An Acad Bras Cienc (2023) 95(1): e20211422 DOI 10.1590/0001-3765202320211422
Anais da Academia Brasileira de Ciências | *Annals of the Brazilian Academy of Sciences*
Printed ISSN 0001-3765 I Online ISSN 1678-2690
www.scielo.br/aabc | www.fb.com/aabcjournal

ENGINEERING SCIENCES

Econometric model of iron ore through principal component analysis and multiple linear regression

BÁRBARA ISABELA DA SILVA CAMPOS, GISELE C.A. LOPES, PHILIPE S.C. DE CASTRO, TATIANA B. DOS SANTOS & FELIPE R. SOUZA

Abstract: Price of iron ore is affected by instabilities of microeconomic balance between supply and demand. Periods of equilibrium adjustment result in huge swings, growth or global recession. They also impact the viability of mineral enterprises and generate consequences to important global economic scenarios. This research aims to evaluate the market variables capable of influencing the price of iron ore through multivariate statistical techniques. Principal component analysis and multiple linear regression, booth multivariate statistical techniques were used. The studied variables were rate export of iron ore and concentrates from Brazil, steel production from China, steel production from Japan, production from Europe, steel production from the United States, steel production from India, steel price, coal price, China's Construction Gross Domestic Production, United States construction index, oil price and global oil production. First three components explained 89.12% of the variability of the data matrix. Multiple linear regression highlighted the significance of five variables. They are export iron ore from Brazil, steel production from China, price of coal, steel production from India and price of steel.

Key words: multivariate statistics, iron ore, principal component analysis, multiple linear regression, iron ore, mineral economy.

INTRODUCTION

Iron ore pricing system has been updated over the past few decades. For a long time, global iron ore market has presented stability in commodity demand. An agreement between buyers and producers created a benchmark in annual price transactions in the 1970s (Gaggiato 2014). Long-term agreement between major steel makers and mining companies determined the iron ore price for that period. This pricing system is known as the Benchmark Price System or Reference Price System.

An important aspect responsible to change the ore price scenario in global market was the process of growth of mining industry, in relation to the stagnation of steel industry and the shift of the world buyer center from Europe and Japan to China. Gradually China is increasing its demand for iron ore and, consequently, the country contributed to increasing the price of the commodity.

Economic viability of mining enterprises directly depends on market price, quality and extent of the reserves. The increasing of iron ore prices made possible the entry of new suppliers in the market, which offered products with quality characteristics different from the traditional commercialized ores.

Then, new trading systems and a growing market negotiated via spot prices with daily variations became a reality and came to compete with the reference price system (Gaggiato 2014). Benchmark Price System has been ended and replaced by a others systems, which are generally based on the spot prices traded on the Chinese market.

The use of statistics focused on economic context is named econometrics. Econometrics consists of the application of mathematical and statistical methods to problems of economics in order to verify hypotheses and predict future trends (Hoffmann 2016). Use of an accurate methodology is essential to anticipate more assertive economic measures based on market fluctuations. The prerequisites for econometrics are basic concepts of statistical estimations including on sampling procedures, estimators, confidence intervals and hypothesis tests, non-parametric statistics (Biage 2012). Regression analysis is one the most used econometric technique. In the present research, the multiple linear regression method was adopted.

Studies conducted by Wårell investigated the impact on the econometric model in the price regime change of iron ore based on monthly data from different periods between 2003 and 2017 (Wårell 2018). The author used linear multiple regression in its analyses and he presented important conclusions, such as the great influence of the Chinese GDP growth on the price of iron ore, considering the analyzed period.

Most of the researches related to price prevision are grouded on economic principle named “Ceteris Paribus”. This principle assumes that the economic instability factors are constant and defined by a average over a time interval. The researchers usually consider that the instability is constant because is more ease to modeling and calibrating the variables. According to (Alameer 2020) and (Li et al. 2020), it is more accurate to select variables capable of impacting the price of the commodity using principal component analysis. The system developed by (Alameer 2020) based on neural networks showed greater accuracy than the time series model for coal.

In this research, a dataset was built and multivariate statistical methods were performed in this research in order to carry out a econometric analysis and investigate the influence of selected variables in iron ore price, considering the years 1991 to 2020. A predictive model was created using appropriate techniques. Statistics is an excellent tool for data collection and data analysis. These methods are very widespread and support scientific research in different areas.

MATERIALS AND METHODS

R Software Version 4.1.2 (Team 2013) was used to carry out principal components analysis (PCA) and multiple linear regression (MLR) in the dataset. The package used for booth technique application was ‘stats’, which is part of R Software (Team 2013). The dataset was constructed and it was verified the need of data standadization througth boxplot analysis. Exploratory analysis and correlation matrix was defined in order to understanding and identifying the relationship between the variables of the data.

Barlett’s Test was carried out in order to verify if there is sufficient correction between the data for the application of multivariate statistical techniques (Bartlett 1951). Then, PCA was carried out with the due of understand the interdependence between the variables. Kaiser’s criterion determined the number of principal components retained in the analysis (Kaiser 1970).

Principal component analysis was used to determining the variables that did not have a great impact on the iron ore price variable (Hotelling 1933). These variables were removed before performing MRL analysis (Hair Jr Joseph et al. 2009) . Multivariate outliers were determined in order to to improve the obtained result (Filzmoser 2004). The values identified were removed from the dataset and multiple regression model was defined and proposed.

In order to validate the model, the residual values, linearity, residual homoscedasticity, residual normality and model accuracy was obtained a analyzed. The most significant variables for prediction of the dependent variable were determined.

The Dataset

The dataset uses information from the Trading Economics website (Trading 2021). The platform provides accurate information for 232 countries, including historical data for more than 300,000 economic indicators, exchange rates, stock market indices, government bond yields and commodity prices. The data are based on official sources and are regularly checked for inconsistencies. Table I and Table II preset the consolidated dataset used in this research.

Independent variables were used to analyze the influence on the annual average value of the iron ore price (Iron_Price), from 1991 to 2020, in US dollars per dry metric ton. The dependent variables are average annual value of iron ore and concentrated exports from Brazil with 62% content (USD) (Br_Iron_Exp); steel production (t) in China (China_Steel), India (India_Steel), Japan (Japan _Steel), Europe (Euro_Steel) and the United States (USA_Steel); annual average value of prices (USDt) of steel (Steel_Price), coal (Coal_Price) and oil (Oil_Price); construction GDP in China (GDPCCCN) (CNY) (China_GDP); US Construction Index (ICCUS) (%) (USA_Constr) and global oil production (bbld) (Glob_Oil_P-rod).

The variables were selected according to its ability to influence the price of iron ore and they were mainly considered by the World Bank's an Econometric Model of the Iron Ore Industry in 1987. To built this data set, Trading Economics website (Trading 2021) were used.

Multivariate analysis techniques

The statistical methods, regarding the analysis of variables, are divided in two statistical areas: univariate statistics (analysis of variables one by one) and multivariate statistics (joint analysis of the variables) (Vicini 2005). Multivariate statistics allows simultaneous investigation of multiple variables, considering each sample element. All variables should be random and correlated and this type of technique provides way of evaluating information, which cannot be obtained and interpreted with the use of uninivariate statistical techniques (Hair Jr Joseph et al. 2009). Principal component analysis (PCA) and multiple linear regression (MLR) are multivariate statistical techniques.

Determination of multivariate outliers

An outlier is an observation so different from other observations. It promotes suspicions that was generated by a distinct mechanism (Enderlein 1987). According to (Krige & Magri 1982), the outliers approach is pointed out by:

Table I. Dataset 1993 - 2020.

Year	Brazilian Iron Ore Exports of 62% (USD Millions)	Chinese Steel Production (kt)	Japanese Steel Production (kt)	European Steel Production (kt)	American Steel Production (kt)	Indian Steel Production (kt)
1991	2,600.21	70,564	109,648	137,408	79,331	17,100
1992	2,381.29	80,037	98,133	132,200	83,102	18,165
1993	2,256.84	89,453	99,624	132,275	87,007	18,155
1994	2,293.93	93,143	98,296	138,972	88,855	19,284
1995	2,547.72	93,842	101,639	155,824	93,602	20,768
1996	2,695.15	100,059	98,803	146,681	94,247	23,755
1997	2,846.10	107,899	104,546	159,919	96,705	24,579
1998	3,252.99	114,063	93,548	159,950	97,294	23,480
1999	2,745.95	123,643	94,192	155,522	96,054	24,269
2000	3,048.19	126,317	106,444	163,012	100,711	26,924
2001	2,931.48	145,224	102,867	158,497	89,710	27,291
2002	3,048.80	180,532	107,745	158,437	91,605	28,814
2003	3,455.88	219,449	110,510	160,656	91,339	31,779
2004	4,758.81	277,691	112,718	194,317	98,522	32,626
2005	7,296.58	353,564	112,472	187,531	93,216	45,780
2006	8,948.82	425,100	116,228	198,592	98,539	49,450
2007	10,557.85	492,697	120,203	210,257	98,182	53,468
2008	16,538.47	498,688	118,740	198,229	91,350	57,791
2009	13,246.87	568,877	87,535	138,958	58,195	63,527
2010	28,911.81	623,810	109,599	172,701	80,495	68,321
2011	41,817.19	684,275	107,594	177,468	86,247	72,206
2012	30,989.22	714,939	107,233	168,650	88,695	77,561
2013	32,491.48	800,984	110,594	153,676	86,878	81,299
2014	25,819.03	815,084	110,664	169,349	88,174	86,530
2015	14,076.10	800,529	105,153	166,105	78,916	89,581
2016	13,289.34	804,825	104,709	162,134	78,475	95,475
2017	19,199.15	867,543	104,661	168,548	81,612	101,453
2018	20,220.36	922,798	104,318	167,732	86,607	109,272
2019	21,819.90	993,411	99,283	160,141	87,761	111,245
2020	24,341.36	1,054,429	83,461	135,388	71,311	95,573

Table II. Dataset 1993 - 2020.

Year	Steel Price (USD/t)	Coal Price (USD/t)	PIB-CC-CN (CNY CMH)	ICC-EUA (%)	Oil Price (USD/Bbl)	Global Oil Production (Mbbl/d)	Iron Ore Price (USD/t)
1991	108.40	39.67	1,270.66	-0.30	19.37	60.13	34.76
1992	108.00	38.56	1,415.00	0.84	19.04	60.10	33.10
1993	112.40	31.33	2,266.50	0.93	16.79	60.17	29.09
1994	122.10	32.30	2,964.70	0.13	15.95	61.18	26.47
1995	126.60	39.37	3,728.80	0.25	17.20	62.43	28.38
1996	127.70	38.07	4,387.40	0.69	20.37	63.82	30.00
1997	133.60	35.10	4,621.60	0.48	19.27	65.80	30.15
1998	129.60	29.23	4,985.80	0.83	13.07	67.02	31.00
1999	129.20	25.89	5,172.10	0.93	17.98	65.90	27.59
2000	132.00	26.25	5,522.30	0.23	28.23	68.34	28.79
2001	127.10	32.31	5,931.70	0.39	24.33	67.92	30.03
2002	140.50	27.06	6,465.50	0.05	24.95	67.05	29.31
2003	146.10	27.95	7,490.80	0.88	28.90	69.19	31.95
2004	238.60	56.73	8,694.30	0.74	37.72	72.25	37.90
2005	225.70	50.82	10,400.50	1.12	53.36	73.52	65.00
2006	239.20	52.73	12,450.10	-0.35	64.27	73.10	69.33
2007	233.90	70.09	15,348.00	-0.14	71.16	72.70	122.99
2008	294.30	138.02	18,807.60	-0.91	96.96	73.58	155.99
2009	241.00	76.16	22,681.50	-1.39	61.77	72.39	79.98
2010	285.20	104.60	27,259.30	-0.53	79.03	74.17	145.86
2011	322.20	129.61	32,926.50	0.35	104.05	74.28	167.75
2012	313.60	101.44	36,896.10	0.48	105.01	76.05	128.50
2013	298.30	90.13	40,896.80	1.00	104.07	76.00	135.36
2014	297.00	75.73	45,401.70	0.62	96.25	77.72	96.95
2015	247.60	62.69	47,761.30	0.82	50.79	79.78	55.85
2016	263.00	70.08	51,498.90	0.78	42.84	80.76	58.42
2017	288.60	94.14	57,905.60	0.23	52.81	81.09	71.76
2018	360.40	113.23	65,493.00	-0.14	68.33	82.98	69.75
2019	332.70	82.19	70,904.30	0.73	61.39	82.34	93.85
2020	339.50	61.98	72,995.70	0.47	41.29	75.91	108.92

1. Use of statistics such as probability charts, histograms and scatterplots;
2. Validation of the sample context considering the domain according to its support and neighborhood. To decide if it really is an anomalous value and needs to be modified or removed;
3. Validation of the possibility of human error in the transcription of the sample value by checking the history of the sample;
4. Validate if the outlier belongs to a previously stipulated confidence interval. If the outlier deviates by a factor close to twice the non-outlier value it is appropriate to remove it from the statistics calculations.

The outliers can negatively influence the analysis and interpretation of the data matrix, therefore its identification is necessary. It can be eliminated depending on the purpose of the analysis and the researcher's experience.

Mahalanobis distance is most widely used for multivariate outlier detection. The distance is calculated from the ith sample element into the average of the data, given by Equation 1.

$$MD_i = \sqrt{(x_i - \bar{x})' S^{-1} (x_i - \bar{x})}\,, \tag{1}$$

where x^i is the ith element sample $\bar{x}$ is vector of means (average) and S is the matrix of variances and convariaces of dataset X.

The distance of the sample elements follows chisquare distribution, with p (number of variables) degrees of freedom. Multivariate outliers are defined as measures that exceed a certain amount of the chisquare distribution (Valadares et al. 2012).

Principal component analysis

Principal component analysis (PCA) is a multivariate statistical method capable of explaining the interdependence between the variables and reducing the dimensionally of the data (Varella 2008). The principal components shall ensure variance similar to the original variables so as to accurately represent the information contained. The technique consists of converting the original variables into new variables named principal components. The principal composts are linear combinations of the original variables (Bouroche & Saporta 1982), see Equation 2.

$$PC_i = e_t^i X = e_{i1}X_1 + e_{i2}X_2 + \cdots + e_{ip}X_p \tag{2}$$

Where PCi is i^{th} the principal component (i = 1, 2, ..., p); $e^t{}_i$ is the transposed eigenvector of the data correlation matrix; X is the vector of the original variables.

The variance associated to each principal component is represented by the associated eigenvalue. The proportion of explained variance of each principal component ins given by Equation 3.

$$P_i = \frac{\lambda_i}{\sum_{i=1}^{p} \lambda_i} \tag{3}$$

Where P_i is the proportion of total variance explained by the i^{th} principal component; p is the number of variables and λ_i is the i^{th} eigenvalue.

The dimensionality of the problem can be archived by discading the principal components with low proportion of variance explained. Kaiser criterion can be used to define the number of retained principal components.

The function used to carry out Principal Component Analysis in R Software Version 4.1.2 was primcomp from Package Stats Version 3.6.2 (Team 2013)

Multiple linear regression

Multiple linear regression is a dependence multivariate statics method method. It is capable of describing the linear relationship between predictive variables (independent variables) with a quantitative response variable (dependent variable) (Hair Jr Joseph et al. 2009). The result is a model that can reasonably predict future situations. The mode is given by Equation 4.

$$Y = \beta_0 + \sum_{i=1}^{p} \beta_i x_i + \varepsilon = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p + \varepsilon \tag{4}$$

Where Y the dependent variable; β_0 the intercept; β_i the partial regression coefficient i; xi the independent or predictive variables; p the number of variables and ε the error.

The analysis of residues in the multiple linear regression assesses the adequacy of the model. The waste is the difference between the expected value and the actual value, a suitable model has residues with a normal and average waste distribution close to zero.

The function used to carry out Multiple Linear Regression in R Software Version 4.1.2 was *lm* from Package Stats Version 3.6.2 (Team 2013).

RESULTS AND DISCUSSION

The dataset is composed by 30 samples 13 numerical variables. Figure 1 shows the boxplot of these variables. The objective of boxplot analysis in this research is verify the scale and distribution of the data. The variables present different scales and distributions, maximum and minimum values of the variables differ significantly. Higher values of steel production in China can be noted in Figure 1.

The data variability shown in Figure 1 suggests the use of the correlation matrix in PCA analysis. It was necessary to establish a standardized covariance pattern, because the difference in data variability can influence the interpretation of the contained information in case of PCA and MLR analysis. Figure 1 presents the boxplot of original data and standardized data.

Statistical exploratory analysis were carried out for each variable of the dataset, see Table III.

Bartlett's Test was perfomed in the dataset and presented a p-value of $1.74x(10)^{-98}$, as p-value is below 5%, the null hypothesis is rejected. The result suggest that there are significant correlations in the data variables. Figure 2 shows the scatterplot of the data and the values of the correlations between the variables.

Significative linear correlations between the variables and the variable iron ore price were are presented in Table IV. The exception is steel production (t) in United States (USA_steel), the only variable with linear correlation less than 0.30. United States steel production presented a weak negative correlation with iron ore price. The negative correlation cam be especially associated to

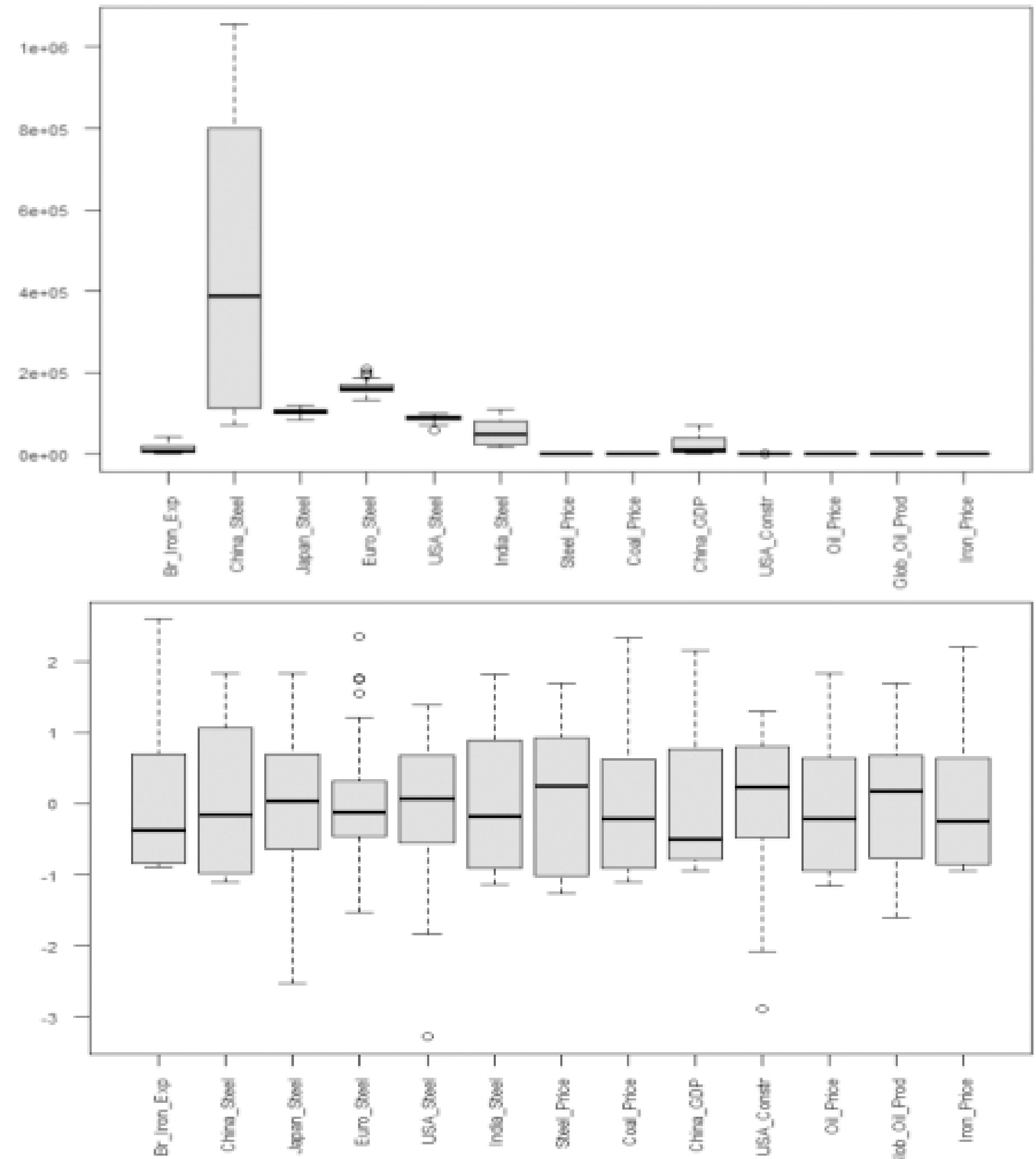

Figure 1. Boxplot of the original and stadardized variables.

USA lost of market share to China and India. United States construction presented significant value of negative correlation with iron price. It is known that the index loses strength when the price of iron ore increases.

Table III. Basic statistic of the data set.

Variable	Minimum	1st Quartile	Median	Average	3rd Quartile	Maximum
Br_Iron_Exp	2,257	2,867	8,123	12,348	19,965	41,817
China_Steel	70,564	116,458	389,332	441,449	779,132	1,054,429
Japan_Steel	83,461	99,369	104,931	104,705	110,295	120,203
Euro_Steel	132,200	154,137	160,398	162,971	169,174	210,257
USA_Steel	58,195	83,888	88,775	88,091	94,086	100,711
India_Steel	17,100	24,347	47,615	53,184	80,365	111,245
Steel_Price	108.00	129.30	236.20	215.50	292.90	360.40
Coal_Price	25.89	33.01	54.73	61.78	80.68	138.02
China_GDP	1,271	5,032	11,425	23,151	39,897	72,996
USA_Constr	-1.39	0.07	0.48	0.34	0.81	1.12
Oil_Price	13.07	19.62	42.06	48.55	67.31	105.01
Glob_Oil_Prod	60.1	66.18	72.55	71.26	75.98	82.98
Br_Iron_Exp	26.47	30.06	57.13	68.49	96.17	167.75

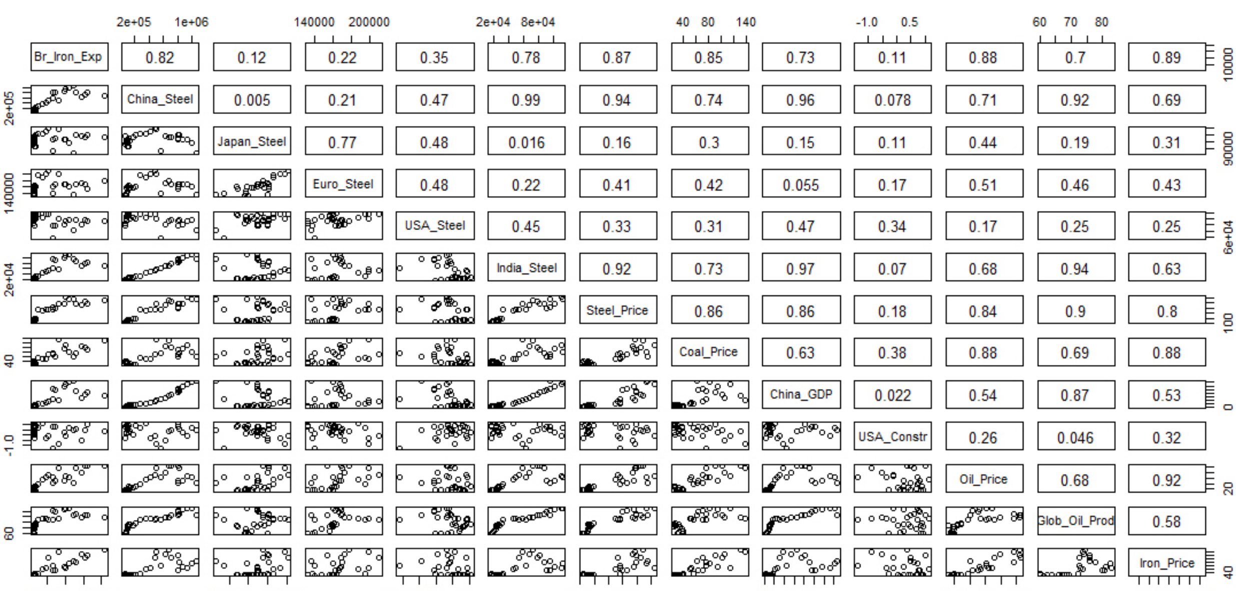

Figure 2. Scatterplot of Dataset.

Principal component analysis was performed in the standardized dataset. According to Kaiser criterion, the components with eigenvalues greater than 1 must to be retained in the analysis. The three fist principal components were retained.

The proportion of each principal component explains of the original data variance and the cumulative proportion are shown in Table V. The values found for principal components 1, 2 and 3

Table IV. Correlations between the variables with Iron Ore Price.

Variables	Linear Correlation With Iron Ore Price
Br_Iron_Exp	0.8879
China_Steel	0.687
Japan_Steel	0.3118
Euro_Steel	0.4324
USA_Steel	-0.2492
India_Steel	0.6284
Steel_Price	0.8036
Coal_Price	0.8826
China_GDP	0.5255
USA_Constr	-0.3187
Oil_Price	0.9157
Glob_Oil_Prod	0.5812

are 59.80%, 18.95% and 10.36% and they represents 89.12% of the total variability of the original data matrix.

The principal components are capable of represent the original variables. The high value of variability in the first two principal components represents strong interdependence between the variables and evidence the oligopolistic behavior of the iron ore market. An oligopolistic behavior is determined by a narrow group of countries.

The variable loadings correspond to the importance of each variable in each principal component. Table VI presents the results of loadings for the three-first principal components.

Figure 3 presents the biplot graph. Biplot presents the two first principal components, that explain 78.75% of the original data variance. A clear change of market behavior is observed from year of 2010.

Most of variables are in the same way of iron price, see Figure 3. Iron price have a rising tend throughout the studied historical series. The variable USA Construction In-dex (ICC-US) (%) (USA_Constr) presents a contrary behavior in relation to iron price. This phenomenon is justified by the increasing of the iron price, which generates a decreasing of the urge to build in USA. But the vector of this variable has small magnitude, then the variable does not have a great influence in the iron ore price to this historical series.

A real estate bubble occurred in United States in year of 2008. Thenceforward, a worldwide financial crisis occurred and steel production from Japan and Europe retracted with small positive oscillations (Trading 2021). In year of 2020, Japan had the worst performance within the analyzed historical series and Europe had the third worst production since 1991. The poor performance of then can be explained by the global context associated to Covid19 pandemic. Figure 3 shows the steel price from Japan and Europe. They partly follows the behavior of steel production from China and India.

Table V. Proportion of each principal component explains of the original data variance.

Principal component	Explained variance(%)	Cumulated explained variance (%)
1	59.8025	59.8025
2	18.9491	78.7516
3	10.3645	89.1161
4	5.2887	94.4048
5	2.2447	96.6495
6	1.1637	97.8132
7	0.8671	98.6803
8	0.6155	99.2958
9	0.3176	99.6134
10	0.2612	99.8746
11	0.0998	99.9744
12	0.0185	99.9929
13	0.0071	100

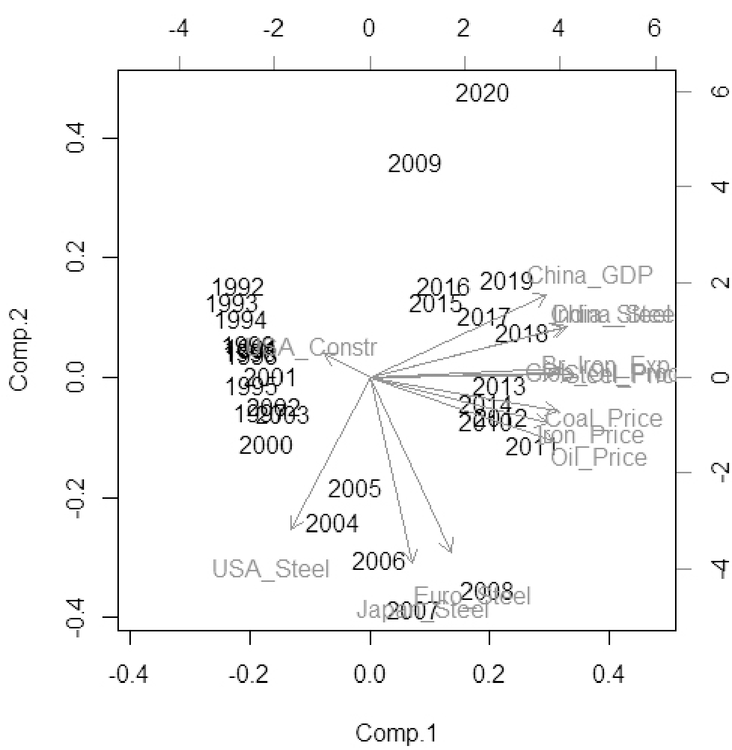

Figure 3. Biplot of two first principal components.

Table VI. Loadings for principal components 1, 2 and 3.

Variable	Component 1	Component 2	Component 3
Br_Iron_Exp	0.326	0.028	0.044
China_Steel	0.339	0.155	-0.146
Japan_Steel	0.074	-0.565	-0.025
Euro_Steel	0.141	-0.531	-0.105
USA_Steel	-0.136	-0.464	-0.374
India_Steel	0.333	0.153	-0.177
Steel_Price	0.351	0.008	-0.062
Coal_Price	0.323	-0.100	0.212
China_GDP	0.304	0.251	-0.239
USA_Constr	-0.079	0.073	-0.742
Oil_Price	0.317	-0.193	0.127
Glob_Oil_Prod	0.320	0.012	-0.261
Br_Iron_Exp	0.307	-0.140	0.230

United States also suffered the effects of the crisis of 2008. Its steel production (USA_Steel) decreased almost 40% between years of 2007 and 2009 (Trading 2021). Recovery measures and a protectionist policy were put into practice by the country, including with surcharges on imported steel. However, the production levels resumed to levels observed in the early 1990s. In addition, the country registered the second lowest steel production in the historical series in year of 2020, because of Covid-19 pandemic. Then this variable has a contrary behavior to iron ore price.

Biplot graph shows two district clouds of sample elements (years), see Figure 3. Years of the 1990s and early 2000s followed the behavior of steel production in the United States. Years of the first decade of the 21st century are in an intermediate position, while the years of the decade that started in 2010 follow the accelerated behavior of growth of Chinese production.

In order to carry out multiple linear regression analysis, the variables US Construction Index (ICC-US) (%) (USA_Constr) , steel production (t) from Japan (Japan_Steel), steel production (t) from Europe (Euro_Steel) and steel production (t) from United States (USA_Steel) were removed. The decision of removing these variables was based on the lower interdependence with the dependent variable iron ore price. Before performance of multiple regression analysis, multivariate outlier detection was carried out. Tree outliers were detected: 2004, 2008 and 2020.

Before performance of multiple regression analysis, multivariate outlier detection was carried out. Tree outliers were detected: 2004, 2008 and 2020.

In year of 2004, China registered the highest inflation since 1997. With the objective control the economic growth, the Chinese Government regulated tax increases, since this event would make it difficult to pay the debts of public companies. The implemented measures included restrictions on credit and investment projects, especially in the real estate market and automobile industry, sectors that essentially depend on steel and on iron ore.

According to (Trevisan 2004), the growing Chinese demand for raw materials affected the prices of some products around the world. In year of 2003, the country consumed 30% of the global steel production, influencing the price of the product on the international market. This scenario could be repeated within 2 years after the end of the COVID-19 pandemic. Probably, the Chinese government will try to control the inflation caused by the large issuance of paper money. Currently, it is estimated that 20% of the dollar in circulation was issued in 2020, this being a historic record. Consequently, several countries may adopt measures to rescue their economics.

Inflation of Chinese economy also marked the year of 2008. Consumer Price Index (CPI) is used to measure inflation trends and it achieved 8.7%, representing the biggest increase in the last twelve years. Then, Chinese government invested in containing price increases, maintaining the stable economic growth, active fiscal policy and relatively open monetary policy. At the end of 2008, China has injected about $ 586 billion to stimulate the economy. In addition, the country changed the agreement system, which took advantage of its monopoly to change longterm to mediumterm agreements. Besides, a strong global financial crisis directly affected commodity prices in this year.

Year of 2020 was noticeable by the Covid-19 pandemic. Brazil increased by 2% the volume exported in mineral products in 2020 over 2019, according to data released by the Brazilian Mining Institute (IBRAM 2021). In the context of iron ore trade between Brazil and China, the Asian country reinforced its position as the main destination for Brazilian iron ore. In 2019, the Asian country accounted for 62% of exports. In 2020, this percentage rose to 72% (ANBA 2020).

The outliers were removed and the multiple linear regression model was obtained. The model is given by Equation 5.

$$\begin{aligned} Iron\_price \quad = \quad & -1,382(10^2) + 8,3910(10^{-4}) \\ & Br\_Iron\_Exp + 4,508(10^{-4}) \\ & China\_Steel - 5,457(10^{-3}) \\ & India\_Steel + 4,693(10^{-2}) \\ & Steel\_Price + 9,242(10^{-1}) \\ & Coal\_Price + 1,018(10^{-4}) \\ & China\_GDP + 1,409(10^{-1}) \\ & Oil\_Price + 3,205 \\ & Glob\_Oil\_Prod + \varepsilon \end{aligned} \tag{5}$$

The residuals consist of the difference between the predicted value and the actual value. The model presented a median equal to -1.432, with minimum value and maxi-mum value equal to -13.215 and 21.741, respectively. Considering this interval, the residuals approach to zero, indicating a good adequacy of the model. The most significant variables in the determination of iron ore price are average annual value of iron ore and concentrated exports from Brazil with 62% content (USD) (Br_Iron_Exp), steel production (t) in China (China_Steel) and annual average value of coal price (Coal_Price), since they presented pvalues less tha 0.05. The significance was measured using the QR decomposition method of resolution for square parameters.

Adjusted R-squared consists of a measure of explanatory power of regression models. The obtained model presented an adjusted R-squared equal to 94.2%, indicating an excellent adequacy of the model.

ANOVA is a statistic in which the variance of a set of observations of adjusted model is analyzed. It was used to made a commentary analysis of variable significance. Table VII presents the results of ANOVA.

Table VII. Loadings for principal components 1 and 2.

Variable	Df	Sum Sg	Mean Sg	F value	Pr(F)	Signif. codes
Br_Iron_Exp	1	40868	40868	384.4012	1.36E-13	***
China_Steel	1	264	264	2.4847	0.132373	
India_Steel	1	2402	2402	22.5894	0.000159	***
Steel_Price	1	1152	1152	10.8323	0.004059	**
Coal_Price	1	556	556	5.2322	0.034496	*
China_GDP	1	262 262	2.4665	0.13371		
Oil_Price	1	21 21	0.2011	0.659154		
Glob_Oil_Prod	1	236	236	2.2175	0.153765	
Residuals	18	1914	106			

ANOVA points out average annual value of iron ore and concentrated exports from Brazil with 62% content (USD) (Br_Iron_Exp), steel production (t) in India (India_Steel), annual average value of prices (USD/t) of steel (Steel_Price) and annual average value of coal price (USD/t) (Coal_Price) as significant predictor variables in the model. These variables are the variables that have the greatest weight in principal component 1.

India is one of the countries with fastest economic growing of the world. Infrastructure and automobile sectors have increased its demand for steel year after year and Indian government promotes incentives to steel industry through investments and political reforms.

According to (T&A 2021), Indian steel production has been growing since its independence. The country gained a prominent position in the global steel landscape due to the establishment of a new state-of-the-art steel plants, the modernization of older plants, the incentive of energy-efficient technologies and retroactive integration to global raw material sources.

Year of 2018, India overtook Japan in steel production raking and became the second largest producing country of the world, only behind China. However, like the others countries, its steel industry was also impacted by Covid-19 pandemic in year of 2020. It is expected that the country will double the current average production until 2031. Figure 4 shows the annual production of steel of the main countries.

China has the fastest growing economy of the world, with an average GDP growth of 9.28% over the last 30 years. On the other hand, the United States, the world's largest economy, has an average GPD equal to 2.30% in the same period.

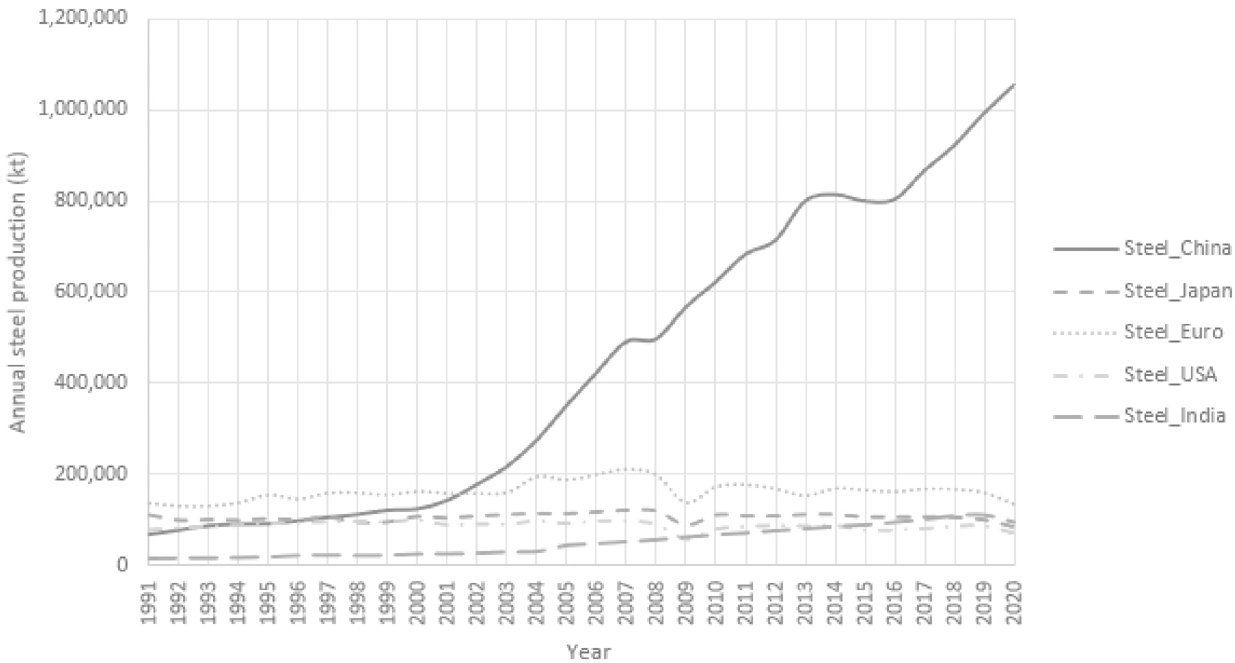

Figure 4. Annual steel production of the main producers.

According to (Nonnenberg 2010), the rise of the Chinese economy is due to multiple factors, including: liberalization process of the price formation system; liberalization of foreign trade; creation of Special Economic Zones (SEZs); absence of intellectual property protection; existence of economies of scale thanks to the gigantic population; existence of a large contingent of low-wage labor; growth of Foreign Direct Investments (FDIs); policies to encourage innovation and transfer and generation of science and technology. Chinese GDP has been boostered by the construction and industry sector.

China has been the largest steel producer of the world in the last two decades. In year of 2015, the Chinese production retracted, because steelmakers were forced to make production cuts due to the decrease of demand, growing losses (mainly motivated by the lowest levels of steel prices in decades) and credit banking services with more restrictions.

In year of 2020, it was the only country among the large producers that increased the steel production, growing of 5.8% when compared to 2019.

Steel is one of the principal components of Chinese civil construction. In year of 2017, the country had more than 300,000 construction companies. The value-added production of the sector represented 3.8% of the Chinese Gross Domestic Product (GDP) in 1978, a rate that rose to 6.7% in 2017, according to China National Bureau of Statistics (Portugueses 2018).

Steel is manufactured using iron ore, coal and lime. Brazil is the second largest global exporter of iron ore and also ranks the position of reserves. In 2019, the export of iron ore had a FOB (Free On Board) value equal to US$ 21.8 billion. In the same year, iron ore occupied the third position in the ranking of the most exported products, behind only soy and oil. China is the main buyer of Brazilian iron ore, accounting 59% of Brazilian exports in 2019. The country is the largest consumer of the commodity in the world and it is among the three largest producers in the world, behind Australia and Brazil.

Coke is a product from mineral coal it is used by steel industry. Thus, the steel industry is largely dependent on coal. Its price fluctuates due to global supply and demand, in addition to production costs. The biggest consumers are China (responsible for half of the world demand), United States and India. In year of 2030, it is estimated that China and the India will account for 60% of the world demand for coal (Rodrigues 2009).

The discussion above allows comprehend the importance of these independent variables, which are the most important variables of principal component 1 and most significant variables in definition of iron ore price.

Figure 5 presents a constant variance of the experimental errors (homocedasticy) and a non-tendency of the residuals for different samples, which confirms that the model has a good fit. The residuals presented a approximately normal distribution, see Figure 5, which is a indication of a good model fit. The normality was confirmed by Shappiro's Normality test, with a p-value equal to 0.1198.

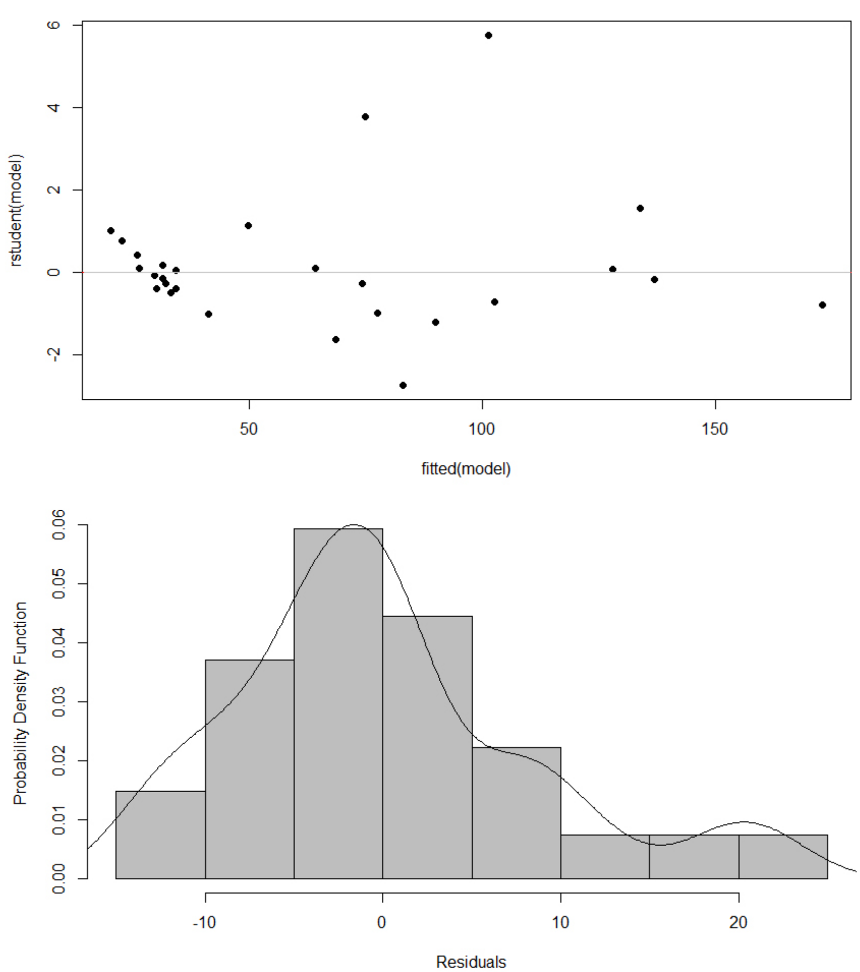

Figure 5. Homocedasticy and histogram of residuals.

CONCLUSIONS

Multivariate analysis allows a grounded study of the relation between economic variables in iron ore context.

The three first principal components are capable of explaining 89.12% of the data variability. Besides, Biplot graph allows visually verify the behavior of the variables in relation to iron ore price.

Linear multiple technique allows define the most significative variables in the variation of iron ore price. They are average annual value of iron ore and concentrated exports from Brazil with 62% content (USD) (Br_Iron_Exp), steel production (t) in China (China_Steel), annual average value of prices (USD/t) of coal (Coal_Price), steel production (t) in India (India_Steel), and annual average value of prices (USD/t) of steel (Steel_Price).

This study demonstrates the relevance of China in the international iron ore market. The country is increasing its production to try to restraint the increase in the commodity prices. For this reason, there is a tendency to reduce prices in the coming years.

India was more relevant than expected. Despite it being a notable steel producer, it was not expected to the country presents more relevant significance than other variables. On the other hand, despite the great influence of the United States economy on various sectors of the world economy, it did not show great relevance in the price of iron ore.

Furthermore, the low influence of annual average value of oil price in iron ore price was not expected, considering principal component analysis and multiple linear regression.

A strong influence of Brazilian iron ore production in the international market was defined. Although the country's exports have a strong link with China and India demands, the reduction in Brazilian production would create a scenario of rising commodity prices. The model created through multiple linear regression, could be used to future predict the iron ore price, once the independent variables can be known or estimated.

Acknowledgments

The authors wish to thank Mining Department from Universidade Federal de Ouro Preto (UFOP) and Post-graduation Program in Mineral Engineering (PPGEM – UFOP). Besides, the authors are grateful for the collaboration of the Research Group on Data Science in Engineering (CIDENG).

REFERENCES

ALAMEER Z. 2020. Multistep-ahead forecasting of coal prices using a hybrid deep learning model. Resources Policy 65. URL https://doi.org/10.1016/j.resourpol.2020.101588. 2021-02-24.

ANBA. 2020. (Câmara de Comércio Árabe Brasileira), Mineração brasileira aumentou exportação em 2020. URL https://anba.com.br/mineracao-brasileira-aumentou-exportacao-em-2020/. 2021-02-24.

BARTLETT MS. 1951. The Effect of Standardization on a X^2 Approximation in Factor Analysis. Biometrika 38(3/4): 337-344. URL http://www.jstor.org/stable/2332580. 2022-09-25.

BIAGE M. 2012. Estatística Econômica e Introdução á Econometria. 3rd ed. Florianópolis: Departamento de Ciências Econômicas/UFSC, 79-97 p.

BOUROCHE J & SAPORTA G. 1982. Análise de dados. 1st ed. Rio de janeiro: Zahar Editores.

ENDERLEIN G. 1987. Hawkins, D. M.: Identification of Outliers. Chapman and Hall, London – New York 1980, 188 S., £ 14, 50. Biometrical Journal 29: 188p.

FILZMOSER P. 2004. A multivariate outlier detection method. In: Proceedings of the seventh international conference on computer data analysis and modeling, vol. 1. p. 18-22. Minsk: Belarusian State University.

GAGGIATO VC. 2014. Do aço ao minério: um novo modelo de avalição da oferta e demanda global e precificação de minério de ferro. Universidade Federal de Minas Gerais.

HAIR JR JOSEPH F, BLACK WILLIAM C, BABIN BARRY J & ANDERSON ROLPH E. 2009. Multivariate data analysis.

HOFFMANN R. 2016. Análise de regressão: uma introdução á econometria.

HOTELLING H. 1933. Analysis of a complex of statistical variables into principal components. J Educ Psychol 24(6): 417-441, 498-520.

IBRAM. 2021. Mineração industrial brasileira fecha 2020 com desempenho positivo. Belo Horizonte. URL https://ibram.org.br/noticia/mineracao-industrial-brasileira-fecha-2020-com-desempenho-positivo/. 2021-02-25.

KAISER HF. 1970. A second generation little jiffy. Psychometrika 35: 401-415.

KRIGE DG & MAGRI EJ. 1982. Studies of the effects of outliers and data transformation on variogram estimates for a base metal and a gold ore body. J Int Assoc Math Geol 14(6): 557-564.

LI D, MOGHADDAM MR, MONJEZI M, JAHED ARMAGHANI D & MEHRDANESH A. 2020. Development of a group method of data handling technique to forecast iron ore price. Appl Sci 10(7): 2364.

NONNENBERG MJB. 2010. China: estabilidade e crescimento econômico. Braz J Political Econ 30: 201-218.

PORTUGUESES P. 2018. Setor de construção da China registra rápido crescimento desde 1978. Beijing: O Diário do Povo Online URL http://portuguese.people.com.cn/n3/2018/0910/c309806-9498945.html. 2021-02-16.

RODRIGUES AFS. 2009. Economia Mineral do Brasil. Série Estatísticas e Economia Mineral. Brasília: DNPM/MME, 764 p. URL https://www.gov.br/anm/pt-br/centrais-de-conteudo/publicacoes/serie-estatisticas-e-economia-mineral/outras-publicacoes-1/2-2-carvao. 2021/02/19.

T&A. 2021. Consulting. URL https://investexportbrasil.dpr.gov.br/arquivos/PesquisasMercado/PMRIndiaIndustriaSiderurgica2017.pdf. 2021/02/16.

TEAM RC. 2013. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL http://www.R-project.org/.

TRADING E. 2021. tradingeconomics. URL https://tradingeconomics.com/. 2021/02/09.

TREVISAN C. 2004. Indústria superaquecida preocupa a China. São Paulo: Folha de S. Paulo. URL https://www1.folha.uol.com.br/folha/dinheiro/ult91u82949.shtml. 2021/02/16.

VALADARES FG, AQUINO A & RABELO R. 2012. Detecção de outliers multivariados em redes de sensores sem fio. In: XLIV Simpósio Brasileiro de Pesquisa Operacional. SBPO.

VARELLA C. 2008. Análise de Componentes Principais. Rio de Janeiro: Instituto de Agronomia, Universidade Federal Rural do Rio de Janeiro/UFRJ, 12 p.

VICINI L. 2005. Análise multivariada da teoria à prática. 1st ed. Santa Maria: UFSM/CCNE, 215 p.

WÅRELL L. 2018. An analysis of iron ore prices during the latest commodity boom. Miner Econ 31(1): 203-216.

How to cite
CAMPOS BIS, LOPES GCA, CASTRO PSC, SANTOS TB & SOUZA FR. 2023. Econometric model of iron ore through principal component analysis and multiple linear regression. An Acad Bras Cienc 95: e20211422. DOI 10.1590/0001-3765202320211422.

Manuscript received on October 27, 2021;
accepted for publication on February 5, 2022

BÁRBARA ISABELA DA SILVA CAMPOS
https://orcid.org/0000-0002-0209-6224

GISELE C.A. LOPES
https://orcid.org/0000-0002-2383-9533

PHILIPE S.C. DE CASTRO
https://orcid.org/0000-0001-7538-422X

TATIANA B. DOS SANTOS
https://orcid.org/0000-0001-5484-6675

FELIPE R. SOUZA
https://orcid.org/0000-0001-6804-9589

Programa de Pós-Graduação em Engenharia Mineral, Universidade Federal de Ouro Preto, Departamento de Engenharia de Minas, Campus Universitário, s/n, Morro do Cruzeiro, 35400-000 Ouro Preto, MG, Brazil

Correspondence to: **Bárbara Isabela da Silva Campos**

E-mail: barbara.isabela@aluno.ufop.edu.br

Author contributions

BÁRBARA ISABELA DA SILVA CAMPOS and GISELE COSTA AGUIAR: Conception or design of the work; Data analysis and interpretation; Drafting the article. PHILIPE SILVA CARDOSO DE CASTRO: Conception or design of the work; Data collection; Data analysis and interpretation; Drafting the article. TATIANA BARRETO DOS SANTOS LOPES: Conception or design of the work; Data analysis and interpretation; Drafting the article; Critical revision of the article; Final approval of the version to be published. FELIPE RIBEIRO SOUZA: Data analysis and interpretation; Drafting the article; Critical revision of the article.