Exploratory Analysis of Big Social Data Using MIC/MINE Statistics

Piyawat Lertvittayakumjorn[1], Chao Wu[1], Yue Liu[1,2], Hong Mi[2], and Yike Guo[1(✉)]

[1] Data Science Institute, Imperial College London, London SW7 2AZ, UK
{pl1515,chao.wu,l.yue,y.guo}@imperial.ac.uk
[2] School of Public Affairs, Zhejiang University, Zhejiang 310058, China
spsswork@163.com

Abstract. A major goal of Exploratory Data Analysis (EDA) is to understand main characteristics of a dataset, especially relationships between variables, which are helpful for creating a predictive model and analysing causality in social science research. This paper aims to introduce Maximal Information Coefficient (MIC) and its by-product statistics to social science researchers as effective EDA tools for big social data. A case study was conducted using a historical data of more than 3,000 country-level indicators. As a result, MIC and some by-product statistics successfully provided useful information for EDA complementing the traditional Pearson's correlation. Moreover, they revealed several significant, including nonlinear, relationships between variables which are intriguing and able to suggest further research in social sciences.

Keywords: Exploratory data analysis · Big social data · Maximal information coefficient · Correlation

1 Introduction

Exploratory Data Analysis (EDA) aims to understand the main characteristics of a dataset and, more importantly, find previously unknown relationships in the data. A well-known quantitative technique for EDA is using *Pearson's correlation* (also known as Pearson product-moment correlation coefficient denoted by r or ρ in literature) because it is easy to compute and, in the case of simple linear regression, its square (i.e., r^2) is a *coefficient of determination* telling noise level in a dependent variable predicted from an independent variable. However, a severe weakness of Pearson's correlation is that it can detect only linear relationships, but in real-world data sets, we are also interested in nonlinear relationships which may be missed if we use only Pearson's correlation to explore data.

In 2011, Reshef et al. proposed a nonlinear correlation measure called *Maximal Information Coefficient (MIC)* [8]. It was designed specifically for exploring large datasets with the ability to capture a wide range of associations regardless of their function types or whether they are formed by functional relationships.

G.L. Ciampaglia et al. (Eds.): SocInfo 2017, Part II, LNCS 10540, pp. 513–526, 2017.
DOI: 10.1007/978-3-319-67256-4_41

By simulation, MIC gave similar correlation scores to equally noisy relationships of various types. In addition, the authors of MIC also proposed *MINE statistics* (Maximal Information-based Nonparametric Exploration) which are by-product statistics of MIC calculation that can provide us more information with regard to the variable pair besides the relationship strength reflected by MIC.

This paper aims to introduce Maximal Information Coefficient (MIC) and MINE statistics to social science researchers as effective EDA tools for big social data. By comparing the MIC scores of variable pairs in a multivariate dataset, we can easily detect the pairs with significant relationships that are worth to investigate in detail. In the case study, we used MIC/MINE statistics to explore a large-scale historical data of country-level indicators to find indicators which are strongly (and unexpectedly) related to two target indicators – *(i)* GDP per capita and *(ii)* Corruption Perceptions Index (CPI). We found that using MIC and MINE statistics provided useful information for EDA complementing the traditional Pearson's correlation. They also revealed several intriguing relationships between variables such as a relationship between CPI and time to import (days). With a relatively high MIC (rank 9, 0.594) and an insignificant $|\rho|$ (rank 62, |-0.608|) in 2014, it is expected to be a noticeably nonlinear relationship, whereas the reason "why it is" could be a future research topic in social sciences.

Overall, the main contribution of this paper is twofold: an experience-based case study of using MIC and MINE statistics for exploring big social data and general guidelines of employing these statistics for EDA based on what we learned from the case study.

2 MIC and MINE Statistics

As discussed earlier, MIC is a nonlinear correlation measure which is able to capture various types of significant relationships and allows us to compare the strength across all of them. The basic idea of MIC is that "if a relationship exists between two variables, then a grid can be drawn on their scatter plot to partition the data and encapsulate that relationship".

The following process is used to calculate the MIC of a bivariate data [8].

1. Create a scatter plot of the bivariate data.
2. Explore all possible grids placed on the scatter plot such that $n_x n_y < B$ where n_x and n_y are the number of the partition bins of the x- and y-axis respectively and B, called a maximal grid resolution, is a function of the number of data points n. The authors of MIC suggested B equal to $n^{0.6}$.
3. Calculate mutual information, I(X;Y), of each grid resolution (n_x-by-n_y) and partition placement (the position where a grid is placed) by

$$I(X;Y) = \sum_{y \in Y} \sum_{x \in X} p(x,y) log \frac{p(x,y)}{p(x)p(y)} \quad (1)$$

where X and Y are discrete random variables by the partitions of x- and y-axis, $p(x,y)$ is a joint probability distribution function of X and Y, and $p(x)$, $p(y)$ are marginal probability distribution functions of X and Y, respectively.

4. Find the highest mutual information score for each resolution (n_x-by-n_y). Then normalise it using Eq. (2) to obtain a modified value between 0 and 1 and fill the value in a characteristic matrix $M = (m_{n_x,n_y})$.

$$m_{n_x,n_y} = \frac{I^*(n_x, n_y)}{log_2(min\{n_x, n_y\})} \tag{2}$$

   where $I^*(n_x, n_y)$ is the best mutual information of the grid resolution n_x-by-n_y among all partition placements.
5. The statistic MIC is the maximum value in the characteristic matrix M.

MIC between two variables ranges from 0 to 1. It tends to 0 for a pair of statistically independent variables, whereas it tends to 1 for a variable pair forming a noiseless relationship, or a superposition of noiseless functional relationships. Also, by simulation, MIC approximately equals to the coefficient of determination r^2 relative to the respective noiseless function. In the case of non-function, MIC scores degrade when noise is added.

Additionally, four MINE statistics computed based on MIC and the characteristic matrix M are defined as follows [8, Supporting material]:

1. **Measure of nonlinearity** – Since MIC is general and roughly equals to r^2 on functional relationships, a measure of *nonlinearity* can be defined by MIC$-\rho^2$, where ρ denotes the Pearson's Correlation. If both MIC and ρ of a variable pair are high (i.e., MIC$-\rho^2 \approx 0$), the relationship is strongly linear. In contrast, in the case of high MIC but low ρ (i.e., large MIC$-\rho^2$), the relationship is strong but not linear.
2. **Maximum Asymmetry Score (MAS)** – MAS is helpful for detecting *deviation from monotonicity* as well as periodic relationships with unknown frequencies that vary over time. It is computed from cells in the characteristic matrix M using the equation below.

$$MAS = \max_{n_x n_y < B} |m_{n_x,n_y} - m_{n_y,n_x}| \tag{3}$$

   MAS is never greater than MIC. The higher the periodic frequencies are, the larger the MAS is.
3. **Maximum Edge Value (MEV)** – MEV is defined by

$$MEV = \max_{n_x n_y < B} \{m_{n_x,n_y} : n_x = 2 \text{ or } n_y = 2\}. \tag{4}$$

   It measures the degree to which the dataset appears to be sampled from a *continuous* function (i.e., closeness to being a function). Since MEV is chosen from a cell in the characteristic matrix, MEV$\leq$MIC.
4. **Minimum Cell Number (MCN)** – MCN is defined as

$$MCN(\epsilon) = \min_{n_x n_y < B} \{log_2(n_x n_y) : m_{n_x,n_y} \geq (1-\epsilon)MIC\}. \tag{5}$$

   It measures the *complexity* of the relationship, in terms of the number of cells (n_x-by-n_y) required to reach the MIC score. For noiseless functions, ϵ should

Table 1. Eight data sources and the numbers of collected indicators

Data source	Description / Institution	No. of indicators
WDI[1]	World Bank Development Index **(Multidisciplinary)**	1,421
ILO[2]	International **Labour** Organization	1,122
OECD[3]	The Organisation for Economic Co-operation and Development **(Multidisciplinary)**	473
WTO[4]	World **Trade** Organization	107
WHO[5]	World **Health** Organization	93
Fund for peace[6]	**Failed State Index** and its relevant measures	14
ITU[7]	The International **Telecommunication** Union	7
Transparency[8]	**Corruption Perceptions Index**	2
Total number of indicators		**3,239**

[1] http://data.worldbank.org/indicator
[2] http://www.ilo.org/ilostat
[3] https://data.oecd.org/
[4] http://stat.wto.org/Home/WSDBHome.aspx
[5] http://www.who.int/gho/en/
[6] http://fsi.fundforpeace.org/
[7] http://www.itu.int/en/ITU-D/Statistics/
[8] http://www.transparency.org/

be set to 0. Otherwise, to provide robustness, the ϵ parameter is suggested to be a function of the MIC in question; for example, $\epsilon = 1-\text{MIC}$. In any case, from Eq. (5), the range of MCN is $[2, log_2(B))$.

In summary, MINE statistics tell us other characteristics of the relationship – nonlinearity (MIC $- \rho^2$), non-monotonicity (MAS), continuity (MEV), and complexity (MCN) – apart from the relationship strength reflected by MIC.

3 A Case Study on Big Social Data

We conducted an interdisciplinary analysis of a large-scale historical data of country-level indicators. The objective of the case study is to find country-level indicators which are strongly and unexpectedly related to two target indicators – *(i)* GDP per capita (current US$) and *(ii)* Corruption Perceptions Index.

3.1 Dataset and Methodology

We collected country-level indicator data from eight trustworthy sources (listed in Table 1) which focus mainly on different aspects of countries such as agriculture, education, energy, finance, health, society, and technology. Regarding the target indicators, we chose GDP per capita (current US$) made available by

World Bank Development Index (WDI) and Corruption Perceptions Index by Transparency International [11]. We considered the relationships between these target indicators in a specific year and the values of other indicators in the previous year because these relationships will be helpful for one-year forecasting which we plan to do in the future. Two iterations of analysis were conducted. The first iteration studied the relationships between the target indicators in 2014 and other indicators, sometimes called explanatory variables, in 2013. The second iteration collectively studied the target variables from 2007 to 2014 and the explanatory variables in the previous year of target records.

MIC/MINE statistics can help us grasp lots of variable pairs in this large-scale dataset. However, computing MIC/MINE exhaustively takes very long time. So, we decided to use the approximate calculation software of MIC/MINE provided at http://www.exploredata.net. The software was implemented following *ApproxMaxMI* algorithm proposed by the authors of MIC/MINE.

3.2 Results and Interpretations

This section reports and interprets interesting EDA results from both iterations. Table 2 and Appendix 1's Table 4 list top 20 relationships (ranked by MIC score) between the target variable GDP per capita and explanatory variables in the first and second iteration respectively, while Appendix 1's Tables 3 and 5 present the same results but for Corruption Perceptions Index. The first and the last columns of these tables present the ranks of the explanatory variables with respect to their MIC and $|\rho|$, respectively. In addition, abbreviations used in these four tables (and following figures) are listed at[1].

To begin with, we can notice that there are two groups of indicators in Table 2. The first group consists of indicators originally related to the target, GDP per capita, by their definitions such as "GNI per capita" and "GDP per capita PPP". These indicators yielded really high MIC score and low MIC $- \rho^2$ meaning that they are strong linear relationships. In contrast, the other indicators such as "Fixed telephone subscriptions", "Agriculture; value added", and "Pupil-teacher ratio; primary" have surprised us because we could not imagine the obvious reasons why they are related to GDP per capita. Moreover, the relationships are not strongly linear as their $|\rho|$ measures were not so large.

To investigate more, we created interesting scatter plots from some of the top 20 relationships using Tableau [10] as shown in Fig. 1. Each dot represents indicator data of a country and its colour signifies the continent of that country. The association between GDP per capita and "Adjusted net national income per capita" (rank 1 in Table 2) displayed in Fig. 1(a) is a clear example of a linear relationship between two variables. On the contrary, the relationship between our target variable and "Agriculture; value added (% of GDP)" (rank 15), shown in Fig. 1(c), looks similar to a rectangular hyperbola. That is why its nonlinearity

[1] Abbreviations used in the tables and the figures: cap. = capita; c.US$ = current US dollar; c.int$ = current international dollar; 2011.int$ = constant 2011 international dollar; inhab. = inhabitants; consump. = consumption; pop. = population.

Table 2. Top 20 relationships ranked by MIC score for the target variable GDP per capita (current US$) (year: 2013 → 2014).

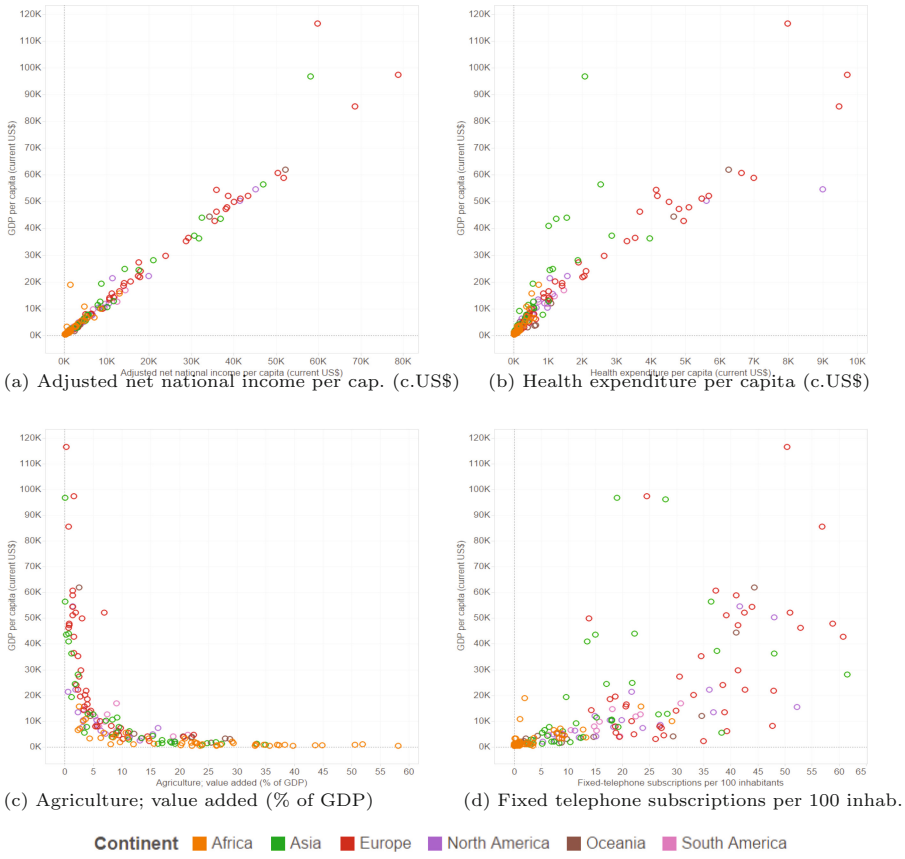
M #	Indicators (Unit)	MIC	MIC $-\rho^2$	MAS	MEV	MCN	ρ	$\lvert\rho\rvert$ #
1	Adjusted net national income per cap. (c.US$)	1.000	0.039	0.027	1.000	4.000	0.980	4
1	GDP per capita (c.US$)	1.000	0.003	0.027	1.000	3.585	0.999	1
1	GNI per capita; Atlas method (c.US$)	1.000	0.034	0.073	1.000	4.000	0.983	2
1	GNI per capita; PPP (2011.int$)	1.000	0.138	0.066	1.000	4.000	0.928	7
5	GNI per capita (2005.US$)	1.000	0.051	0.038	0.987	4.322	0.974	6
6	GNI per capita; PPP (c.int$)	0.981	0.134	0.036	0.971	4.585	0.921	9
7	Adj. net national income per cap. (2005.US$)	0.968	0.015	0.050	0.955	4.322	0.976	5
8	GDP per capita; PPP (c.int$)	0.964	0.108	0.028	0.964	4.585	0.925	8
8	GDP per capita; PPP (2011.int$)	0.964	0.117	0.022	0.964	4.585	0.920	10
10	GDP per capita (2005.US$)	0.961	−0.001	0.056	0.961	4.585	0.981	3
11	Health expenditure per cap.; PPP (2011.int$)	0.906	0.096	0.052	0.896	4.585	0.900	12
12	Household final consump. per cap. (2005.US$)	0.903	0.123	0.072	0.903	4.322	0.883	14
13	Health expenditure per capita (c.US$)	0.901	0.086	0.023	0.883	4.585	0.903	11
14	GDP per person employed (2011.PPP$)	0.893	0.101	0.051	0.893	4.459	0.890	13
15	Agriculture; value added (% of GDP)	0.817	0.530	0.045	0.817	4.459	−0.536	76
16	Fixed-telephone subscriptions per 100 inhab	0.815	0.381	0.156	0.800	4.459	0.658	38
17	Fixed telephone subscriptions per 100 people	0.787	0.315	0.120	0.787	4.585	0.687	35
18	Percentage of Individuals using the Internet	0.785	0.228	0.025	0.766	4.459	0.747	22
19	Electric power consumption (kWh per capita)	0.780	0.301	0.061	0.768	4.322	0.692	34
20	Pupil-teacher ratio; primary	0.770	0.491	0.070	0.754	4.322	−0.529	82

scores, MIC $-\rho^2$, is distinctively higher than other indicators. This demonstrates the ability of MIC to detect nonlinear relationships in big data. If we considered only the rank from $|\rho|$, we would miss interesting relationships.

The relationship between GDP per capita and "Fixed-telephone subscriptions per 100 inhabitants" (rank 16), shown in Fig. 1(d), got the highest MAS value in Table 2. It means that this relationship had the highest degree of non-monotonicity among the 20 relationships. It might be because we cannot decide definitely whether the right half of Fig. 1(d) shows the increasing or decreasing trend. In other words, it is unclear to state that this relationship is monotonic.

Comprehensive EDA not only tells us the shape of data but also guides us the proper means for further data analysis. For example, in Fig. 1(b) showing the relationship between GDP per capita and "Health expenditure per capita (current US$)" (rank 13), when the value on X-axis grows larger, the line of green dots (countries in Asia) is separate from the others and makes the plot seem like a superposition of two linear relationships. It suggests that if we create an OLS regression model for GDP per capita, we should have a linear term for health expenditure per capita and another term to control for location and capture different behaviours between continents.

With regard to another target variable – Corruption Perceptions Index (CPI), several measures in its top 20 relationships (ranked by MIC score) are also the top 20 s of GDP per capita. They are probably correlated as a group or a chain reaction. However, we found that the top 20 MIC scores of CPI are markedly lower than the ones of GDP per capita. This might be because CPI was not



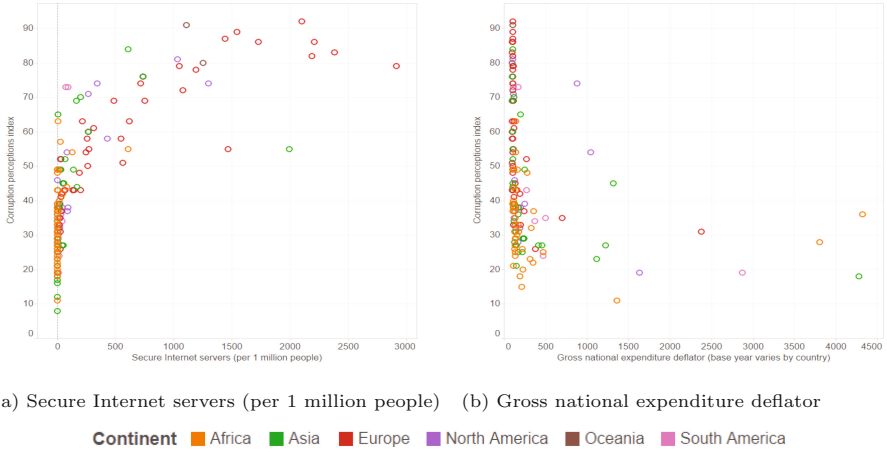
(a) Adjusted net national income per cap. (c.US$) (b) Health expenditure per capita (c.US$)

(c) Agriculture; value added (% of GDP) (d) Fixed telephone subscriptions per 100 inhab.

Fig. 1. Scatter plots between GDP per capita (current US$) in 2014 (Y-axis) and some explanatory indicators in 2013 (X-axis).

computed from hard empirical data, but from surveys and assessments of corruption, gathered by various reliable institutions [11] causing higher uncertainty in CPI than GDP per capita.

Figure 2 shows two interesting scatter plots selected from the top 20 relationships of CPI. By definition, the lower the CPI is, the higher perceptions of corruption exist in that country. Figure 2(a) which looks like a linear or a logarithm function shows the relationship between CPI and the number of "Secure Internet servers" (rank 4). This relationship is unexpected because the number of secure Internet servers does not have a direct link to corruption. So, it is intriguing to figure out the reason behind this association. Besides, Fig. 2(b) illustrates an association between "Gross national expenditure deflator (base year varies by country)" and CPI (rank 10). This plot is interesting because most of the dots form a vertical straight line showing no relationships between the variables. The ρ of this relationship is close to zero because it shows neither increasing nor



(a) Secure Internet servers (per 1 million people) (b) Gross national expenditure deflator

Fig. 2. Scatter plots between Corruption Perceptions Index in 2014 (Y-axis) and some explanatory indicators in 2013 (X-axis).

decreasing trend. However, this relationship obtains relatively high MIC score and it results in a large MIC $- \rho^2$ even if the plot looks like a straight line overall. This case suggests that we should be careful while interpreting the statistics. Please see more discussion on this issue in Sect. 4.

In iteration 2, we collectively computed MIC score and MINE statistics for the target indicators from 2007 to 2014. We found that MIC scores in iteration 2 were generally less than the scores in iteration 1. Also, many indicators from the top 20 s of iteration 1 did not appear in the top 20 s of iteration 2 as those relationships were strong for only a specific year. To get a high MIC in iteration 2, the indicators must correlate with the target for almost all of the analysed years which is quite difficult except for the indicators originally related by definitions.

On the whole, MEV (the continuity measure) and MCN (the complexity measure) did not play a significant role in exploring our interdisciplinary dataset in both iterations. We notice that MEV usually approximated to MIC. After creating a plot between MIC and MEV from all pairs of variables, we found that the plot was obviously linear with MIC $= 0.9779$ and $\rho = 0.9978$. The line of best fit is MEV $= 0.9953$MIC-0.0042 with $r^2 = 0.9957$. In other words, MEV $\approx$ MIC in this case study. Actually, MEV is not always a redundant measure of MIC. For instance, a noiseless circle with $n = 1,000$ yields MIC $= 0.71$ and MEV $= 0.32$ [8, Supporting material]. Nonetheless, this kind of relationships does not exist in our real-world dataset with more than thousand pairs of variables.

Considering MCN within each iteration, we cannot observe significant differences in complexity between scatter plots whose MCNs are unequal. However, the MCN values in iteration 1 are noticeably less than the ones in iteration 2. The ranges of MCN in iteration 1 and 2 are [3.58, 4.58] and [5.32, 6.49], respectively. This might be because iteration 2 has more dots than iteration 1. Hence, a more complex grid is required to capture the real-world relationships and reach

the MIC scores, while the upper bound of MCN, which is $\log(B) = log(n^{0.6})$, allows MCN to grow larger as n increases in iteration 2.

4 Discussions

Since MIC was intentionally designed for rapid exploration of multidimensional datasets, it performed well in the case study, capturing not only linear but also quadratic, logarithmic, and hyperbolic relationships. It seems that MIC is a perfect tool for EDA; however, it still has some limitations.

- Unlike Pearson's correlation (ρ), MIC does not imply the tone of relationship, i.e., whether it is increasing or decreasing.
- MIC, as well as ρ, supports only numeric data. We noticed different behaviours of continents in the case study not because of MIC but because of the different colours we set in the scatter plots.
- MIC is sensitive to noises and sometimes detects false-positive relationships. In other words, it sometimes gives a variable pair with no veritable association too high MIC score [9]. This might be the reason for the relatively high MIC score (as well as the high nonlinearity score) of Fig. 2(b). In our case study, even the lowest MIC value from both iterations is 0.0868 (8.68% from the perfect score), while the lowest $|\rho|$ is only 0.0003 (0.03%).

Based on what we learned from the case study and the limitations, we provide general guidelines for using MIC/MINE statistics to explore data as follows:

- Analyse MIC/MINE statistics with ρ to get more comprehensive information for the sketch of the relationship. Note that MIC does not totally replace ρ.
- Do not conclude the exact shape of the relationship from only the measures until seeing the scatter plot.
- Use MIC/MINE and ρ to help us select some (interesting) relationships to investigate graphically. The easiest way is selecting the relationships with a relatively high MIC score as we did in the case study. Another recommended way is plotting a scatterplot matrix (SPLOM) of the measures (as we show in Appendix 2) and manually selecting the relationships located in interesting areas of subplots in the SPLOM such as high MIC and MIC$-\rho^2$ or selecting the outlier dots compared to other relationships.
- Do not forget to use other visual variables of scatter plot such as shapes and colours to represent categorical features.

5 Related Work

Exploratory data analysis in social sciences – Several measures have been used as correlation measures in quantitative social research such as Lambda coefficient for nominal-level data, Kendall's and Spearman's correlation for ordinal-level data, and Pearson's correlation (ρ) for interval-level and ratio-level data. So far, however, there has been little discussion about nonlinear associations

in social science research, even in recent research methodology textbooks [2,4]. Only a few papers turn to concern nonlinear associations such as the paper by Boutyline and Vaisey which used ρ for numeric variables and conducted general (not only linear) dependency analysis in an appendix using mutual information to confirm the results [3]. We hope that our paper will build more awareness of nonlinear relationships in social data among social science researchers.

Applications of MIC/MINE statistics – MIC has been employed in several data analysis studies. For example, Tan et al. conducted model selection based on the MIC between regression residuals and explanatory variables [6]. Zhao et al. combined MIC with affinity propagation clustering to perform feature selection [12]. MIC was also deployed to support construction and analysis of biological networks as it could capture a wide range of relationships [1,5,7]. Nonetheless, we have not seen any methodology or application papers using MINE statistics in their work. Moreover, even if a few papers tried applying MIC on social science datasets as small examples [8,12], to the best of our knowledge, this paper is the first study that places great emphasis on interpreting the meaning of MIC and MINE statistics for practical use in social science research.

6 Conclusion

This paper aims to use a case study of analysing a large-scale country-level indicator dataset to introduce MIC and MINE statistics for exploring big social data. Based on the case study, MIC was a powerful nonlinear correlation measure which detected several unexpected relationships and complemented the traditional linear correlation (Pearson's ρ). Also, two MINE statistics, MAS (for non-monotonicity) and MIC $- \rho^2$ (for nonlinearity), helped us imagine the scatter plots more accurately, while the other two statistics – MEV and MCN – did not show their full discrimination power in our big real-world dataset. To overcome some limitations of MIC/MINE statistics, we also provide general guidelines for using the statistics in this paper. For future work, we plan to conduct interdisciplinary forecasting of target indicators with the aid of MIC/MINE statistics. We believe that it could yield accurate prediction results as all related domains of country-level data are included in the model, not only the specific domain of target indicators.

Appendix 1: Supplementary Results

Table 2 in Sect. 3.2 and Appendix 1's Table 4 list top 20 relationships (ranked by MIC score) between the target variable GDP per capita and explanatory variables in the first and second iteration respectively, while Appendix 1's Tables 3 and 5 present the same results but for Corruption Perceptions Index. Each of the tables has nine columns. The first and the last columns present the ranks of the explanatory variables with respect to their MIC and $|\rho|$, respectively, as

Table 3. Top 20 relationships ranked by MIC score for the target variable Corruption Perceptions Index (year: 2013 → 2014).

M #	Indicators (Unit)	MIC	MIC $-\rho^2$	MAS	MEV	MCN	ρ	$\lvert\rho\rvert$ #
1	Corruption Perceptions Index	0.918	−0.074	0.041	0.918	4.000	0.996	1
2	Household final consump. per cap. (2005.US$)	0.721	0.015	0.114	0.701	4.170	0.840	7
3	Adj. net national income per cap. (2005.US$)	0.717	−0.001	0.113	0.700	4.170	0.847	4
4	Secure Internet servers (per 1 million people)	0.655	0.138	0.054	0.655	4.322	0.719	31
5	GNI per capita (2005.US$)	0.649	−0.058	0.050	0.623	4.170	0.840	6
6	Legitimacy of the State	0.645	−0.161	0.076	0.645	4.322	−0.898	2
7	Failed States Index Rank	0.602	−0.089	0.079	0.602	4.322	0.831	8
8	Failed States Index Total	0.601	−0.168	0.078	0.601	4.322	−0.877	3
9	Time to import (days)	0.594	0.224	0.117	0.594	4.322	−0.608	62
10	Gross national expenditure deflator	0.590	0.519	0.139	0.590	4.170	−0.267	261
11	GDP per capita (2005.US$)	0.587	−0.056	0.061	0.587	4.322	0.802	14
12	Health expenditure per capita (c.US$)	0.585	−0.003	0.066	0.585	4.322	0.767	25
13	Public services	0.585	−0.079	0.049	0.585	4.322	−0.815	11
14	Burden of customs procedure; WEF (1 to 7)	0.583	−0.070	0.053	0.583	4.170	0.808	13
15	GNI per capita; PPP (2011.int$)	0.581	0.064	0.046	0.581	4.170	0.719	32
16	Price level ratio of PPP conversion factor (GDP) to market exchange rate	0.581	−0.050	0.121	0.581	3.585	0.794	17
17	Improved water source (% of pop. with access)	0.578	0.234	0.036	0.578	4.322	0.587	70
18	Percentage of Individuals using the Internet	0.577	−0.033	0.071	0.577	4.322	0.781	22
19	Fixed-broadband subscriptions per 100 inhab	0.577	−0.042	0.131	0.577	4.322	0.786	20
20	Adj. net national income per cap. (c.US$)	0.575	−0.090	0.062	0.562	4.322	0.815	10

Table 4. Top 20 relationships ranked by MIC score for the target variable GDP per capita (current US$) (year: 2006–2013 → 2007–2014).

M #	Indicators (Unit)	MIC	MIC $-\rho^2$	MAS	MEV	MCN	ρ	$\lvert\rho\rvert$ #
1	GDP per capita (c.US$)	0.993	0.010	0.020	0.992	6.492	0.992	1
2	GNI per capita; Atlas method (c.US$)	0.974	0.015	0.019	0.973	6.459	0.979	3
3	Adjusted net national income per cap. (c.US$)	0.954	−0.001	0.020	0.954	6.358	0.977	5
4	GNI per capita (2005.US$)	0.946	−0.009	0.026	0.944	6.209	0.977	4
5	GNI per capita; PPP (2011.int$)	0.943	0.100	0.022	0.943	6.248	0.918	7
6	GDP per capita (2005.US$)	0.931	−0.036	0.026	0.931	6.459	0.984	2
7	GDP per capita; PPP (c.int$)	0.930	0.118	0.026	0.929	6.459	0.901	10
8	GDP per capita; PPP (2011.int$)	0.926	0.109	0.040	0.925	6.459	0.904	8
9	Adj. net national income per cap. (2005.US$)	0.925	−0.025	0.018	0.925	6.170	0.975	6
10	GNI per capita; PPP (c.int$)	0.918	0.110	0.018	0.917	6.459	0.899	11
11	Household final consump. per cap. (2005.US$)	0.863	0.047	0.028	0.863	6.209	0.903	9
12	Health expenditure per cap. (c.US$)	0.838	0.066	0.023	0.838	6.426	0.879	12
13	Health expenditure per cap.; PPP (2011.int$)	0.837	0.080	0.020	0.837	6.426	0.870	13
14	GDP per person employed (2011.PPP$)	0.821	0.074	0.026	0.819	6.392	0.864	14
15	Agriculture; value added (% of GDP)	0.780	0.494	0.056	0.780	6.392	−0.534	65
16	Electric power consumption (kWh per capita)	0.766	0.228	0.055	0.766	6.170	0.733	27
17	Automated teller machines per 100000 adults	0.713	0.291	0.077	0.713	6.358	0.650	42
18	Energy use (kg of oil equivalent per capita)	0.711	0.215	0.052	0.709	6.248	0.704	34
19	Air and GHG emissions - CO_2 (tonne_cap)	0.683	0.229	0.073	0.683	6.000	0.674	39
20	Fixed telephone subscriptions per 100 people	0.682	0.115	0.029	0.682	6.459	0.753	22

Table 5. Top 20 relationships ranked by MIC score for the target variable Corruption Perceptions Index (year: 2006–2013 → 2007–2014).

M #	Indicators (Unit)	MIC	MIC $-\rho^2$	MAS	MEV	MCN	ρ	$\lvert\rho\rvert$ #
1	Corruption Perceptions Index	0.865	−0.116	0.019	0.865	5.907	0.990	1
2	Household final consump. per cap. (2005.US$)	0.623	−0.101	0.042	0.620	6.044	0.851	5
3	Adj. net national income per cap. (2005.US$)	0.613	−0.122	0.049	0.613	5.907	0.857	4
4	Secure Internet servers (per 1 million people)	0.613	0.151	0.049	0.613	6.209	0.680	38
5	GNI per capita (2005.US$)	0.578	−0.142	0.036	0.578	5.954	0.848	6
6	Adj. net national income per cap. (c.US$)	0.568	−0.113	0.068	0.568	6.170	0.825	10
7	Public services	0.561	−0.142	0.046	0.557	6.209	−0.839	7
8	GDP per capita (2005.US$)	0.560	−0.103	0.059	0.560	6.248	0.814	13
9	Electric power consumption (kWh per capita)	0.554	0.142	0.084	0.551	6.044	0.642	49
10	Health expenditure per capita (c.US$)	0.553	−0.053	0.044	0.551	6.248	0.778	23
11	Failed States Index Total	0.551	−0.214	0.028	0.549	6.209	−0.875	3
12	Legitimacy of the State	0.540	−0.226	0.046	0.539	6.209	−0.875	2
13	Health expenditure per cap.; PPP (2011.int$)	0.536	−0.104	0.077	0.533	6.248	0.800	19
14	Agriculture; value added (% of GDP)	0.531	0.178	0.086	0.529	6.170	−0.595	64
15	GNI per capita; Atlas method (c.US$)	0.530	−0.128	0.040	0.530	6.209	0.811	16
16	Failed States Index Rank	0.529	−0.126	0.045	0.529	5.322	0.809	17
17	Percentage of Individuals using the Internet	0.520	−0.153	0.052	0.518	6.248	0.821	12
18	GNI per capita; PPP (2011.int$)	0.517	−0.008	0.045	0.517	6.087	0.725	31
19	GDP per capita (c.US$)	0.515	−0.111	0.042	0.513	6.248	0.791	21
20	Internet users (per 100 people)	0.514	−0.160	0.050	0.514	6.248	0.821	11

both measures reflect the relationship strengths (in different ways). The second column shows the variables' names and units of measurement. The third column reports the MIC scores of the relationships, whereas the fourth to seventh list their MINE statistics. The eighth column shows the Pearson's ρ scores ranging from -1 to 1. Lastly, abbreviations used in these tables are listed at[11].

Appendix 2: Relationships among the statistics

A scatterplot matrix (SPLOM) of measures in the case study (MIC, MIC-ρ^2, MAS, MEV, MCN, ρ, and $|\rho|$) is shown in Fig. 3. We can notice the linear relationship between MIC and MEV as well as the separation of MCN values of both iterations in this scatterplot matrix.

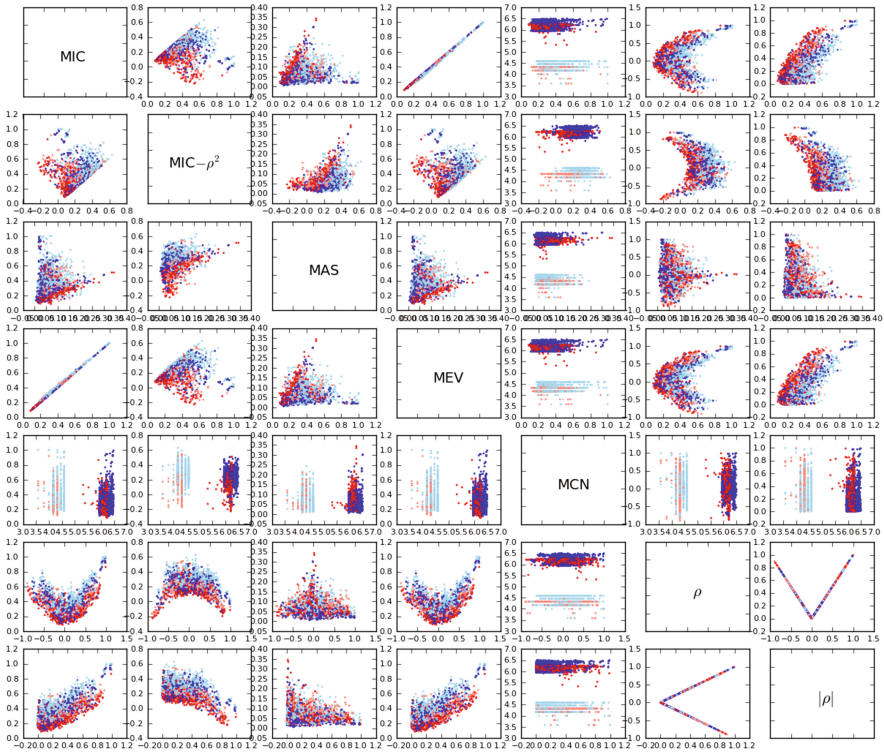


Fig. 3. A scatterplot matrix (SPLOM) of measures in the case study. Each dot represents a relationship of a variable pair. The blue dots are for the target variable GDP per capita (current US$). The red dots are for the target variable Corruption Perceptions Index (CPI). The light colours represent the relationship pairs from the first iteration, while the dark colours represent the relationship pairs from the second iteration.

References

1. Akhand, M., Nandi, R., Amran, S., Murase, K.: Gene regulatory network inference incorporating maximal information coefficient into minimal redundancy network. ICEEICT **2015**, 1–4 (2015)
2. Bhattacherjee, A.: Social science research: principles, methods, and practices (2012)
3. Boutyline, A., Vaisey, S.: Belief network analysis: a relational approach to understanding the structure of attitudes. Am. J. Sociol. **122**(5), 1371–1447 (2017)
4. Neuman, L.W.: Social research methods: qualitative and quantitative approaches, 7th edn. Pearson Education Limited, Essex (2014)
5. Paul, A.K., Shill, P.C.: Reconstruction of gene network through backward elimination based information-theoretic inference with maximal information coefficient. In: icIVPR 2017, pp. 1–5 (2017)
6. Qiuheng, T., Jiang, H., Yiming, D.: Model selection method based on maximal information coefficient of residuals. Acta Math. Sci. **34**(2), 579–592 (2014)
7. Rau, C., Wisniewski, N., Orozco, L., Bennett, B., Weiss, J., Lusis, A.: Maximal information component analysis: a novel non-linear network analysis method. Front. Genet. **4**, 28 (2013)
8. Reshef, D.N., Reshef, D.N., Reshef, Y.A., Finucane, H.K., Grossman, S.R., Mcvean, G., Turnbaugh, P.J., Lander, E.S., Mitzenmacher, M., Sabeti, P.C.: Detecting novel associations in large data sets. Science **334**, 1518–1524 (2011)
9. Simon, N., Tibshirani, R.: Comment on "Detecting novel associations in large data sets". In: Reshef., D.N., et al. Science, 16 December 2011. ArXiv e-prints (2014)
10. Tableau Software: Answer questions as fast as you can think of them with tableau (2016). http://www.tableau.com/trial/tableau-software
11. Transparency International: Corruption perceptions index 2016: Frequently asked questions (2017). https://www.transparency.org/news/feature/corruption_perceptions_index_2016
12. Zhao, X., Deng, W., Shi, Y.: Feature selection with attributes clustering by maximal information coefficient. Procedia Comput. Sci. **17**, 70–79 (2013)