# A Comparison of Monotonic Correlation Measures with Outliers

AHMED ALSAYED[1*], GIANCARLO MANZI[1,2]
[1]Department of Economics, Management and Quantitative Methods
[2]Data Science Research Center
University of Milan, ITALY
*Ahmed.alsayed@unimi.it

*Abstract.* - This paper aims at examining the performance of a recently proposed measure of dependence – the Monotonic Dependence Coefficient – MDC - with respect to classical monotonic correlation measures like Pearson's $r$, Spearman's $\rho$, and Kendall's $\tau$, using simulated outlier contaminated and non-contaminated data sets as well as a contaminated real dataset, considering three different cases. This comparison aims at checking how and when these coefficients detect dependence relationships between two variables when outliers are present. Several scenarios are created with multiple values for the dependence measures, outlier contamination fractions and data patterns. The basic simulated dataset is generated from a bivariate standard normal distribution. Using values generated from the exponential, power-transformed, lognormal, and Weibull distributions, added to the basic generated dataset, we transform the contaminated data, allowing for multiple patterns. The main findings tend to favour the Spearman's $\rho$ coefficient for most of the simulated scenarios, especially when the outlier contamination is taken into account, whereas MDC performs better than $\rho$ in non-contaminated data. However, in the real data scenario Spearman's $\rho$ outperforms the other measures in two out of three cases, whereas MDC performs better in the other case.

*Key-Words:* Outliers; Correlation Coefficient; Monotonic Dependence; Monte Carlo Simulation; Environmental Quality; Economic Growth.

## 1 Introduction

Dependence analysis is one of the most important research topics not only in statistical literature, but also in many other fields of science. Several statistical approaches have been proposed since the dawn of statistical science to model the relationship between continuous, discrete or ordinal variables, while the increasing availability of ordinal datasets, especially in social sciences, has recently contributed to the development of new reliable methods also for qualitative variables. Among the most important dependence analysis methods, monotonic dependence methods, such as Pearson's $r$, Spearman's, and Kendall's $\tau$, are also the most used. Pearson's product moment correlation coefficient $r$ is generally appropriate to detect linear relationship between two continuous variables, whereas the Spearman's rank correlation coefficient $\rho$ is used to measure the strength of the association between two ordinal variables. Another very popular coefficient is Kendall's $\tau$, which is used as an alternative to $\rho$. However, most of the times i.e. in practical situations when both measures are not too close to 1 in absolute values - $\tau$ has a stable relationship with $\rho$ [1, 2].

Among recently proposed methods, the Monotonic Dependence Coefficient [3] (MDC), based on the Lorenz and concordance curves, and built by comparing the observed values of the dependent variable and the corresponding values of the independent variables properly reordered according to the relationship with the independent variable, is suitable to be applied when the dependent variable is continuous or discrete and the independent variable is at least of ordinal nature. This method has been proved to be an effective competitor of the Pearson and Sperman's coefficients, especially in the case of non-normally distributed data and when some important information is lost [4].

In this literature stream Bishara and Hittner [5] compared the bias in point estimates among five correlation approaches: Pearson's $r$, Spearman's $\rho$, the bootstrap estimate, the Box–Cox transformation family, and a normalizing transformation correlation after the ranked-based inverse transformation, using the *rankit* equation [6]. They performed this comparison using Monte Carlo simulation for different scenarios with normal and non-normal data, various combinations of distribution shapes and sample sizes. The degree of inaccuracy was evaluated using both bias and Root Mean Square Error (RMSE).

The main finding in Bishara and Hittner's work was that the Spearman's $\rho$ coefficient and the rankit correlations, being similar methods, perform better in reducing the bias and providing more robust estimates. Rankit correlations also minimized the RMSE for most sample sizes, except for the smallest samples, for which bootstrapping was more effective. Generally, these results justify the use of carefully chosen alternatives to Pearson's $r$ when data are non-normally distributed. The non-normality of the data (particularly distributions having heavy-tails) can cause a biased and overestimated evaluation of the correlation. Some correlation coefficients can mitigate the bias problem better than others, depending on the choice of alternatives for the sample size and distribution shape.

Together with the sample size and the type of underlying distribution of the data, in some specific contexts the existence of outliers in the dataset can become a crucial issue for estimation, particularly when variables are continuous [7]. Outliers can be the main cause of under- or over-estimation for the parameter of interest. In such cases, some of the correlation coefficients might fail in detecting the real dependence relationship between the variables and some corrections are needed [8].

It is not infrequent that real data sets contain outliers presenting unusually large or small values when compared with the majority of the observations. The existence of outliers may cause a negative effect on the output, particularly on correlation and regression measures or measures based on distributional assumptions. However, when they are not caused by measurement errors, outliers may be informative about some characteristics of the observations. For example, in psychology, social research or demographics, outlier analysis helps detecting people clustered apart, having a different social behaviour with respect to the majority of the population, as is the case of tiny ethnic minorities or migrants [9].

In the case of dependence analysis, due to the effects of the outliers in detecting the real relationship between variables, several statistical methods have been deployed to adjust the estimates. Although some methods are powerful with large normally distributed data, they might be sensitive to outliers or extreme values, and may be problematic when applied to non-normal data or small samples [10]. Moreover, some simpler measures like the median are more robust to outliers than others and might be more suitable in such cases.

The usual definition of outliers is that observations have unusually large/small values compared to other observations in a data set. Barnett and Lewis [10] defined an outlier as the observation appearing to be inconsistent with the remainder of the others. Grubbs [11] defined the outlying observation as a value appearing to deviate markedly from other elements of the sample.

Outliers could occur as incorrectly recorded data, or come from heterogeneous data sets or a different underlying population. Most likely, outliers are present when data are collected from heterogeneous groups having different characteristics regarding a certain variable.

Outliers may be deleterious for data analysis for several reasons: (i) they could increase the error in the estimates, (ii) reduce the power of statistical tests, (iii) decrease the normality of the data distribution, (iv) influence the estimated coefficients, and, therefore, (v) provide distorted information.

The main purpose of this study is to compare the performance of four statistical methods in detecting the dependence relationship, namely, the Monotonic Dependence Coefficient (MDC), Pearson's $r$ Spearman's $\rho$ and Kendall's $\tau$ for contaminated and non- contaminated data sets. This comparison is performed via a Monte Carlo simulation study and an application on a real dataset. These methods are quite effective when working with large dataset that are fairly normally distributed, but many distributions of real-world data do not follow the normal distribution. They are often highly skewed because of the inclusion of some extreme values, being the skewness very often positive, so that the distribution resembles a lognormal distribution. This is why the lognormal distribution is often used in practice [12]. Findings from this study might help to propose some guidelines to the use of a proper dependence analysis method, which could be robust against the presence of outliers and/or skewed data, and give advice in case of violation of statistical test assumptions.

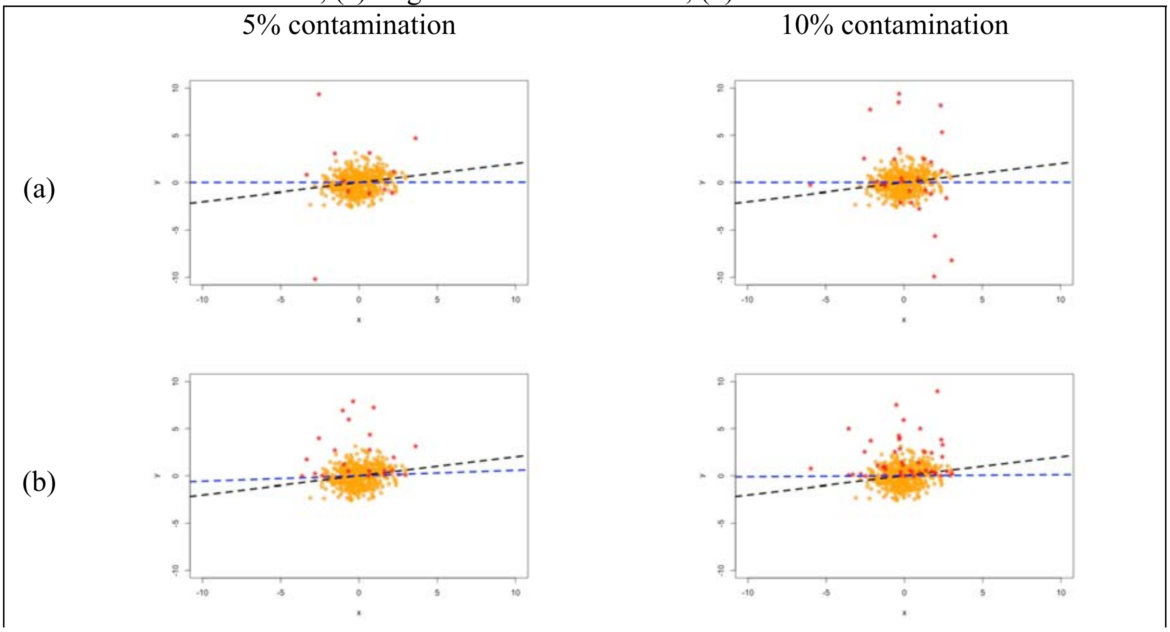
The rest of the paper is organized into four sections: the next section illustrates the simulation study, followed by a section describing its results. Section 4 presents an application to a real situation and the last section concludes the paper.

## 2 Simulation Study

The assessment of the performance and comparisons among the four dependence measures were carried out through Monte Carlo (MC) simulation. In this simulation we applied MC on MDC, Pearson's $r$, Spearman's $\rho$, and Kendall's $\tau$ on outlier contaminated and non-contaminated data sets. All the four correlation coefficients have a direct or inverse monotonic relationship ranging from −1 to +1, assuming the value zero with no correlation. We analyse the behaviour of all four dependent measures in different experimental scenarios, according to the sample size, the percentage of the contamination in the data and the type of distribution generating the outliers.

The simulation dataset was built in two steps [13]. In the first step, we generated correlated data for two variables $x$ and $y$ from a bivariate standard normal distribution, having zero skewness. An MC process with 10,000 iterations was performed generating random samples of size $n = 500$ having two particular correlation structures. Finally, the four dependence methods were applied to evaluate the relationship between $x$ and $y$ by averaging over the MC samples. In the second step, we generated a contamination correlated data set amounting to 5% and 10% of the whole sample size for $x$ and $y$ by first generating a bivariate standard normal distribution with zero mean parameter and correlation matrix $\begin{pmatrix} 1 & r \\ r & 1 \end{pmatrix}$. Two values of $r$ (0.2, 0.8) were considered. Contaminating data for the variable $y$ were added to these values using the exponential distribution, the power transformation $w = y^5 + y^3 + y$, the lognormal distribution and the Weibull distribution with scale parameter equal to 2.1 and shape parameter equal to 1.1. Examples of generated data are presented in the scatterplots provided in Figures 1 and 2. Yellow points represent the values generated from a bivariate standard normal, whereas the red stars represent the contamination outlier observations. Black and blue dash lines indicate the simple regression lines before and after the contamination, respectively. In general, adding the random contamination to the bivariate normal dataset will reduce the values of the regression coefficients for all scenarios and the relationship between $x$ and $y$ goes toward the horizontal direction after adding the contaminated data (Fig. 1 and 2, (a) and (b)). The lognormal and the Weibull outlier scenarios are the least influenced by the contamination as the patterns of the two lines are almost indiscernible (Fig. 1 and 2, (c) and (d)).

Fig. 1. Scatter plots for simulated data ($r = 0.20$; $n = 500$). (a) Power transformation; (b) Exponential transformation; (c) Lognormal transformation; (d) Weibull transformation.



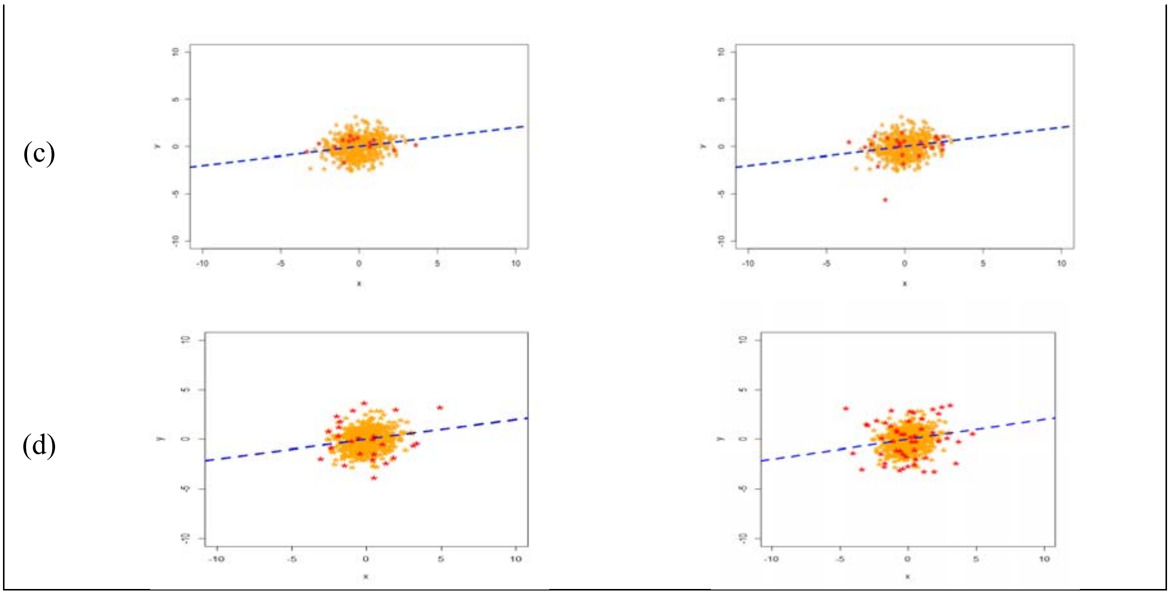
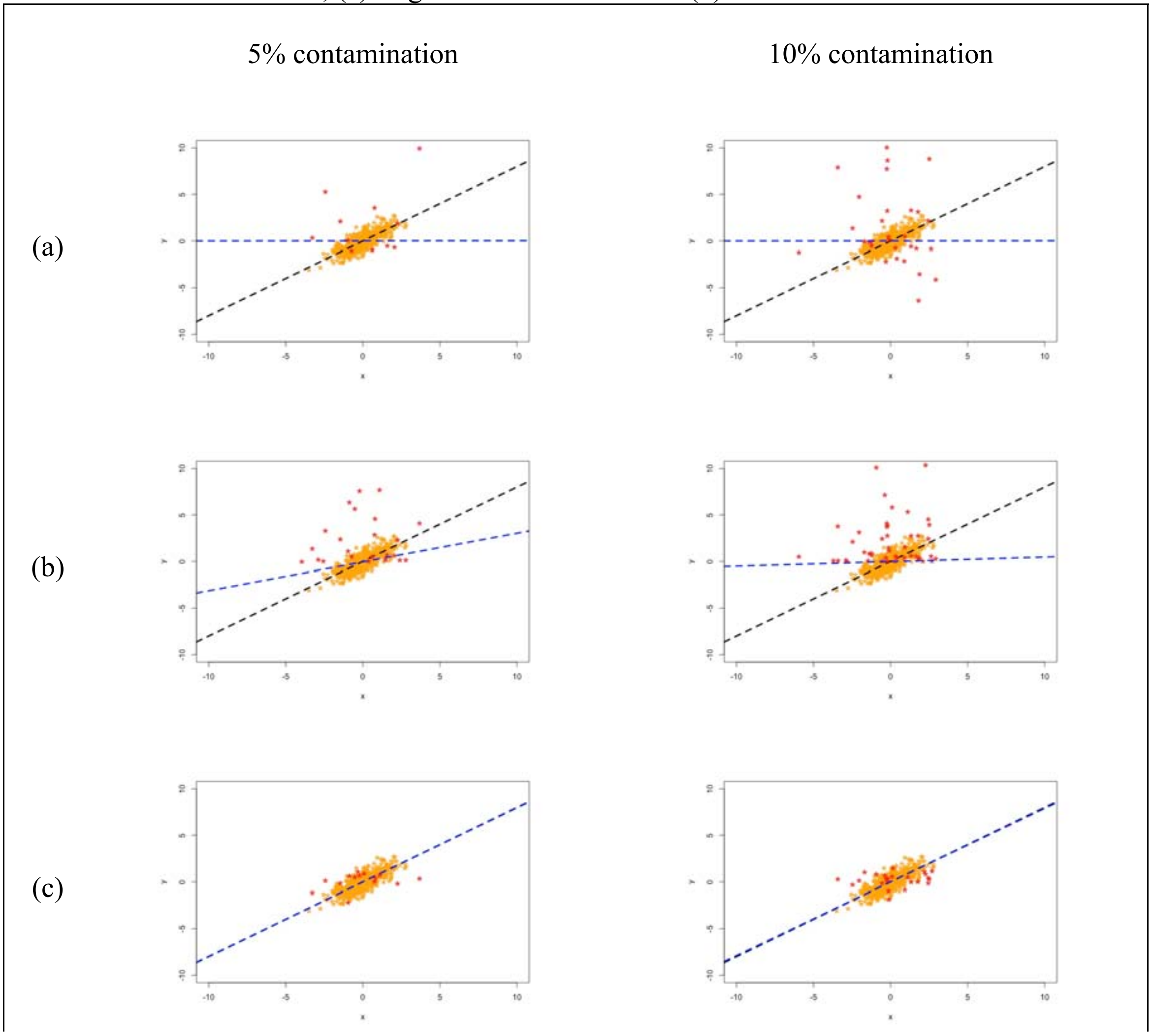
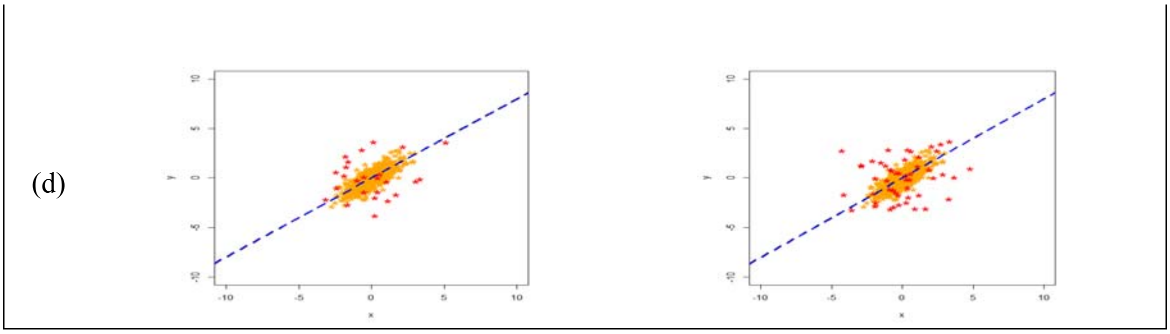


Fig. 2. Scatter plots for simulated data ($r = 0.80$; $n = 500$). (a) Power transformation; (b) Exponential transformation; (c) Lognormal transformation (d) Weibull transformation





# 3 Results on Simulated Data

Our simulation procedure considers samples generated from a bivariate standard normal distribution. Then, contamination data are added to the values of these samples. Results of the MC simulations on the performance of the four statistical measures considered – MDC, $r$ $\rho$ and $\tau$ - in detecting the dependence relationship on contamination and non-contamination datasets generated as described in the previous section are shown in Table 1 and Table 2. The evaluation of the performances of these methods is determined by their ability to detect the dependence relationship between $x$ and $y$ in the contaminated data sets. The benchmark is the value of $r$ in the non-contaminated data set, so all the results are compared with it.

Table 1. Simulation results. $r = 0.20$, $n = 500$. Most successful results for each scenario in bold – Validation percentages in brackets

| | MDC | $r$ | $\rho$ | $\tau$ | MDC | $r$ | $\rho$ | $\tau$ |
|---|---|---|---|---|---|---|---|---|
| **Without outliers** | | | | | | | | |
| | **0.199**<br>(0.995) | 0.2<br>(1.00) | 0.191<br>(0.955) | 0.128<br>(0.64) | | | | |
| **With outliers** | | | | | | | | |
| | Percentage of contamination | | | | | | | |
| **Transformation** | 5% | | | | 10% | | | |
| Power transformation | 0.079<br>(0.395) | 0.016<br>(0.080) | **0.170**<br>(0.850) | 0.120<br>(0.60) | 0.080<br>(0.400) | 0.019<br>(0.095) | **0.165**<br>(0.825) | 0.113<br>(0.565) |
| Exponential | 0.154<br>(0.77) | 0.076<br>(0.38) | **0.178**<br>(0.89) | 0.120<br>(0.60) | 0.129<br>(0.645) | 0.051<br>(0.255) | **0.167**<br>(0.835) | 0.110<br>(0.55) |
| Lognormal | 0.176<br>(0.880) | 0.159<br>(0.795) | **0.177**<br>(0.885) | 0.119<br>(0.595) | 0.157<br>(0.785) | 0.133<br>(0.665) | **0.164**<br>(0.820) | 0.111<br>(0.555) |
| Weibull (scale 2.1, shape 1.1) | **0.192**<br>(0.96) | 0.183<br>(0.915) | 0.182<br>(0.91) | 0.121<br>(0.605) | **0.186**<br>(0.93) | 0.170<br>(0.85) | 0.175<br>(0.875) | 0.116<br>(0.58) |

Table 2. Simulation results $r = 0.80$, $n = 500$. Most successful results for each scenario in bold– Validation percentages in brackets

| | MDC | $r$ | $\rho$ | $\tau$ | MDC | $r$ | $\rho$ | $\tau$ |
|---|---|---|---|---|---|---|---|---|
| **Without outliers** | | | | | | | | |
| | **0.799**<br>(0.998) | 0.8<br>(1.000) | 0.785<br>(0.981) | 0.59<br>(0.74) | | | | |
| **With outliers** | | | | | | | | |
| | Percentage of contamination | | | | | | | |
| **Transformations** | 5% contamination | | | | 10% contamination | | | |
| Power transformation | 0.324<br>(0.405) | 0.067<br>(0.084) | **0.728**<br>(0.910) | 0.550<br>(0.688) | 0.315<br>(0.393) | 0.076<br>(0.095) | **0.678**<br>(0.847) | 0.517<br>(0.65) |
| Exponential | 0.615<br>(0.769) | 0.302<br>(0.378) | **0.732**<br>(0.915) | 0.551<br>(0.689) | 0.514<br>(0.643) | 0.201<br>(0.251) | **0.686**<br>(0.858) | 0.517<br>(0.646) |
| Lognormal | 0.706<br>(0.882) | 0.639<br>(0.798) | **0.725**<br>(0.906) | 0.546<br>(0.683) | 0.632<br>(0.790) | 0.531<br>(0.664) | **0.672**<br>(0.840) | 0.508<br>(0.635) |
| Weibull (scale 2.1, shape 1.1) | **0.770**<br>(0.963) | 0.733<br>(0.916) | 0.750<br>(0.938) | 0.557<br>(0.696) | **0.744**<br>(0.930) | 0.679<br>(0.849) | 0.719<br>(0.899) | 0.528<br>(0.660) |

Findings of Table 1 and Table 2 show that generally each method presents correlation values to detect the relationship between $x$ and $y$ lower to the benchmark values $r = 0.2$ $r = 0.8$ and, for each method, increasing the contamination percentage ends up in further lowering the measures' values. When there is no contamination and presence of asymmetric outlying observations, MDC is the best method to detect the relationship between $x$ and $y$ with respect to $r$, as its values on the non-contaminated dataset are $\{0.199, 0.799\}$ which are almost identical to the values $\{0.2, 0.8\}$ of the Pearson's $r$ coefficient.

MDC had less variability than $r$ throughout the different scenarios as compared to other methods in the presence of outliers. Passing from a contaminated percentage of 5% to 10% in the data set provided similar findings of the performance of the four methods in both cases for $r = \{0.2, 0.8\}$. However $\rho$ performed decidedly better than MDC in the first two scenarios, but in the third and fourth scenarios the difference between the two coefficients was very tiny. In the lognormal scenario (ρ = 0.2, 5% contaminated data) MDC had practically the same behaviour of $\rho$. MDC behaved better on the third and fourth scenarios as the lognormal and Weibull distributions are from the normal distribution family.

## 4 Results on Real Data

In this section we present results on the performance of MDC, $r$, $\rho$ and τ by using a contaminated real data set. We considered a contaminated data set containing two variables: the per-capita gross domestic product (GDP) measured in USD$ (y), and the amount of Carbon dioxide emission ($CO_2$), measured in metric tons per capita (x) for Luxembourg from 1960 to 2008. Data were obtained from the World Bank website [14] [15].

Before performing the analysis using the four coefficients, we should identify the outliers in the data set as it is important to enhance the efficiency of the estimated coefficients. For that purpose, the Mahalanobis distance method has been applied to detect the outliers in the relationship between GDP and $CO_2$, as it provides a robust estimation by using the minimum covariance determinant (MCD) estimator. Moreover, this method identifies the leverage points and the residuals outliers at the same time. The Robust Mahalanobis Distance is constructed as

$$RD = \sqrt{\left(x_i - \mu_{MCD}\right) \sum\nolimits_{MCD}^{-1} \left(x_i - \mu_{MCD}\right)'} \qquad (1)$$

where $\mu_{MCD}$ and $\sum_{MCD}^{-1}$ are respectively the mean vector and the covariance matrix estimated by MCD [16].

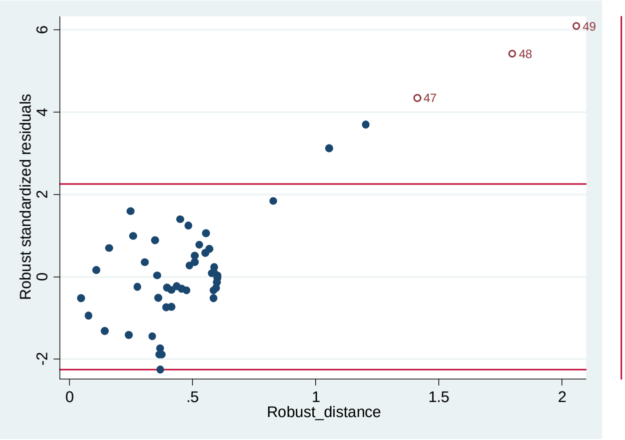


Fig. 3. Diagnostic Mahalanobis distances for the relationship between $CO_2$ and GDP.

The diagnostic outliers plot is shown in Figure 3. The scatterplot shows the values of the Mahalanobis distance on the horizontal axis, while on the vertical axis the values of standardized residuals are plotted. The values above 2.5 and lower -2.5 in the horizontal line are considered as outliers. Figure 3 clearly shows that the data contains some outliers.

We performed the analysis on three different measurement scales, similar to the classification in [3]:

a) The case with continuous values for $x$ and $y$, as provided in the original data to represent the real situation of the analysis.
b) The case with the average groups values for $x$ (five groups with group average $CO_2$ emission); we classified the $x$ variable into five groups, the intervals being determined by using the max-min/5 rule (bottom row of Table 3).
c) The case with ordinal values for the $x$ variable encoded as group rank (first row of Table 3).

The purpose of this analysis of the three situations is to check the validations of the four coefficients into detecting the relationship among the three cases as pieces of information are lost.

Table 3 Frequency Distribution of Groups and Average of $CO_2$

| Groups | 1 | 2 | 3 | 4 | 5 |
|---|---|---|---|---|---|
| $CO_2$ | Less | 25422- | 48622- | 71821- | More than |

| | | | | | |
|---|---|---|---|---|---|
| intervals | than 25422 | 48621 | 71821 | 95020 | 95020 |
| Number (n) | 29 | 10 | 5 | 3 | 2 |
| Actual average | 8906.70 | 39942.96 | 52934.98 | 81803.77 | 112560.17 |

Obtained coefficients' values for the three coefficients for the cases a), b) and c) are displayed in Table 4.

With regard to case (a), $\rho$ and MDC achieves slightly similar results (-0.844, -0.828) respectively. Such findings suggest the presence of a high concordance between $y$ and $x$. Kendall $\tau$ presents the smallest value (-0.571). This result for the real data is coherent with the results obtained in the simulation study.

With regard to case (b) all the four coefficients values decreased. However, $\rho$ outperformed to detect the dependence relationship between the two variables. MDC and Pearson's $r$ became weaker to detect the dependence relationship because the actual means of $x$ are replaced by equidistant values. On the other hand, $\tau$ did not decrease much as the other coefficients because the involved variables are of a different nature. In the last case (c) there are no substantial differences with respect to the results obtained in (b) except on $r$ as it results slightly decreased.

Table 4. Simulation results on a contaminated real data. (a) Continuous variable; (b) Average values in $CO_2$ emissions groups; (c) Ranks related to groups. Most successful results for each scenario in bold.

| | (a) | (b) | (c) |
|---|---|---|---|
| MDC | **-0.844** | -0.614 | -0.614 |
| $r$ | -0.682 | -0.597 | -0.588 |
| $\rho$ | -0.828 | **-0.689** | **-0.689** |
| $\tau$ | -0.571 | -0.522 | -0.522 |

## 5 Conclusion

In the presence of a vast literature concerning the dependence analysis methods, it is always problematic to choose the best performing method depending on the real situation at hand, especially when outliers are present. Within the scope of our work, we have tried to address this point by comparing, under several distinct data patterns and different outlier contamination scenarios, four monotonic methods, the Pearson's correlation coefficient, the Spearman's rank correlation coefficient, the Kendall correlation coefficient and a recently proposed method, the Monotonic Dependence Coefficient. Even though these coefficients seem to cover positively several situations of dependence analysis, they present some drawbacks especially in cases where one variable is continuous and the other is of another nature and when outliers are present. Our findings showed that the Spearman's $\rho$ method is more efficient in terms of detecting the dependence relationship for contaminated data sets. However, MDC is very close to it when outliers having a similar distribution (lognormal or Weibull) to that of the non-contaminated values (normal) are present and performs better with non-contaminated data. The results on real data are coherent with the results obtained in the simulation study.

We argue that the simulation study we have designed to analyse the performance in a dependence problem could have a more general application scope, e.g. concerning other kinds of statistical dependence techniques and investigations. Our simulation study permitted us to give some useful insights to the potential users on the choice of the most performing dependence methods.

**Acknowledgment:**

**The author would like to thank the two anonymous reviewers for all of their careful, constructive and insightful comments in relation to this work. The author would like to thank the Editor of the Journal for the editorial work and the useful remarks and comments.**

*References:*

[1] Kendall MG. *The Advanced Theory of Statistics*, vol. 1, fourth ed. London: Charles Griffin & Company, 1948.

[2] Fredricks GA, Nelsen RB. On the relationship between Spearman's rho and Kendall's tau for pairs of continuous random variables. *J Stat Plan Inference*. 2007; 137: 2143-2150.

[3] Ferrari PA, Raffinetti E. A Different Approach to Dependence Analysis. *Multivar Behav Res*, 2015; 50(2): 248-264.

[4] Raffinetti, E, Ferrari, PA. New Perspectives for the MDC Index in Social Research Fields. In Morlini, I., Minerva, T., Vichi, M. (Eds.): *Advances in Statistical Models for Data Analysis*, Zurich, Switzerland: Springer Verlag: 211-219, 2015.

[5] Bishara AJ, Hittner JB. Reducing bias and error in the correlation coefficient due to nonnormality. *Educ Psychol Meas*, 2015; 75(5): 785-804.

[6] Bliss CI. *Statistics in Biology*. New York (NY): McGraw-Hill; 1967.

[7] Rousseeuw PJ, Leroy AM. *Robust Regression and Outlier Detection*. New York (NY): John Wiley & Sons; 1987.

[8] Abdullah MB. On a Robust Correlation Coefficient. *The Statistician*, 1990; 39: 455-460.

[9] Osborne JW, Overbay A. The Power of Outliers (and Why Researchers Should Always Check Them). *Practical Assessment, Research and Evaluation*, 2004; 9(6): 1-8.

[10] Barnett V, Lewis T. *Outliers in statistical data.* 3rd edition, 1994, Chichester (UK): John Wiley & Sons.

[11] Grubbs FE. Procedures for detecting outlying observations in samples. *Technometrics*, 1969, 11:1 - 21.

[12] Iglewicz B, Hoaglin D. *How to detect and handle outliers*. 1993, Milwaukee (WI): ASQC Quality Press.

[13] Vale, C., & Maurelli, V. (1983). Simulating multivariate non-normal distributions. *Psychometrika*, 48(3), 465–471.

[14] Al Sayed, A. R., Isa, Z., & Kun, S. S. (2018). Outliers Detection Methods in Panel Data Regression: An Application to Environment Science. *International Journal of Ecological Economics & Statistics*, 39(1), 73-86.

[15] Al Sayed, A. R., & Sek, S. K. (2013). Environmental Kuznets curve: evidences from developed and developing economies. *Applied Mathematical Sciences*, 7(22), 1081–1092.

[16] Rousseeuw, P. J., and B. C. van Zomeren. 1990. Unmasking multivariate outliers and leverage points. *Journal of the American Statistical Association* 85: 633–639.