$ denotes the angle between two gold standard-based SRD values, namely $SRD_{0,j}$ and $SRD_{0,m}$, while $SRD_{j,m}$ stands for the summed difference between the rankings of the j -th and m -th variables. Figure 3 represents the dissimilarity of the variables between two axes in terms of parallel coordinates visualization by measuring the angles. The method is described in Algorithm 1. As the mutual similarity between items is not necessarily revealed by the SRD method, the calculation of the deviation angle between the SRD values measures the alignment of the values and thus the differences in similarities if the triangle equality is satisfied.

Figure 3. Placement of the axes based on the angle of deviation calculated with the law of cosines. The angle of deviation between the SRD values (distance from the reference axis) of one axis and its neighboring axis can be determined by constructing a triangle with the SRD value (one angle as a reference) between the two axes as the third side. If the angle of deviation is considerably small, then the variables can be considered mutually similar.

In the proposed algorithm, first, the SRD values of the variables are determined before the angles between the SRD values are calculated by Equation (9) and the variables iteratively compared. On the x-axis, the normalized SRD values between 0 and 100 are plotted. The gold standard and the variable with the smallest SRD value are defined as the starting reference axes to the left- and right-hand sides, respectively. The angle of deviation between the variable and the reference determines the side on which the variable should be placed on. If the angle of deviation with regard to the variable and the reference axis on the left-hand side is less than the other, the variable is placed to the left-hand side and selected as the following reference and *vice versa*.

Algorithm 1 SRD-based visualization of parallel coordinates.

```

 $N \leftarrow \dim(\mathbf{X}', 1);$ 
 $n \leftarrow \dim(\mathbf{X}', 2);$ 
 $\rho \leftarrow \text{gen\_reference}(\mathbf{X}')$ 
 $\text{SRD} \leftarrow [0, \text{SRD}(\mathbf{X}', \text{ref})];$ 
 $\text{lft} \leftarrow 1;$ 
 $\text{rgh} \leftarrow 2;$ 
for ( $j = 3$  to  $n$ ) do  $\text{aleft} \leftarrow \text{srd\_angle}(\text{SRD}_{0,j}, \text{SRD}_{0,\text{lft}}, \text{SRD}_{\text{lft},j})$ 
 $\text{aright} \leftarrow \text{srd\_angle}(\text{SRD}_{0,j}, \text{SRD}_{0,\text{rgh}}, \text{SRD}_{\text{rgh},j})$ 
if ( $\text{aleft} < \text{aright}$ ) then  $\text{lft} = j; \text{place\_left}(\text{SRD}_{0,j});$ 
else  $\text{rgh} = j; \text{place\_right}(\text{SRD}_{0,j});$ 
end if
end for

```

A third possibility to determine differences between the variables is by calculating the disarray value using classical multidimensional scaling (MDS) [6] and the rank correlation. A similarity matrix is required that can be defined by the rank correlation coefficients of the variables, which is converted into dissimilarity by the equation $1 - \tau_{j,m}$ before being converted into a dissimilarity matrix. Then the data are converted into a one-dimensional distance matrix with MDS.

3.3. MATLAB and Octave Source Codes

The angle-based visualization of the SRD technique is provided in a script (.m extension). The functions that build up the visualization technique are also included in the toolbox. The scripts are only dependent on 'tiedrank' and 'corr' functions. The following functions are included in the toolbox:

- `parcoord(args)` for the visualization of parallel coordinates (in general);
- `SRD(args)` for the calculation of the gold standard-based SRD and SRD matrix;
- `srd_angle(args)` for the calculation of the angle of deviation between the j -th and m -th variables;
- `parcoord_angles(args)` for the placement of angle-based SRD in parallel coordinates;
- `pc_srd_demo.m` for a demo that presents the usage of the functions, creating the figures presented in this research.

All the functions require the gold standard and matrix of SRD value vectors to be included separately. The angle-based SRD visualization is also provided a compact function:

```
[tau, intersections] = parcoord_srd_angle(data, golden, axesnames, goldenname);
```

where 'data' is either raw or transformed, the 'golden' gold standard is either ideal or generated, 'axesnames' denotes the labels of the variables, and 'goldenname' stands for the label of the gold standard.

The 'io' and 'nan' packages are required for the demo and the functions to be functional in Octave.

4. Case Study of Greenhouse Gas Emissions

To introduce the capabilities of the SRD-based parallel coordinates, we employ the algorithm on the Climate Analysis Indicators Tool (CAIT) [29] database of Climate Watch, which contains a category type of aggregated greenhouse gases (all GHG) that includes carbon dioxide, methane, nitrous oxide and fluorinated gases. Carbon dioxide (CO₂) represents another emissions category.

Table 1 contains all emissions categories as well as their IDs and codes. All categories are provided in metric tons of CO₂ equivalent divided by the population. Indicators are presented in the following formulae: category code + 'e' + emissions type, where 'e' denotes the category and emissions separator.

Table 1. Category list. This table contains the ID as well as the name and code of the categories.

ID	Category Name	Category Code
1	Agriculture	AGR
2	Building	BLD
3	Bunker Fuels	BNK
4	Electricity/Heat	ELH
5	Energy	ENG
6	Fugitive Emissions	FUE
7	Industrial Processes	IND
8	Land-Use Change and Forestry	LCF
9	Manufacturing/Construction	MAN
10	Other Fuel Combustion	OFC
11	Total excluding LCF	TOT
12	Transportation	TRP
13	Waste	WAS

Moreover, the emissions data include the land use, change and forestry (LCF) indicator provided by the Food and Agriculture Organization of the United Nations [30] in their Food and Agriculture statistics (FAOSTAT) emissions database, as well as combustion data recorded by the Organization for Economic Co-Operation and Development (OECD) [31]. The economic indicators are requested from the World Bank Open Data database, namely,

the gross domestic product (GDP), population, rural population percentage, GDP growth and urban population growth rate [32]. The emissions and the GDP are divided by the population, which is removed afterwards. False entries with high proportion of missing data as well as countries with fewer than 500,000 inhabitants are excluded.

First, the categories characterizing the climate change are reduced: constants (all zeroes) as well as variables with low standard deviation are eliminated, and the variables are arranged in alphabetic order. The total number of nations and categories are 147 and 20 (excluding population), respectively.

Regarding the visualization of parallel coordinates, the red axis is considered to be the gold standard or origin axis. The quantiles (25%—yellow, 50%—magenta, 75%—blue, 100%—green) and the first decile (10%—red) of the ranks are colored to distinguish between their positions. The rank numbers for each variable are plotted on the y-axes, and the x-axis depends on the input data. In the case of Figure 4, the indicators are provided in alphabetical order for the purpose of visualization. The neighboring indicators are in an arbitrary ordering without regard to their relationships to each other; therefore, the minimization of the total crossings is not probable. This method yields 76,509 intersections for a ranking of 147 members. The quantiles change position with each step further from the origin; therefore, the ordering is inefficient. Especially between GDP, GDP growth, industry, all GHGs categories, and from the land use, change and forestry to the total emissions category, the intersections propose that these are non-similar neighbors. Even in this inefficient layout, the very similar variables (e.g., LCFEAllGHG and LCFECO₂) can be detected, as can reverse rankings (e.g., OFCAllGHG and RuralPOP), or very different ones (e.g., GDPGrowth and INDeAllGHG), as well

Figure 4. Ranked climate indices in alphabetical order in a parallel coordinate system. The legend presents the color of the first decile (D1) and the quantiles (Q1–Q4) for the ranks. The total number of intersections is 76,509.

The MDS-based analysis focuses on the lower-dimensional representation of the pairwise dissimilarity matrix. Kendall's τ presents similarities; therefore, the inversion yields a dissimilarity that can be used to calculate the MDS value of a variable. The method analyzes the similarity of one variable to all others.

The parallel visualization of MDS yields 53,591 intersections, and both negative and positive dissimilarities. The MDS-based ordering is depicted in Figure 5. On the left-hand side, various categories are placed in the vicinity of each other, creating a cluster whose quantiles remain in a similar position, reducing the number of intersections.

Figure 5. Ranked climate indices based on the dissimilarity metrics of multidimensional scaling in a parallel coordinate system. Color codes are explained in Figure 4. The total number of intersections is 53,591.

The grouping suggests that these retain a close connection to each other, primarily energy-related variables and GDP, total emissions, industry and transportation categories. This result might indicate how closely the sectors relate to each other. As the figure proceeds towards the positive limits of the x-axis, the intersections between the categories become more and more convoluted. The distance between the categories is either minute or enormous. The introduction of circular economy may depend on the sectors that are distant from each other. The waste, other fuel sources, industry CO₂, and growth indicators may be the key to starting the transition to a circular economy, as these sectors are relatively “independent” of each other. Waste is a consequence of human consumption, and changing the recycling industry barely affects the products themselves, as the core materials rarely change. On some occasions, the waste can be decommissioned, e.g., communal waste can be burned or fermented to methane.

To depict the SRD-based parallel coordinates, we performed SRD with the total of all GHGs emissions category as the gold standard before sorting the categories based on the SRD value and drafting Figure 6. The correlations gradually diminish as the normalized SRD values increase. Above 35, the number of intersections in the rankings of the variables escalates. The proximity of the axes shows the close degree of similarity based on their distance from the ideal ranking, although they are not necessarily mutually similar. The total number of intersections is 53,591.

Beginning from the left-hand side, the number of intersections are considerable in the range of 92 and 66, and after that, follows a group of variables that can be mutually similar. On the right-hand side, the quantiles are swapped in terms of position, although there are exceptions, as the SRD values are not 100 (not perfectly inverse). Given that the total emissions indicator acts as the reference, it is placed at the zero SRD value mark. The following indicators are the total greenhouse gas emitted by the energy sector, as well as the CO₂ emissions of total and energy sectors. Electricity, transportation and manufacturing are grouped, expressing a connection to the GDP indicator. The industry, building and bunker fuels are grouped as well. CO₂ emitted from industrial activity and waste are connected as well. Other fuel sources, agriculture and GDP growth are considered outliers. The land use, change and forestry indicators, urban growth, and rural population are ranked in reverse order to the gold standard.

Figure 6. Ranked climate indices based on the SRD in a parallel coordinate system. Color codes are explained in Figure 4. The total number of intersections is 52,200.

In Figure 7, the angle-based ordering of the SRD values is depicted. The total of the GHG emissions category axis is denoted with red dashes at the x-axis so that the gold standard is quickly identifiable. On the left-hand side, the position of the quantiles is reversed, indicating that the left-hand side of the figure contains reversely ranked categories, whereas, on the right-hand side, the quantiles remain within the same range, with a negligible degree of fluctuation. It is only reasonable that the randomly and reversely ranked categories are placed in the reverse direction, as, if one side is filled with similarly correlating variables, the angles between the remaining non-correlating variables may be less than each other than to the other side. The total crossings can be further reduced (to 52,200) if the asymmetric distances are plotted to the right and the left from the gold standard, whose SRD value is 0. The total number of intersections is determined as 52,200, which is 24,309 less than in the alphabet-based approach.

Figure 7. Ranked climate indices based on the angles between neighboring SRD values in a parallel coordinate system. Color codes are explained in Figure 4. Number of total crossings is 50,252.

In the vicinity of the gold standard in Figure 7, all greenhouse gas emitted by the energy sector, CO₂ emissions of transportation, manufacturing, and GDP are followed by industry bunker fuels and waste, all of which are mutually similar. Such a relation implicitly suggests the significant contribution of the heavy industry to the total emissions. Because of that, the economic models of nations should be shifted towards circular sustainable economies, as circular economies may provide such benefits as mitigation, reduced costs of manufacturing, recycling and efficient waste management [33–35]. On the left-hand side, the total CO₂ emitted by energy and electricity, respectively, are similar variables. It is in the best interest of mankind to reduce CO₂ emissions due to global warming and health issues that are caused by pollution and increasing temperature. The most polluting sector of all is the energy sector, where fossil fuels are used for on-site and industrial energy generation. An example is Germany, which has made significant progress to introduce renewable energy to the sector in an attempt to reduce CO₂ emissions. A net-positive change occurred, whereas the growth of the industry has not been altered, as it has provided an immense amount of jobs [36]. The only problem with such energy is its storage, which is not yet cost effective at the present [37].

For the reduction in emissions from bunker fuels, shipping could be optimized based on the operational and charter cost of vessels, port handling and fees, fuel consumption, and inventory management [38]. As shipping provides the most merchandise to the service sector, unjustified consumption poses health risks to those living in cities with harbors [39] and steadily increases the polluting agents in the air.

A strong manufacturing industry is a prerequisite of a profitable economy, but without a circular one, the processes may cause a detrimental effect on the environment. As one of the most polluting countries, China has been identified as being capable of reducing 78 % of its CO₂ emissions if proper actions are taken [40]. For green technology to progress, optimization and perseverance are better driving forces than strongly profit-oriented approaches masked by greenwashing.

The left-hand side consists of indicators that may prove useful in determining either the service or the agricultural sector, but are not essential to the industry sector. Industry is a requirement for the service sector, and agriculture is boosted with manufactured machines. The GDP growth is reversely ranked along with agriculture, yet it is clear that the two magnitudes barely correlate. In many developing countries, the agriculture is based on subsistence farming, and a high proportion of the population may solely feed themselves and their families. The population of these countries live below the poverty line and can only contribute marginal profits to the “big” economy [41]. The figure contains 50,252 intersections, 1948 less than the classical SRD-based ordering.

Several sectors of the economy are provided in the analysis. The sectors are in exponential growth in general, which does not consider sustainability and circular economy. Unfortunately, the phenomenon of greenwashing [42] has arisen under the current regulations. Greenwashing companies are not environmentally friendly, but aim to obtain more profit. These organizations may exaggerate emissions reductions and mislead consumers. Huge contributors of greenhouse gases are the building and construction sectors. The problems and solutions are being thoroughly researched [43]. To reduce emissions, the introduction of biofuel to heavy machinery and energy generation is recommended. The electrification of processes and the application of carbon capture technologies is also proposed. Through embracing the recommended solutions, the company policies become ever-more aligned with the Paris Agreement for both short- and long-term periods to avoid greenwashing. An example of greener building is passive houses. The maintenance cost in terms of the environmental footprint of office buildings is now reduced [44], and passive houses are under strict regulation of heat insulation and electricity consumption [45]. The actual manufacturing emissions of the required materials for passive buildings may be overlooked [46]. However, from the CO₂ emissions of the buildings category, the correlation steeply decreases.

5. Discussion

As mentioned above, the ordering of the variables is a major factor that reduces the number of intersections. It is not recommended to visualize the parallel coordinates based on an arbitrary ordering, such as the alphabet, as the number of intersections is large, which is 76,509 in this exact example. Each ordering method reduces the number of intersections significantly (each by about 25,000), proving that the ordering is a requirement to good visualization.

Although the angle-based ordering algorithm provided fewer intersections than the multidimensional scaling-based method, the gain is only marginal. However, the hierarchical ordering-based multidimensional scaling was not compared to SRD. We compared a similar rank correlation-based multidimensional scaling technique to the novel SRD-based parallel coordinates algorithm; the multidimensional scaling performs slightly worse. It is important to note that another analysis may not yield such enormous or minute distances between the axes. By dropping the fixed-position of the axes, the multidimensional scaling-based method may reduce the resemblance of the figure and provide information on the relationship between the axes.

The inclusion of the SRD method in the visualization of the parallel coordinates provided a satisfactory reduction in the number of intersections, which was reduced and compared to multidimensional scaling. “Dynamics Visualization based on Parallel Coordinates” (DYVIPAC) focuses on a set of problems and not on general use. However, the SRD angle method-based visualization alone is not capable of catching complex biological or mathematical problems, or dynamics analyses. It is a technique that focuses on the correlation and relationship between the variables.

The parallel coordinates technique is widely used, e.g., in ecology, where a complete framework was created to ease the visualization and clarification of water quality as well as invasive species proliferation. The visualization is interactive, and the framework provides features, such as line coloring, data filtering, and highlighting as well as tables. In the SRD-based parallel coordinates visualization toolbox, we provide a general SRD-based visualization with angle-based ordering that can be further extended in MATLAB or Octave, depending on the requirements of the analyst.

6. Conclusions

The visualization of parallel coordinates in itself is extended as the representation of the sum of ranking differences (SRD); the order of the variables is calculated by the SRD method. The SRD and the parallel coordinates technique are modified by placing the variables not only on the right-hand side of the gold standard, but on the left-hand side as well, thereby increasing the information content of the visualized relationship system (i.e., showing that variables adjacent to each other are similar).

We developed a method for classifying the arrangement according to SRD as a distance norm for triangular inequality (and the cosine theorem). The variables are iteratively placed based on the angle of deviation between the selected side reference variables on left- and right-hand side. The parallel coordinates-based visualization technique is preferably applicable, as it significantly reduces the quantity of the intersections in the figure.

The algorithm inherits the limitations of the parallel coordinates, e.g., negative correlations can be easily overestimated visually, and the hardships of reading the visualization may increase when adding a significant amount of variables or objects. Non-linear correlations may not be appropriately interpreted in the original method, and this requires further research concerning the visualization of SRD values. Although the original parallel coordinates method suffers from lack of proper ordering, it is solved by applying the SRD algorithm. However, these limitations are intertwined with advantages. A great deal of information is presented by the visualized data that can be evaluated in a short matter of time, and it is easy to learn. The method excels at providing information on the correlation between the rankings and detects outliers in the rankings, making it a candidate for efficient cluster analysis.

Another limitation is that only two mutually similar groups can be established, due to the shortage of axes. In this case, the probability values being mutually similar may pave the way for more information on the exact relationships between the variables. It is also important to note that parallel coordinates may be used besides other visualization techniques.

As for future research, the determination of non-linear correlations will be built in the SRD framework; moreover, the behavior of aggregation is to be researched, where the properties of the distributions of the ranking distances during data fusion are to be determined.

Author Contributions: Á.I. wrote the original draft, and prepared the MATLAB and Octave codes. K.H. came up with and conceptualized the idea as well as writing and proofreading the article. J.A. conceptualized the idea, developed the methodology, supervised and corrected the code and article, as well as acquired funding. All authors have read and agreed to the published version of the manuscript.

Funding: The contributions of János Abonyi and Ádám Ipkovich to this research were funded by the National Laboratory for Climate Change (NKFIH-471-3/2021), and that of Károly Héberger by the “Development of soft sensor models based on *in vitro* experiments and creation of related metrics”, project no. OTKA 134260 supported by the Ministry of Innovation and Technology of Hungary from the National Research, Development and Innovation Fund, financed under the K type funding scheme.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The MATLAB and Octave codes described in Section 3.3 and the database (co2_data.xlsx) are accessible at <https://github.com/abonyilab/parcoord> (accessed on 31 October 2021), written by János Abonyi and Ádám Ipkovich, 10 October 2021.

Conflicts of Interest: The authors declare that they have no known competing financial interest or personal relationships that could have appeared to influence the work reported in this paper.

References

1. Wegman, E. Hyperdimensional data analysis using parallel coordinates. *J. Am. Stat. Assoc.* **1990**, *85*, 664–675. [\[CrossRef\]](#)
2. Inselberg, A. The plane with parallel coordinates. *Vis. Comput.* **1985**, *1*, 69–91. [\[CrossRef\]](#)
3. Johansson, J.; Forsell, C. Evaluation of Parallel Coordinates: Overview, Categorization and Guidelines for Future Research. *IEEE Trans. Vis. Comput. Graph.* **2016**, *22*, 579–588. [\[CrossRef\]](#)
4. Kendall, M. *Rank Correlation Methods*; Oxford University Press: Oxford, UK, 1990.
5. Zhou, Z.; Ye, Z.; Yu, J.; Chen, W. Cluster-aware arrangement of the parallel coordinate plots. *J. Vis. Lang. Comput.* **2018**, *46*, 43–52. [\[CrossRef\]](#)
6. Seber, G.A.F.E. *Multivariate Observations*; Wiley Series in Probability and Statistics; John Wiley & Sons: Hoboken, NJ, USA, 1984; pp. 139–146. [\[CrossRef\]](#)
7. Ellis, G.; Dix, A. Enabling Automatic Clutter Reduction in Parallel Coordinate Plots. *IEEE Trans. Vis. Comput. Graph.* **2006**, *12*, 717–724. [\[CrossRef\]](#)
8. Héberger, K. Sum of ranking differences compares methods or models fairly. *TrAC—Trends Anal. Chem.* **2010**, *29*, 101–109. [\[CrossRef\]](#)
9. Héberger, K.; Kollár-Hunek, K. Sum of ranking differences for method discrimination and its validation: Comparison of ranks with random numbers. *J. Chemom.* **2011**, *25*, 151–158. [\[CrossRef\]](#)
10. Héberger, K. Method and Model Comparison by Sum of Ranking differences in Cases of Repeated Observations (Ties). *Chemom. Intell. Lab. Syst.* **2013**, *127*, 139–146. [\[CrossRef\]](#)
11. Vathy-Fogarassy, Á.; Abonyi, J. *Graph-Based Clustering and Data Visualization Algorithms*; Springer: Berlin/Heidelberg, Germany, 2013.
12. Tóth, G.; Amari-Amir, S. Seriation, the method out of a chemist’s mind. *J. Chemom.* **2018**, *32*, e2995. [\[CrossRef\]](#)
13. Dörgő, G.; Sebestyén, V.; Abonyi, J. Evaluating the interconnectedness of the sustainable development goals based on the causality analysis of sustainability indicators. *Sustainability* **2018**, *10*, 3766. [\[CrossRef\]](#)
14. Oyedele, O.F. Extension of biplot methodology to multivariate regression analysis. *J. Appl. Stat.* **2020**, *48*, 1816–1832. [\[CrossRef\]](#)
15. Nie, M.; Meng, L.; Chen, X.; Hu, X.; Li, L.; Yuan, L.; Shi, W. Tuning parameter identification for variable selection algorithm using the sum of ranking differences algorithm. *J. Chemom.* **2019**, *33*, e3113. [\[CrossRef\]](#)

16. Chen, X.; Xu, Y.; Meng, L.; Chen, X.; Yuan, L.; Cai, Q.; Shi, W.; Huang, G. Non-parametric partial least squares–discriminant analysis model based on sum of ranking difference algorithm for tea grade identification using electronic tongue data. *Sens. Actuators B Chem.* **2020**, *311*, 127924. [\[CrossRef\]](#)
17. Rácz, A.; Bajusz, D.; Héberger, K. Consistency of QSAR models: Correct split of training and test sets, ranking of models and performance parameters. *SAR QSAR Environ. Res.* **2015**, *26*, 1–18. [\[CrossRef\]](#)
18. Roy, K.; Mitra, I.; Ojha, P.; Kar, S.; Das, R.; Kabir, H. Introduction of rm2(rank) metric incorporating rank-order predictions as an additional tool for validation of QSAR/QSPR models. *Chemom. Intell. Lab. Syst.* **2012**, *118*, 200–210. [j.chemolab.2012.06.004. \[CrossRef\]](#)
19. West, C.; Khalikova, M.A.; Lesellier, E.; Héberger, K. Sum of ranking differences to rank stationary phases used in packed column supercritical fluid chromatography. *J. Chromatogr. A* **2015**, *1409*, 241–250. [\[CrossRef\]](#)
20. Nowik, W.; Héron, S.; Bonose, M.; Tchaplá, A. Separation system suitability (3S): A new criterion of chromatogram classification in HPLC based on cross-evaluation of separation capacity/peak symmetry and its application to complex mixtures of anthraquinones. *Analyst* **2013**, *138*, 5801–5810. [\[CrossRef\]](#)
21. Vastag, G.; Apostolov, S.; Perišić-Janjić, N.; Matijević, B. Multivariate analysis of chromatographic retention data and lipophilicity of phenylacetamide derivatives. *Anal. Chim. Acta* **2013**, *767*, 44–49. [\[CrossRef\]](#) [\[PubMed\]](#)
22. Andrić, F.; Bajusz, D.; Rácz, A.; Šegan, S.; Héberger, K. Multivariate assessment of lipophilicity scales—Computational and reversed phase thin-layer chromatographic indices. *J. Pharm. Biomed. Anal.* **2016**, *127*, 81–93; [\[CrossRef\]](#)
23. Brownfield, B.; Kalivas, J.H. Consensus Outlier Detection Using Sum of Ranking Differences of Common and New Outlier Measures Without Tuning Parameter Selections. *Anal. Chem.* **2017**, *89*, 5087–5094. [\[CrossRef\]](#)
24. Sziklai, B.R.; Héberger, K. Apportionment and districting by Sum of Ranking Differences. *PLoS ONE* **2020**, *15*, e0229209. [\[CrossRef\]](#) [\[PubMed\]](#)
25. Sziklai, B.R. Ranking institutions within a discipline: The steep mountain of academic excellence. *J. Inf.* **2021**, *15*, 101133. [\[CrossRef\]](#)
26. West, C. Caroline West Statistics for Analysts Who Hate Statistics, Part VII: Sum of Ranking Differences (SRD). *LCGC N. Am.* **2018**, *36*, 2–6.
27. Bajusz, D.; Rácz, A.; Héberger, K. Why is Tanimoto index an appropriate choice for fingerprint-based similarity calculations? *J. Cheminform.* **2015**, *7*, 20. [\[CrossRef\]](#) [\[PubMed\]](#)
28. Griffin, H.D. Graphic Computation of Tau as a Coefficient of Disarray. *J. Am. Stat. Assoc.* **1958**, *53*, 441–447. [\[CrossRef\]](#)
29. Climate Watch. *GHG Emissions*; World Resources Institute: Washington, DC, USA, 2020.
30. Food and Agriculture Organization. *Food and Agriculture Statistics (FAOSTAT) Emissions Database*; United Nations: Rome, Italy, 2020.
31. Dumont, J.C.; Zurn, P. Organisation for Economic Co-operation and Development (OECD) & International Energy Agency (IEA). *In CO₂ Emissions from Fuel Combustion*; OECD: Paris, France, 2019.
32. World Bank. *GDP, Population, Urban Population Growth, Rural Population, GDP per Capita Growth, Surface Area Indicators*; World Bank Group Archives: Washington, DC, USA, 2021.
33. Gallego-Schmid, A.; Chen, H.M.; Sharmina, M.; Mendoza, J.M.F. Links between circular economy and climate change mitigation in the built environment. *J. Clean. Prod.* **2020**, *260*, 121115. [\[CrossRef\]](#)
34. Durán-Romero, G.; López, A.M.; Beliaeva, T.; Ferasso, M.; Garonne, C.; Jones, P. Bridging the gap between circular economy and climate change mitigation policies through eco-innovations and Quintuple Helix Model. *Technol. Forecast. Soc. Chang.* **2020**, *160*, 120246. [\[CrossRef\]](#)
35. Lausset, C.; Cherubini, F.; Oreggioni, G.D.; del Alamo Serrano, G.; Becidan, M.; Hu, X.; Rørstad, P.K.; Strømman, A.H. Norwegian Waste-to-Energy: Climate change, circular economy and carbon capture and storage. *Resour. Conserv. Recycl.* **2017**, *126*, 50–61. [\[CrossRef\]](#)
36. Hinrichs-Rahlwes, R. Renewable energy: Paving the way towards sustainable energy security: Lessons learnt from Germany. *Renew. Energy* **2013**, *49*, 10–14. [\[CrossRef\]](#)
37. Trainer, T. Some problems in storing renewable energy. *Energy Policy* **2017**, *110*, 386–393. [\[CrossRef\]](#)
38. Pasha, J.; Dulebenets, M.; Kavoosi, M.; Abioye, O.; Theophilus, O.; Wang, H.; Kampmann, R.; Guo, W. Holistic tactical-level planning in liner shipping: An exact optimization approach. *J. Shipp. Trade* **2020**, *5*, 8. [\[CrossRef\]](#)
39. Sofiev, M.; Winebrake, J.; Johansson, L.; Carr, E.; Prank, M.; Soares, J.; Vira, J.; Kouznetsov, R.; Jalkanen, J.P.; Corbett, J. Cleaner fuels for ships provide public health benefits with climate tradeoffs. *Nat. Commun.* **2018**, *9*, 406. [\[CrossRef\]](#)
40. Yang, J.; Cheng, J.; Huang, S. CO₂ emissions performance and reduction potential in China's manufacturing industry: A multi-hierarchy meta-frontier approach. *J. Clean. Prod.* **2020**, *255*, 120226. [\[CrossRef\]](#)
41. Siphesihle, Q.; Lelethu, M. Factors affecting subsistence farming in rural areas of nyandeni local municipality in the Eastern Cape Province. *S. Afr. J. Agric. Ext.* **2020**, *48*, 92–105. [\[CrossRef\]](#)
42. Netto, S.; Sobral, M.; Ribeiro, A.; Soares, G. Concepts and forms of greenwashing: A systematic review. *Environ. Sci. Eur.* **2020**, *32*, 19. [\[CrossRef\]](#)
43. Johnsson, F.; Karlsson, I.; Rootzén, J.; Ahlback, A.; Gustavsson, M. The framing of a sustainable development goals assessment in decarbonizing the construction industry—Avoiding “Greenwashing”. *Renew. Sustain. Energy Rev.* **2020**, *131*, 110029. [\[CrossRef\]](#) [\[PubMed\]](#)

-
44. Airaksinen, M.; Matilainen, P. A Carbon Footprint of an Office Building. *Energies* **2011**, *4*, 1197. [[CrossRef](#)]
 45. Mihai, M.; Tanasiev, V.; Dinca, C.; Badea, A.; Vidu, R. Passive house analysis in terms of energy performance. *Energy Build.* **2017**, *144*, 74–86. [[CrossRef](#)]
 46. Stephan, A.; Crawford, R.H.; de Myttenaere, K. A comprehensive assessment of the life cycle energy demand of passive houses. *Appl. Energy* **2013**, *112*, 23–34. [[CrossRef](#)]