Improving with Metaheuristics the Item Selection in Parallel Coordinates Plot

David Cordero-Machuca, Juan-Fernando Lima, and Marcos Orellana(✉)

Universidad del Azuay, Cuenca 010204, Ecuador
{david.corderom,flima,marore}@uazuay.edu.ec

Abstract. Data visualization is one of the most powerful techniques to analyze and obtain reliable results based on the displayed outputs since it allows humans to improve decision-making by visually analyzing data behavior. Nevertheless, it could be disrupted by high data amounts in the visualization, as is the case with Parallel Coordinate Plot (PCP), where data behavior and associations of volumes of data are difficult to identify. This paper aims to reduce this issue with PCP and take advantage of metaheuristics for optimization problems through a Simulated Annealing (SA) algorithm. The proposed method was developed and tested using air pollution and meteorological variables. The obtained results presented a reduction in data volume, thus helping represent the most relevant data for the final user.

Keywords: Metaheuristic · Parallel coordinates plot · Filtering

1 Introduction

In recent years, the large volumes of daily generated data have opened the path for exploring and analyzing big data [13]. Big data allows using historical data to identify patterns useful for recent events, enabling a data analyst to make quick decisions based on big data analysis [1]. On the other hand, the constant growth of data has certain drawbacks, such as an increase in the algorithmic complexity required to process large amounts of data. This issue has pressured companies and academics to develop new systems and methodologies while generating innovative results [4].

Metaheuristic algorithms are processes that generate high-quality solutions by choosing a satisfactory solution through an iterative process with clear guidelines from a pool of possible solutions. They are implemented to avoid high algorithmic complexity. These algorithms include high-level and low-level procedures, such as a simple local search or a construction method [17] Metaheuristic algorithms are divided into two groups: the first group is single point-based methods, where each search space is gradually developed and includes some of the most known algorithms: Simulated Annealing (SA), trajectory/local search methods, tabu search, and simple evolutionary strategies [18]. The second group is population-based algorithms which include multiple trial points in the search

J. Herrera-Tapia et al. (Eds.): TICEC 2022, CCIS 1648, pp. 186–200, 2022.
https://doi.org/10.1007/978-3-031-18272-3_13

space. Their results depend on their collective behavior, such as ant colony optimization, particle swarm optimization, and evolutionary strategies [18].

Data visualization is one of the most common techniques used to exploit big data and includes filtering irrelevant data, detecting multi-variable relationships, interacting with data representation, and observing data subsets in detail [9], making it easy to identify patterns, and is essential for information analysts [11]. Although previously, data visualization techniques were empirical and used to give a general idea of data representation, the difficulty of producing a quick analysis has required the generation of diverse data analysis techniques [5]. A lack of knowledge about new techniques, or the habit of using standard data visualization techniques, generates limitations for big data representation [22]. A clear example of the general lack of knowledge of visualization techniques is the Parallel Coordinates Plot (PCP), which is relatively unknown outside of the data visualization community [14,25]. This technique has some advantages, such as multi-variable relationships in a small space, data patterns between variables, and manipulation of data visualization by the user [10,12]. However, its main disadvantage is the difficulty of interpreting relationships between variables when handling big data [10,15]. Hence, to solve the data dispersion issue, Albazzaz and Wang [2] propose eliminating abnormal data and reducing the dimensions involved. Thus, we propose a metaheuristic algorithm that uses SA to show high levels of data concentration, based on various local results, to solve the disadvantage of PCP previously discussed.

The remaining paper is organized as follows: Sect. 2 presents the related work. Problem Formulation is in Sect. 3 which is divided into three stages: i) The data binning discretization, ii) The generation of alternative solutions, and iii) The optimal solution. Section 4 contains the results; and finally, Sect. 5 the conclusions and future works are presented.

2 Related Works

Data generally presents problems such as missing values, mixed formats, replicated entries, or lack of integrity rules [24]. It is important to apply preprocessing data techniques as the quality of results depends on cleaning the data of issues like those previously mentioned. In the field of data visualization, it is estimated that data scientists spend more than 50% of their time preparing data to be analyzed [24]. To deal with the PCP data dispersion problem, it is necessary to only reduce the data to the one relevant to the study. Meera and Sundar [19] propose a feature selection to reduce processing time and extract data from a database with big data, as it is a Hybrid method between Particle Swarm Optimization and Grammatical Evolution (PSO-GE).

Although PCP is a great advantage in representing multi-variables, its efficiency decreases when huge amounts of data are manipulated, thus presenting an overlapping issue. In order to reduce PCP representation complexity and make data easy to read, Berthold and Lawrence [6] apply fuzzy rules and even delimit the dimensions to be used to three, up to twenty. Also, variable selection reduces the amount of data processed. A good practice is establishing conditions and observations that are applied before analyzing and selecting variables [27].

Optimization problems can take advantage of metaheuristics. Abedinia et al. [21] propose a Shark Smell Optimization metaheuristic algorithm simulating shark behavior. The algorithm selects the best solution or "prey" by detecting higher odor concentration and forecasting the levels of solar energy by linking atmospheric components through a neural network.

Another research related to climate change, including metaheuristic algorithms and artificial intelligence tools, is the study by Dehghani et al. [8]. This study finds relationships between variables and data behavior and employs them to predict groundwater level behavior through a hybrid model on climatic variables. On the other hand, the analysis of climate variables does not exclusively use metaheuristics but also other artificial intelligence techniques. To find the best policy for drought regeneration systems, Mumtaz et al. [3] proposed the use of a Kernel Ridge Regression (KRR) model to split data and posed wet and dry scenarios, the Multivariate Empirical Mode Decomposition (MEMD) method to delimit multivariable climate indices, a Simulated Annealing (SA) model to define the most appropriate decomposed Intrinsic Mode Functions (IMFs) for the training period and feature selection strategy, and Random Forest (RF) to make decisions on a forecasting model. Another example of metaheuristics use is the one presented by Mohamed [20], where the algorithm seeks a relevant data subset by applying three different metaheuristic techniques: Particle Swarm Optimization (PSO), Cuckoo Search (CS), and Artificial Bee Colony (ABC).

Moreover, an example of the adaptability of metaheuristic algorithms to any optimization problem is the proposal by Bahadir and Serdar [16]. They employ SA to solve a p-median problem using a probabilistic metaheuristic and find the best threshold value for bi-level segmentation of gray-scale images, detecting optimal contour for edge-based images. In a different usage context, SA is used to find the optimal neighbor based on the power flow characteristics, within a real power system of nonlinear order and with a large combinatorial problem [26].

As observed, most of the studies implemented metaheuristics for global optimization. For this reason, our proposal contemplates using SA to define a rule for filtering relevant data in each selected variable. Thus, we use atmospheric variables, resulting in an understandable example of PCP optimization after filtering relevant data.

3 Problem Formulation

PCP is a visualization technique representing the relationships among related variables on the same graph. Usually, human intervention over the visualization graphs [7] results in data dispersion problems on PCP. These data dispersion issues highly increase the difficulty of interpreting a PCP, frequently becoming impractical to read. For this reason, this proposal aims to limit the amount of data presented by applying a SA metaheuristic algorithm. This section is divided into three parts: i) Visualization issue on PCP, ii) PCP visualization solution, and iii) Threats to validity.

3.1 Visualization Issue on PCP

To demonstrate the data dispersion issues, the analyzed dataset is managed by the municipality of Cuenca, Ecuador, which collects the atmospheric data measured in the city. A subset of this data, air pollution in the month of July 2018, has been chosen based on its completeness compared to other subsets. It includes five variables of air pollution: Carbon Monoxide (CO), Ozone (O_3), Nitrogen Dioxide (NO_2), Particulate Matter ($PM_{2.5}$), Sulfur Dioxide (SO_2), and a meteorological variable named air temperature ($°C$).

Then, preprocessing techniques were applied to the data. Table 1 presents the value range of variables, considering that the normality test failed for all of them. The PCP shown in Fig. 1 illustrates the selection of 3909 records belonging to the raw data for the month of July.

Table 1. Descriptive statistics of variables used for creating the visualization.

	Temperature ($°C$)	O_3 (ug/m^3)	CO (mg/m^3)	NO_2 (mg/m^3)	SO_2 (mg/m^3)	$PM_{2.5}$ (mg/m^3)
Min	5.30	4.74	0.25	0.0003	7.77	0.26
25%	12.20	11.40	0.55	5.19	9.95	5.29
Median	14.00	20.34	0.67	10.00	11.65	7.30
Mean	14.13	22.03	0.73	11.77	14.73	8.48
75%	16.00	30.44	0.83	16.23	15.87	9.82
Max	21.60	71.45	2.47	75.89	86.63	96.16

3.2 PCP Metaheuristic Solution

The proposed algorithm can filter records, thus reducing data presented to the most significant ones, to plot a clear image to show data behavior and relationships between variables. To aid human judgment, our proposed method aims to find the most relevant data in each variable based on user selection. Initially, each variable has the same amount of selected records. However, the data of

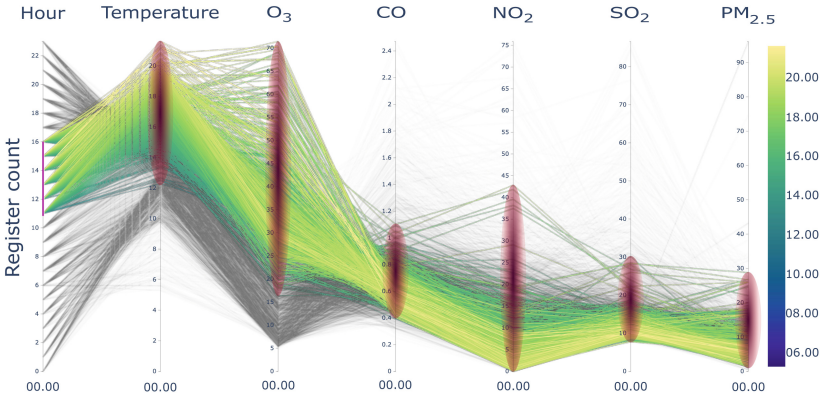


Fig. 1. PCP issue from July 2018.

a variable can be relevant in a different frequency and amplitude. The appropriate data selection method consists of three phases: i) The data distribution discretization phase receives prepossessed data as input, selects the ideal bin size, and detects the group of bars in the chosen variable. Then, ii) the alternative solutions phase calculates the minimum and maximum data from the selected dataset after the arithmetic means of each bin are calculated to decide if the current mean is unique; this value is added to the possible solutions array (threshold array). Finally, iii) the optimal solution phase chooses the better solution, the closest value to the frequency and amplitude conditions. The last two phases employ the SA algorithm, which has been divided into two algorithms for a better understanding: Algorithm 1 is responsible for finding alternative solutions, and Algorithm 2 is responsible for finding the optimal solution, obtaining between the two the SA algorithm and optimal threshold value.

To observe the data used in this study, Fig. 2 illustrates data for each variable in blue bars before it represents bins, and Fig. 3 depicts the dispersion of the raw data for the $PM_{2.5}$ variable.

Data Binning Discretization. Histograms are a type of graph that permit finding anomalies in data, such as dispersion, quantity counts of a bar (height or frequency) and the interval of a bar (width or amplitude). For this reason, histograms must be refined additionally for data preprocessing, as can be observed in Fig. 3.

Due to the dispersion presented was necessary to find a way to approach the histograms, one possibility being the division of the data into bins. There are different techniques to get a proper bin size, such as the equation of Sturge, normal of Scott, rule of Rice, and the Freedman-Diaconis equation [23]. To get a smoothness histogram, the Freedman-Diaconis equation was used in our dataset, as shown in Fig. 4 where the raw data of $PM_{2.5}$ are represented in Fig. 4a, and the division of that data into bins is represented in Fig. 4b.

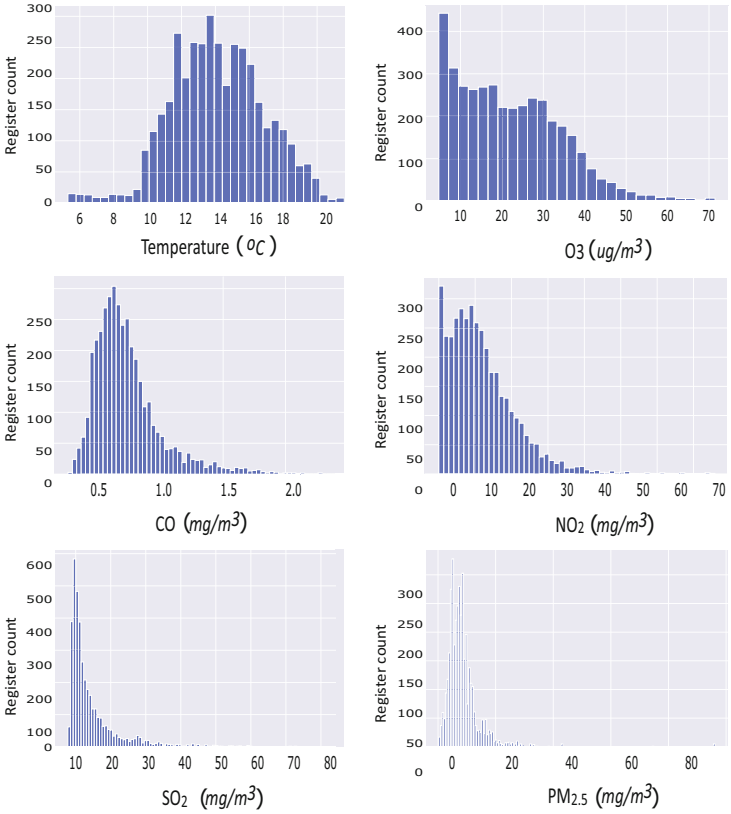


Fig. 2. Frequency and amplitude of selected variables. (Color figure online)

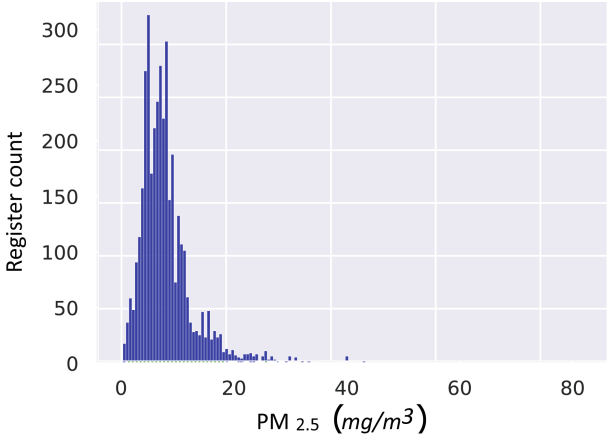
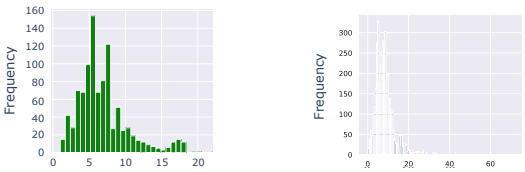


Fig. 3. $PM_{2.5}$ data dispersion.

Sahann et al. [23] show that the number of bins directly influences the distribution histogram. However, the same authors emphasize that the number of bins reaches a limit where no more bins can be added as the error rate stops decreasing. Based on the tests, the Freedman-Diaconis equation was chosen because it was balanced and sticks to the best bin size division for ideal results.



(a) Raw data of ($PM_{2.5}$). (b) Bins data of ($PM_{2.5}$).

Fig. 4. Use of Freedman-Diaconis equation into ($PM_{2.5}$) variable data.

Once the data has been divided into bins, the bin groups (a sequence of bins) and their boundaries must be identified. The proposed algorithm went through the generated bins sequentially to determine the formed data groups. If the height of the current bar was more significant than or equal to the value of the Filtered Amount Criteria ()FAC), the bar was added to the current bin group (which is named in sequential order from zero to n). On the other hand, if the current bar register count (height) was not more significant than the value of Eq. 1, the bar was not added and the new bin group was defined. This process was repeated until all the bars in the histogram of each chosen variable were covered, producing an algorithmic complexity of $O(n^2)$.

$$FAC = \frac{\sum_{i=1}^{m} h_i}{(x_{max} - x_{min})} \tag{1}$$

The mentioned equation was applied to each selected variable, where the height summation is repeated from $i = 1$ to the last bin height h_i of the actual variable. m represents the total register count for the actual variable, x_{max} represents the maximum record register, and x_{min} represents the minimum record.

Generation of Alternative Solutions. After dividing the data into bins and detecting the number of groups, the most relevant data was filtered by the two by the two algorithms representing the SA algorithm. Algorithm 1 defines a threshold value, where each record that exceeds this value is considered relevant data within the dataset in the histogram of each variable. Algorithm 1 explores the average values of each bin group for each variable. To reduce the complexity of the algorithm, initial control parameters are established, such as the possible solutions, which must be unique, and how many bin groups the algorithm can go through ("*nNeighbours*"). For our meteorological data, the value of "*nNeighbours*" was established as five neighbors since this value achieved the best test results.

In Algorithm 2, the input data refers to the list of variables to be analyzed $varList$, the current variable on analysis is $curVarible$, the complete data frame is $fulData$, the selected data frame is $selData$, the number of established neighbors is $nNeighbors$, and the number of bins from the actual variable is $nBinGroups$. As output data, the algorithm returns a vector of possible solutions $posibleSolutions$ and a vector of possible solutions expressed as a percentage $posibleSolutionsPct$.

Algorithm 1. Alternative solutions.

Require: $varList, curVarible, fulData, selData, nNeighbors, nBinGroups$
Ensure: $posibleSolutions, posibleSolutionsPct$
$minDataSel \leftarrow minseltData[]$
$maxDataSel \leftarrow maxseltData[]$
$minPerSel \leftarrow getPercenile(fulData[curVarible], minDataSel)$
$maxPerSel \leftarrow getPercenile(fulData[curVarible], maxDataSel)$
for seq in nNeighbor **do**
  $perMin \leftarrow 0$
  $perMax \leftarrow 100$
  if $minPerSel + seq \geq 0$ **then**
    $perMin \leftarrow minPerSel + seq$
  end if
  if $maxPerSel + seq \leq 100$ **then**
    $perMax \leftarrow maxPerSel + seq$
  end if
  $varMinPerSel \leftarrow getPercenile(fulData[curVarible], perMin)$
  $varMaxPerSel \leftarrow getPercenile(fulData[curVarible], perMax)$
  $tempData[] \leftarrow fulData[(fulData[curVarible] \geq varMinPerSel)$
  $and(fulData[curVarible] \leq varMaxPerSel)]$
  $n \leftarrow getHist(tempData[curVarible])$
  $tempSolution \leftarrow getAvg(n)$
  $tempPct[] \leftarrow getPct(curVarible, tempData[], nNeighbors)$
  if $tempSolutionnotinposibleSolutions$ **then**
    $posibleSolutions.add(tempSolution)$
    $posibleSolutionsPct.add(tempPct)$
  end if
end for

Optimal Solution. Algorithm 2 selects the optimal threshold value as the good enough solution. This threshold was found in the vector of bin groups chosen as possible solutions for each variable. This selection was produced through the application of predefined rules and fulfilled by the use of percentages such as the percentage of minimal width for acceptable data (*reqMinWidth*), the percentage of maximal width for acceptable data (*reqMaxWidth*), and the ideal amount of data (*reqHeight*) in each bin group chosen by each variable as a possible solution. Also, it is important to emphasize that the predefined rules were established

according to the expertise of the data analyzer. Using the current dataset, and after testing possible best values for each rule, the best values were 5%, 42% and 10% for the variables *reqMinWidth*, reqMaxWidth, *reqHeight* respectively.

Therefore, those variables were taken as input data for the possible solutions *posibleSolutions*, the percentage of the possible solution *posibleSolutionsPct*, the minimum width required by the user for the selected data $reqMinWidth$, the maximum width required by the user for the selected data $reqMaxWidth$, and the height needed for the user for the selected data $reqHeight$.

Algorithm 2. Optimal solution.

Require: $posibleSolutions, posibleSolutionsPct$
Require: $reqMinWidth, reqMaxWidth, reqHeight$
Ensure: $optimalSolution$
for current in posibleSolutions **do**
  if $current \geq reqMinWidth$ **then**
    if $current \leq reqMaxWidth$ **then**
      if $current \geq reqHeight$ **then**
        $optimalSolution \leftarrow posibleSolutions[current]$
      end if
    end if
  end if
end for

Since the metaheuristic algorithm is executed on each analyzed variable, and each contains filtered data with different indexes, a final step is required for the selection of tuples to be displayed in the last PCP. For this reason, the union and the intersection of the filtered data indexes of each variable were tested, giving as best result in the merge of these indexes. The merge presented a loss of 6%, the minor data loss compared to the intersection loss of 11% of the indexes.

3.3 Threats to Validity

Although the study was performed and focused on multiple domains, some threats to its validity were identified, which are listed below to give an understanding and how to address them.

Threshold Value. The threshold values calculated as alternative solutions and the good enough solution for the SA proposed algorithm are based on their mean value, the most widely used metric. Notwithstanding, the threshold value can be calculated in other mathematical ways, such as the median, mode, or any other matter the data analyst considers. This study used the mean as the threshold value for each analyzed variable, meaning that whether different values could have produced better results or not is currently unknown.

Number of Bins. Due to the dispersion presented in the data used in our study, an approximation technique for histograms was sought to smooth the data. The method chosen was dividing the data into bins using the Freedman-Diaconis equation [23]. This equation provided the ideal size of bins to be applied to the data. However, it may not be the best in other data fields. For this reason, it is advisable to try other equations for obtaining the ideal bin size, such as the equation of Sturge, normal of Scott, normal of Rice, ruler of Rice, and Scot [23].

Control Variables. An expert in the problem domain must adjust the control variables in the SA algorithm according to whether or not they reduce data. The control variables must be adjusted to reduce the data to the most representative data, based on the judgment of a subject matter expert. However, this judgment may differ from the variable adjustment judgment of another subject matter expert. The variables; percentage of minimal width for acceptable data (*reqMinWidth*), percentage of maximal width for acceptable data (*reqMaxWidth*) and the ideal amount of data (*reqHeight*) set the initial rules for the algorithm by a percentage on each of them. The percentage per variable established in our proposal is based on the data analyzed; however, this data may be considered deficient depending on the expert analyst.

4 Results

The $PM_{2.5}$ air pollution variable was used to exemplify the first step of our proposal, where raw data was preprocessed, and then the data distribution discretization stage was applied. The steps presented in this section showed the results of all chosen variables for each step, not only the $PM_{2.5}$ variable.

Results in Data Distribution Discretization. While the $PM_{2.5}$ variable was used to follow the current process, the rest of the variables were not presented. Raw data of each missing variable was given against the data selected by the user and divided into the ideal number of bins. As shown in Fig. 5, the data represented on blue bars correspond to raw data, while green bars correspond to data selected by the user, which in our case includes the hour range from 11:00 to 16:00.

Results in Alternative Solutions. Once variables were preprocessed and their bins were defined, the possible solutions were established using Algorithm 1. As described in Sect. 3.2, each possible solution corresponds to an average value per bin, where it was added only if it was unique on the array of possible solutions for each variable. In Fig. 6, possible solutions are represented with a green color scale.

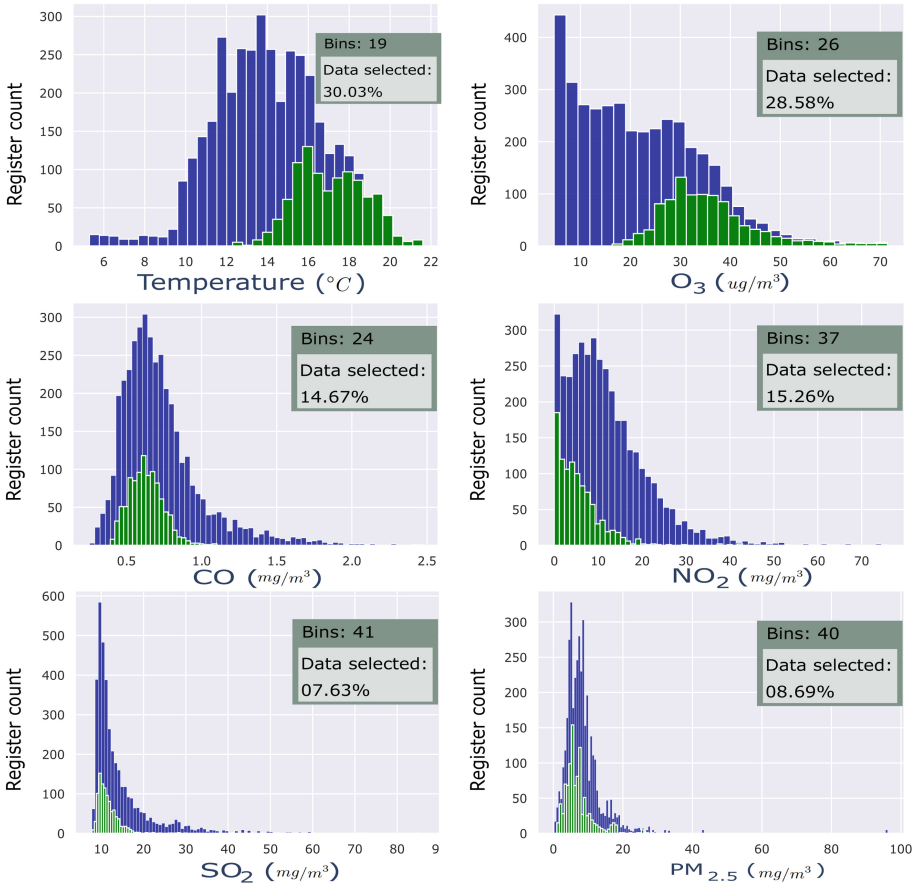


Fig. 5. Raw data versus selected data. (Color figure online)

Results in Optimal Solution. After applying the predefined rules, as mentioned in Sect. 3.2, the best possible solution was chosen. This optimal result reflects a threshold value used to filter the most relevant data from the users' initial selection. All data and bins that exceed the optimal solution value were considered necessary for marking a pattern of behavior within PCP. In contrast, the values below this mark were neglected for the analysis by the data expert. In Fig. 6, each optimal solution is presented with a red line per variable, where the left column shows the possible solutions (green color scale lines) and the optimal solution (red color line), while the right column shows the interaction of the possible solutions with the optimal solution of each variable.

In Fig. 7, the last PCP is shown, but only with the selection of the tuples obtained by the union of indexes of the analyzed variables, thus representing the data filtered by the metaheuristic algorithm.

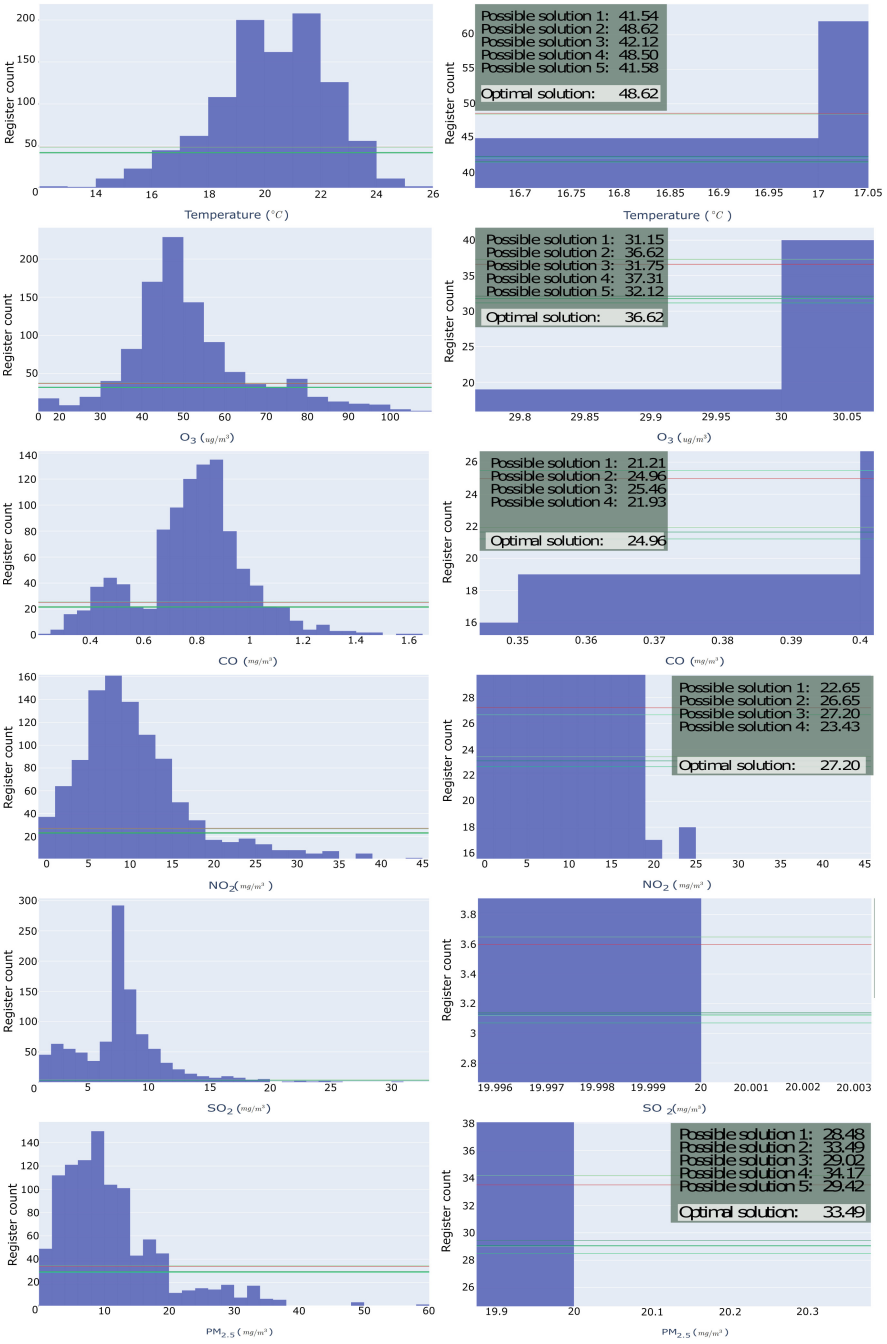


Fig. 6. Alternative and optimal solution. (Color figure online)

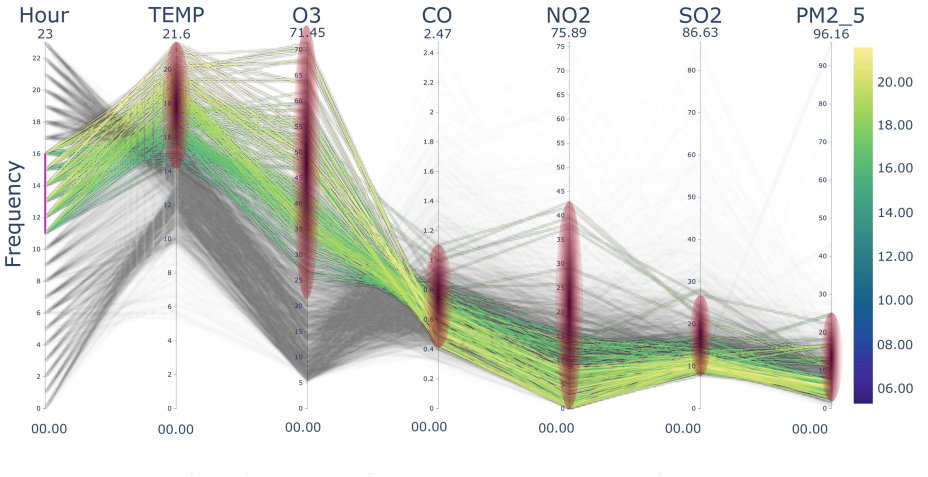


Fig. 7. PCP after being filtered by metaheuristic from July 2018.

5 Conclusions

Visualizing a plot that presents correlation and data behavior from large amounts of data can be a hurdle in human decision-making; this is an issue when employing PCP. For this reason, our proposal seeks to mitigate the difficulty of working with large amounts of data by optimizing the PCP, using Simulated Annealing, a metaheuristic algorithm, to reduce the amount of data in each analyzed variable involved in PCP by filtering relevant data. The proposed algorithm divided selected raw data into ideal bin groups, then found threshold values as possible solutions and established a threshold value as a good enough resolution. Thus, the algorithm filtered all the records that did not exceed the optimal threshold value. For a better understanding, the overall algorithm was divided into two sub-algorithms, the first for finding a set of possible solutions and the second for detecting the optimal value for the threshold value on each variable. In our particular case, the proposed solution for the PCP problem of extensive data analysis is explored through the use of four air pollution variables and one meteorological variable measured in the city of Cuenca, taking the air pollution $PM_{2.5}$ variable as the primary reference. Finally, the data reduction by applying the metaheuristic algorithm fulfills its objective, showing a clear improvement when interpreting the original data selection for July, between 11:00 and 16:00.

This proposal can be applied in any field with high amounts of data. It allows the user to reduce vast amounts of data to the most relevant and present it on a comprehensible PCP. The proposed SA algorithm has adjustment variables to find the best threshold value for each input variable. In future studies, the proposed algorithm will test with a larger dataset produced using a more significant number of sensor measurements in multiple locations of a country. Another

adequate study for confirming the effectiveness of our proposal will be in the commerce field, as it generates enormous correlational data volume.

References

1. Agresi, A.: An Introduction to Categorical Data Analysis, 3rd Edn. Wiley (2018)
2. Albazzaz, H., Wang, X.Z.: Historical data analysis based on plots of independent and parallel coordinates and statistical control limits. J. Process Control **16**, 103–114 (2006). https://doi.org/10.1016/j.jprocont.2005.05.005
3. Ali, M., Deo, R.C., Maraseni, T., Downs, N.J.: Improving spi-derived drought forecasts incorporating synoptic-scale climate indices in multi-phase multivariate empirical mode decomposition model hybridized with simulated annealing and kernel ridge regression algorithms. J. Hydrol. **576**, 164–184 (2019). https://doi.org/10.1016/j.jhydrol.2019.06.032
4. Ansari, S., Mohanlal, R., Poncela, J., Ansari, A., Mohanlal, K.: Importance of big data. In: Handbook of Research on Trends and Future Directions in Big Data and Web Intelligence, pp. 1–19. IGI Global (2015)
5. Berinato, S.: Visualizations that really work. Harvard Bus. Rev. **94**(6), 93–100 (2016)
6. Berthold, M.R., Hall, L.O.: Visualizing fuzzy points in parallel coordinates. IEEE Trans. Fuzzy Syst. **11**, 369–374 (2003). https://doi.org/10.1109/TFUZZ.2003.812696
7. Dasgupta, A., Kosara, R.: The importance of tracing data through the visualization pipeline. In: Proceedings of the 2012 BELIV Workshop: Beyond Time and Errors-Novel Evaluation Methods for Visualization, pp. 1–5 (2012)
8. Dehghani, R., Poudeh, H.T., Izadi, Z.: The effect of climate change on groundwater level and its prediction using modern meta-heuristic model. Groundwater Sustain. Develop. **16**, 100702 (2022). https://doi.org/10.1016/j.gsd.2021.100702
9. Diehl, S.: Past, present, and future of and in software visualization. In: Past, Present, and Future of and in Software Visualization, pp. 3–11 (2015)
10. Fan, J., Li, R.: Statistical challenges with high dimensionality: feature selection in knowledge discovery. arXiv preprint arXiv:math/0602133 (2006)
11. Feng, M., et al.: Big data analytics and mining for effective visualization and trends forecasting of crime data. IEEE Access **7**, 106111–106123 (2019). https://doi.org/10.1109/ACCESS.2019.2930410
12. Groves, R.M., Fowler Jr, F.J., Couper, M.P., Lepkowski, J.M., Singer, E., Tourangeau, R.: Survey Methodology, vol. 561. Wiley (2011)
13. Gupta, A., Deokar, A., Iyer, L., Sharda, R., Schrader, D.: Big data & analytics for societal impact: recent research and trends. Inf. Syst. Front. **20**(2), 185–194 (2018)
14. Heinrich, J., Weiskopf, D.: State of the art of parallel coordinates. In: Eurographics (State of the Art Reports), pp. 95–116 (2013)
15. Johansson, J., Ljung, P., Jern, M., Cooper, M.: Revealing structure within clustered parallel coordinates displays. In: IEEE Symposium on Information Visualization, 2005. INFOVIS 2005, pp. 125–132. IEEE (2005)
16. Karasulu, B., Korukoglu, S.: A simulated annealing-based optimal threshold determining method in edge-based segmentation of grayscale images. Appl. Soft Comput. **11**, 2246–2259 (2011). https://doi.org/10.1016/J.ASOC.2010.08.005
17. Lev, B.: Meta-heuristics: advances and trends in local search paradigms for optimization. Interfaces **30**(4), 94 (2000)

18. Maier, H.R., et al.: Evolutionary algorithms and other metaheuristics in water resources: current status, research challenges and future directions. Environ. Model. Softw. **62**, 271–299 (2014). https://doi.org/10.1016/J.ENVSOFT.2014.09.013
19. Meera, S., Sundar, C.: A hybrid metaheuristic approach for efficient feature selection methods in big data. J. Ambient Intell. Humaniz. Comput. **12**(3), 3743–3751 (2020). https://doi.org/10.1007/S12652-019-01656-W
20. Mohamed, N.S., Zainudin, S., Othman, Z.A.: Metaheuristic approach for an enhanced MRMR filter method for classification using drug response microarray data. Exp. Syst. Appl. **90**, 224–231 (2017). https://doi.org/10.1016/J.ESWA.2017.08.026
21. Abedinia, O., Nima Amjady, N.G.: Solar energy forecasting based on hybrid neural network and improved metaheuristic algorithm. Comput. Intell. **34**, 241–260 (2018). https://doi.org/10.1111/COIN.12145
22. Perkhofer, L.M., Hofer, P., Walchshofer, C., Plank, T., Jetter, H.C.: Interactive visualization of big data in the field of accounting. J. Appl. Account. Res. **20**, 497–525 (2019). https://doi.org/10.1108/JAAR-10-2017-0114
23. Sahann, R., Müller, T., Schmidt, J.: Histogram binning revisited with a focus on human perception. In: 2021 IEEE Visualization Conference (VIS), pp. 66–70 (2021). https://doi.org/10.1109/VIS49827.2021.9623301
24. Sataloff, R.T., Johns, M.M., Kost, K.M.: Data Cleaning 2019. ACM Books Series (2019)
25. Siirtola, H., Räihä, K.J.: Interacting with parallel coordinates. Interact. Comput. **18**(6), 1278–1309 (2006)
26. Sousa, T., Soares, J., Vale, Z.A., Morais, H., Faria, P.: Simulated annealing metaheuristic to solve the optimal power flow. In: 2011 IEEE Power and Energy Society General Meeting, pp. 1–8 (2011). https://doi.org/10.1109/PES.2011.6039543
27. Weidele, D.K.I.: Conditional parallel coordinates. In: 2019 IEEE Visualization Conference (VIS), pp. 221–225 (2019). https://doi.org/10.1109/VISUAL.2019.8933632