Data Mining for Sustainability Analysis: An Education Approach

Mustafa S. Al-Tekreeti[1], Salwa M. Beheiry[2] and Ayman Alzaatreh[2]

[1]Doctoral Candidate, Engineering Systems Management PhD program, American University of Sharjah, UAE

b00049253@alumni.aus.edu

[2]Associate Professor of Civil Engineering at the American University of Sharjah, UAE

[3]Associate Professor of Mathematics and Statistics at the American University of Sharjah, UAE

Abstract

Our planet's population is increasing at a rapid pace and with it the demand for food and resources. Environmental Sustainability (ES), a part of Sustainable Development (SD) concepts and techniques, is key in mitigating the effects of resource overuse. Several indicators have been identified and used to develop ES measures such as an environmental performance index and an environmental vulnerability index. These indices are used to evaluate countries and provide support for decision-making regarding national mitigation strategies and climate risks. This paper describes an educational approach to raise ES awareness and improve SD analytical skills among doctoral level students in Engineering Systems Management. The data used in this paper is obtained from existing ES indices and available data. The students use data mining and analytics techniques to evaluate the data, find relationships, and draw conclusions. These techniques and conclusions are then shared in class presentations and conference publications. Data mining converts raw data into useful information that can be understood by different audiences. It can be used to persuade policymakers about the importance of sustainable strategies for a country, a society, or certain groups or individuals' welfare by highlighting meaningful patterns and trends in ES. The paper also aims to investigate possible correlations among environmental indices and their underlying indicators.

1 Introduction

The resources of planet earth are limited. Consuming the resources in an unsustainable way will deplete them. Major resource consumers are human activities. It is expected that the world's population will reach 9.7 billion by 2050 (Sarkodie et al., 2019). With the population increasing, the demand on resources will increase as well which will impose tremendous pressure on the earth's resources. Human activities cause another major problem: climate change. It affects human health through extreme weather resulting from increasing Greenhouse Gases (GHGs). In response to these threats, governments initiate national development strategies to manage their resources in an effective way and reduce GHGs emissions. The aim for those strategies is to enforce sustainability into society to conserve resources while maintaining living standards through Sustainable Development (SD) (Wang & Li, 2019). According to the World Commission on Environment and Development, SD "meets the needs of the present without compromising the ability of future generations to meet their own needs" (WCED, 1987, p43). Through SD, finite resources can be preserved such that future generations' needs are met.

This interpretation of SD is frequently associated with social and economic development, which should be part of Environmental Sustainability. Since the SD definition was first published and slowly become accepted, the concept of Environmental Sustainability (ES) emerged with it and has its own merits. To

evaluate a nation's ES, several indices and indicators have been developed, such as: environmental performance index, environmental vulnerability index, and ecological footprint, which all measure it. The aim of environmental sustainability is to maintain natural capital in a balanced condition, where both economic development and human consumption depend on the sustainable use of planet earth's resources. Without sustainability, the vitality of the freshwater system, oceans, land, and the atmosphere will be compromised and impact on human life in a negative way. The United Nations' millennium development goals emphasize global collaboration as essential to ensure that environmental sustainability is applied in order to maintaining biodiversity and to minimize the losses of essential global natural resources (Olafsson et al., 2014). An example of these resources is the planet's forests, which not only provide inputs for national economic processes but also play an essential role in mitigating anthropogenic climate change through carbon sequestration (Stewart, 2003).

Sustainability indices provide raw data for each country about their performance in conserving resources, reducing GHGs emissions, and renewable energy uses. On other hand, the environmental performance index gives data relating to air pollution and child mortality (Olafsson et al., 2014). However, these data need to be processed to make a meaningful interpretation of the sustainability situation and support the necessary policies that promote environmental conservation. One way to process raw data is data mining. Data mining is the process of identifying essential knowledge about and relationships between data variables to draw patterns and useful information about the data (Di Blas, 2015).

The source of the data used in this paper is the United Nations Development Programme (UNDP) website. This website is open-source and provides data related to humanity and its welfare. The data combines different environmental sustainability indices for several countries, and is part of Human Development Reports published by UNDP. This report measures the key dimensions of human development in health, decent living standards and knowledge (OCHA, 2016). The aim of this paper is to find the possible relationships between different ES variables and environmental threats to human welfare (mortality rate due to air pollution and unsafe water) as an education approach, to identify the hidden factors and correlations that control them through data mining in order to facilitate the interpretations for the data to policy makers and scholars interested in environmental sustainability.

2 Literature Review

2.1 Environmental Sustainability and its Indices

The term ES was first conceptualized by researchers at the World Bank. Subsequently, it changed to environmentally sustainable development (Serageldin and Streeter, 1993). The final concept was developed by Goodland to be environmental sustainability. ES is a concept entwined with SD through focusing on the social and economic development of a country. ES defined as "improve human welfare by protecting the sources of raw materials used for human needs and ensuring that the sinks for human waste are not exceeded, in order to prevent harm to humans" (Goodland, 1995, p. 3). Goodland identifies ES concepts as a set of four constraints on major activities that control the human economic system, which are: on the source side the non-renewable and renewable resources use, and waste assimilation and pollution on the sink side (Goodland, 1995). In contrast, Holdren et al. (1995) defined ES as a concept that focuses on bio-geophysical aspects, where the meaning of biophysical sustainability is to improve or maintain the integrity of earth's

life-supporting systems through conservation which means the proper use of water, air, and land resources while improving the social and economic aspects of the current and future generations.

After defining the concept of ES by a number of scholars, several indicators emerged to measure it. Several studies have proposed different methods to measure ES and how to construct its indicators using qualitative and quantitative techniques (Cano-Orellana & Delgado-Cabeza, 2015). Those indicators focus on the current sustainability status for a country using reference values or trends generated based on historical conditions (Milman & Short, 2008). Examples for those indicators are: Environmental Vulnerability Index (EVI), Environmental Performance Index (EPI), and Ecological Footprint (EF)

2.2 Data Mining

Data mining, according to Han et al. (2001), is "the process of discovering interesting knowledge, such as patterns, associations, changes, anomalies and significant structures, from large amounts of data, stored in large databases, data warehouses or other information repositories" (Han et al., 2011, p.1). Analyzing the data through statistical software will help to find possible correlations among raw data as well as distinctive patterns between data variables (Wu et al., 2013). Grossman et al. (2013) stated that data mining involves: handling large amounts of quantitative data, discovering patterns, not needing a hypothesis (as results are driven by data). Major tasks regarding data mining can be listed as follow: class description (summary of collection data), data association (finding correlations within data sets), prediction (finding possible future values), clustering (grouping similar data), etc. Data mining uses data to extract useful information and transform it into understandable and structured information for future use. It helps business stakeholders to understand the meaning of complex data from their company and overcome any challenges to ease the interpretation of the data's information (Grossman et al., 2013).

3 Methodology

This paper provides quantitative research analysis using data from the UNDP report in order to find possible relationships between several ES variables and environmental threat variables which are considered the aim of this paper. Different statistical tools are used to analyze the data and investigate the hidden correlations among the variables. Those analyses are used as learning techniques to show how to investigate datasets and conclude useful information from them. This research tries to answer the following question:

- Are there any relationships between the mortality rate due to air pollution, unsafe water and the ES variables?

3.1 Data Description

The data used for the analysis in this paper comes from the UNDP report which is open-source. The data combines different ES indices from different databases such as: World Bank, Human Development Report Office (HDRO), Food and Agriculture Organization (FAO), and World Health Organization (WHO) for 189 countries and regions. The data is divided into two groups: ES indices and environmental threats to human welfare that form 10 indicators. Based on the country's human development level, the data is categorized into four groups: very high human development (58 countries and regions), high human development (52 countries and regions), medium human development (38 countries and regions), and low human development (37 countries and regions).

ES 7 indicators are described and measured as follows. Fossil fuel energy consumption: calculate the percentage of total consumption of energy that comes from using fossil fuels (e.g. oil, coal, natural gas, and petroleum); renewable energy consumption: calculate the percentage share of renewable energy (such as solar, wind, hydroelectric, and geothermal) in the final total energy consumption; carbon dioxide emissions: the carbon emission generated from human activities such as burning fossil fuels, cement production and gas flaring as well as the carbon emissions from forest degradation. The data in this index is expressed in two ways: tones per capita, and in kilograms per Purchasing Parity Power (PPP) of GDP in US dollars. The forest area index is expressed in two ways: the percentage of total land covered by trees natural or planted that stand for more than 5 meters and excludes the land covered by trees in agriculture and urban use (parks and gardens), and the percentage change of this forest area in the period from 1990 to 2015. Fresh water withdrawals: the percentage of freshwater withdrawals (freshwater taken from surface or ground water sources) of total renewable water resources. The withdrawals include irrigation, industrial use, and domestic use, etc. (UNDP, 2018).

The indices for environmental threats to human welfare are: mortality rate attributed to indoor and outdoor air pollution expressed per 100,000 population, mortality rate attributed to using unsafe water or sanitation and hygiene services expressed per 100,000 population, and red list index: the value that measures the aggregate risk extinction across a species group. It shows the trend in overall species that are at risk of extinction through tracking the changes in species number in each category that is in extinction risk. The value of this index ranges from 0 (all categorized species are extinct) to 1 where all categorized species are of least concern (low risk of extinction). This index helps the government to track their progress in reducing the loss in biodiversity (UNDP, 2018). Table 1 shows a summary of the data variables used in this study.

Table 1:Summary of Data Variables

Environmental Sustainability (ES) Indices	Environmental Threats to Human Welfare
Fossil Fuel Energy Consumption	Mortality Rate Attributed to Indoor and Outdoor Air Pollution
Renewable Energy Consumption	Mortality Rate Attributed to Using Unsafe Water
Carbon Dioxide Emissions: Tones Per Capita, and in kilograms Per Purchasing Parity Power (PPP) of GDP	Red List Index
Forest Area: Total Land Covered and Forest Area Percentage Change	
Fresh Water Withdrawals	

4 Data Analysis and Discussion

As mentioned in the literature review, the main tasks for data mining are: data description, data association, clustering, and prediction. Those tasks will be followed in this data analysis to investigate the dataset as a learning approach. The data is analyzed using Statistical Analysis System (SAS) software. This software can perform different statistical analysis varying from simple statistics to multivariate analysis. Since the

focus of this research is on data mining, the following analysis will be used: simple graphs (scatter plot) between both mortality rate due to air pollution and unsafe water and the other variables, cluster analysis to show how the ES variables are grouped, conducting Factor Analysis (FA) to highlight the latent variables that affect specific group of ES variables, performing Pearson and canonical correlation between mortality rate due to air pollution and unsafe water as one set and the other ES variables as another set, and conduct multivariable regression. Due to the large amount of incomplete data in medium and low human development countries that causes inaccuracy in the analysis, the analysis will focus on the first two groups: very high and high human development countries. As stated in the methodology, this analysis is used as an educational approach to highlight how to use data mining techniques to structure useful information that can be easy to interpret. Each analysis tool/technique will reveal specific items of information and/or confirm the finding of the previous ones.

4.1 Data Association

One of the tasks in data mining is to find the possible associations between the variables. There are several ways to investigate the possible patterns among the data variables, such as using graphs that underlie the possible patterns between variables (scatter plots). Figure 1 shows the possible relationships between mortality rate (air pollution and unsafe water) and other ES variables. Increasing the energy consumption of fossil fuel will increase the mortality rate due to air pollution due to the increase in hazed gas emissions for burning fossil fuels which lead to serious illnesses of the human respiration system, because the emissions that are related to fossil fuel burning will lead to serious illnesses of the human respiratory system. Particulate matter (PM10/2.5) is also a fossil fuel triggered problem that causes health issues. It is airborne and varies in size and chemical and physical composition (Zhao et al., 2013). The size of those particles is small enough to penetrate the human lungs and lead to severe health risks. Another gas emitted from burning fossil fuels and affects human health is Nitrogen oxides (NOx). It is inorganic gases formed by the amalgamation of oxygen with nitrogen available in the air, which causes destructive effects to the human bronchial system (Zhao et al.,2018). Since scatter plots do not show all possible patterns among the variables, another analysis can be used to find this. Pearson correlation is an analysis tool that measures the linear correlation between two variables to investigate the possible relationship between them (Benesty et al., 2009). The importance of Pearson correlation for the data will be indicated through the p-value. If the p-value of the correlation is less than $\alpha = 0.05$, then the correlation is significant, otherwise it is not.

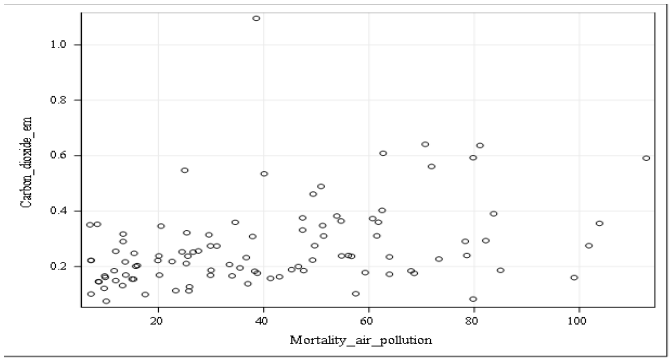

Figure 1: Scatter Plot between Mortality Rate (Air Pollution) and Carbon Dioxide Emission Kg per PPP of GDP

The correlation results shown that fossil fuel energy consumption, carbon dioxide emission Kg per PPP of GDP, and freshwater withdrawals significantly correlate to mortality rate due to air pollution. Increasing the freshwater withdrawals will lead to a decrease of the flora that cover the land due to low precipitation and decrease of the groundwater level that leads to increased desertification which causes major threats to human health due to dust and other pollutant particles. In addition, lowering the flora cover will limit the role of trees in capturing carbon dioxide from the air which by extension will cause health issues to humans. On the other hand, the mortality rate due to unsafe water is correlated negatively with the red index. This indicates that decreasing the species that have a risk of extinction will increase the mortality rate due to unsafe water. This is because more diverse animal species living near or in the water sources will lead to the transfer of infections from those animals due to their feces and/or corpses and this will increase the mortality rate due to unsafe water.

A different method in correlation is called Factor Analysis (FA). It describes the correlation among variables in terms of a potential latent (unobserved) variable called a factor. It is a way to show hidden patterns between the variables (Costello & Osborne, 2005). Figure 2 shows the path diagram for the FA. The first factor can be called the energy factor, since it is correlated to the following variables: fossil fuel energy consumption, carbon dioxide emission Kg per PPP of GDP (both positively correlated to energy factor), and renewable energy consumption (negatively correlated to energy factor). The second factor can be called the forest area factor. Because both forest area (negatively correlated to forest factor) and changes in forest area (positively correlated to forest factor) variables are correlated to it. Factor three is called the red index factor, since it positively correlates with only one variable which is the red index.

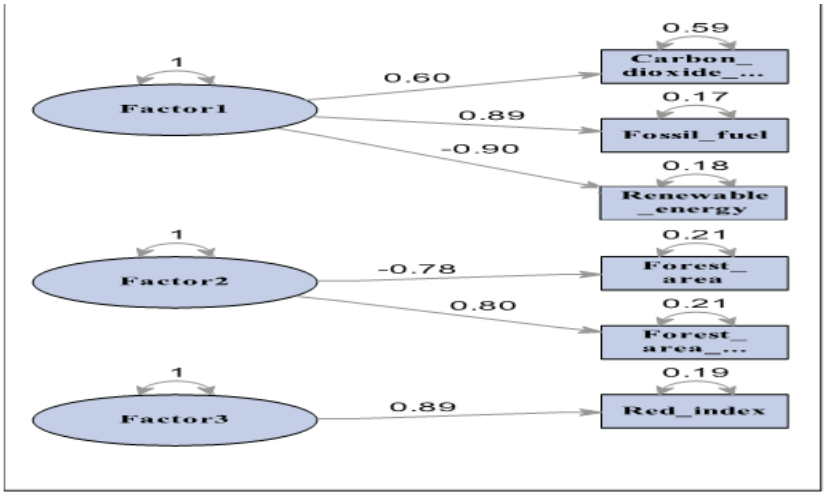

Figure 2: Factor Analysis Path Diagram

4.2 Data Clustering

Another task in data mining is represented by clustering. It divides the variables into groups based on its similarity (Xu & Wunsch, 2008). Clustering will help to understand how the variables are grouped to identify the common characteristics among them. Figure 3 show how the variables are clustered. It can be seen that the variables are grouped into three distinct clusters. Forest area and change in forest area variables are grouped in one cluster. Freshwater withdrawals, carbon dioxide emission per capita and Kg per PPP of

GDP variables are clustered in one group. Energy consumption from fossil fuel sources, renewable sources and red index variables are grouped into one cluster.

4.3 Data Prediction

The final task in data mining is prediction. Since there are two dependent variables the best technique is, therefore, multivariate regression. It predicts the behavior of several dependent variables associated to changes in independent variables (Imai, 2011). The two dependent variables in this analysis are mortality rate due to air pollution and unsafe water, and the other variables are treated as independent variables. The fitness of multivariate regression model indicted by R-squared which is 0.7854. This model shows that the only significant predictors (which p-value less than α=0.05) for the mortality rate due to air pollution are fossil fuel energy consumption, carbon dioxide emission Kg per PPP of GDP and freshwater withdrawals. The fitness of this model indicated by R-squared which is 0.7854. The prediction equation for the mortality rate air pollution will be as follows:

$$\begin{aligned} Mortality\ Air\ Pollution & \\ &= -37.19696 + 0.41461\ (fossil\ fuel\ energy\ consumption) \\ &- \ 142.45164\ (carbon\ dioxide\ emission\ Kg\ per\ PPP\ of\ GDP) \\ &+ \ 0.05134\ (freshwater\ withdrawals), \end{aligned} \tag{1}$$

On the other hand, two variables are significant predictors of mortality rate due to unsafe water: forest area and red index variables. The fitness of this model indicated by R-squared which is 0.7567. Then the prediction equation for the mortality rate unsafe water will be:

$$Mortality\ Unsafe\ Water = \ 0.92020 + 0.01031(forest\ area) \ - \ 1.62837(red\ index), \tag{2}$$

Therefore, using the above equations can help to predict future changes for mortality rate whether form air pollution or unsafe water.

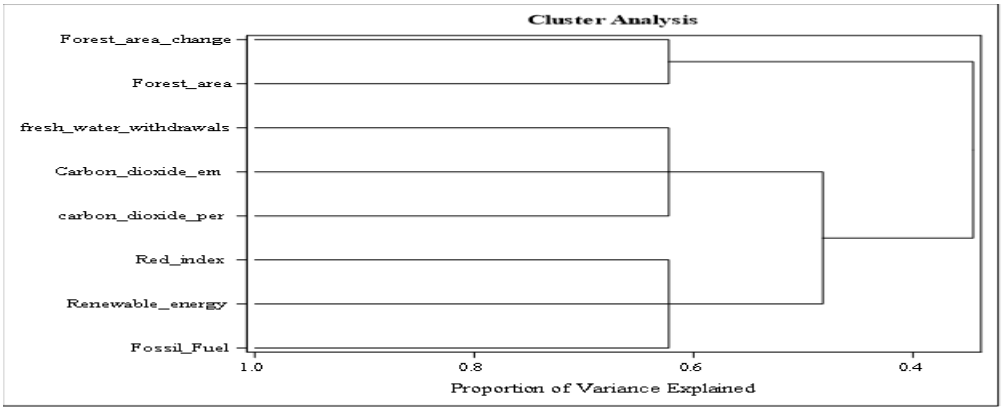

Figure 3: Cluster of Data Variables

5 Conclusion

Data mining is considered a powerful tool to analyze the data and understand the possible patterns among the variables. From analyzing the data for this paper, several important items of information have been revealed. There is a positive relationship between mortality rate due to air pollution and fossil fuel energy consumption, carbon dioxide emission per capita and Kg per PPP of GDP, and freshwater withdrawal variables. In addition, forest area and renewable energy consumption correlate negatively to the mortality rate due to air pollution. On the other hand, mortality rate due to unsafe water is negatively correlated with the red index variable. From cluster analysis, variables are clustered into three groups based on similarity. Forest area and change in forest area is the first group; freshwater withdrawals, carbon dioxide emission per capita and Kg per PPP of GDP is the second group; and consumption of energy from renewable sources and fossil fuel sources and red index is the third group. This indicates that there is a common relationship between them which could be investigated in detail in the future. The predictors for mortality rate air pollution from multi-variant regression analysis are: fossil fuel energy consumption, carbon dioxide emission Kg per PPP of GDP and freshwater withdrawals. However, the predictors for mortality unsafe water are forest area and red index variables. Those items of information can guide future research among environmental scholars to investigate in detail the cause of those relationships and what are the reasons behind those findings.

Moreover, the use of data analytics as a progressive and open-minded pedagogical tool has great potential in transforming engineering education at both the graduate and undergraduate levels. In this particular study, the data was carefully examined, and connections and correlations found. From this connection, conclusions were drawn, and new analysis triggered and explored. This work was conducted in partial fulfillment of a doctoral-level course in sustainability analytics, aiming to teach students the use of data mining and analytics to explore sustainable development issues, and how to find answers to theses from data and scholarly work.

References

Benesty, J., Chen, J., Huang, Y., & Cohen, I. (2009). Pearson correlation coefficient. In Noise reduction in speech processing (pp. 1-4). Springer, Berlin, Heidelberg.

Cano-Orellana, A., & Delgado-Cabeza, M. (2015). Local ecological footprint using Principal Component Analysis: A case study of localities in Andalusia (Spain). Ecological Indicators, 57, 573-579.

Costello, A. B., & Osborne, J. (2005). Best practices in exploratory factor analysis: Four recommendations for getting the most from your analysis. Practical assessment, research, and evaluation, 10(1), 7.

Di Blas, N. (2015). Exploratory portals of research data in education. Research on Education and Media, 7(2), 21-27.

Goodland, R. (1995). The concept of environmental sustainability. Annual review of ecology and systematics, 26(1), 1-24.

Grossman, R. L., Kamath, C., Kegelmeyer, P., Kumar, V., & Namburu, R. (Eds.). (2013). Data mining for scientific and engineering applications (Vol. 2). Springer Science & Business Media.

Han, J., Kamber, M., Pei, J. (2011). Data Mining: Concepts and Techniques. In The Morgan Kaufmann Series in Data Management Systems, Morgan Kaufmann Publishers.

Holdren, J. P., Daily, G. C., & Ehrlich, P. R. (1995). The meaning of sustainability: biogeophysical aspects. Defining and measuring sustainability: the biogeophysical foundations, 3-17.

Imai, K. (2011). Multivariate regression analysis for the item count technique. Journal of the American Statistical Association, 106(494), 407-416.

Milman, A., & Short, A. (2008). Incorporating resilience into sustainability indicators: An example for the urban water sector. Global Environmental Change, 18(4), 758-767.

OCHA. (2016). UNDP Human Development Reports Office (HDRO). Retrieved from https://data.humdata.org/organization/undp-human-development-reports-office.

Olafsson, S., Cook, D., Davidsdottir, B., & Johannsdottir, L. (2014). Measuring countries′ environmental sustainability performance–A review and case study of Iceland. Renewable and Sustainable Energy Reviews, 39, 934-948.

Sarkodie, S. A., Strezov, V., Weldekidan, H., Asamoah, E. F., Owusu, P. A., & Doyi, I. N. Y. (2019). Environmental sustainability assessment using dynamic Autoregressive-Distributed Lag simulations—Nexus between greenhouse gas emissions, biomass energy, food and economic growth. Science of the Total Environment, 668, 318-332.

Serageldin, I., Streeter, A. (1993). Valuing the environment: proceedings of the First Annual Conference on Environmentally Sustainable Development. Environmentally Sustainable Development Proceedings Series No. 2, The World Bank, Washington, D.C.

Stewart, K. (2003). Selling forest environmental services. Economic Botany, 57(4), 659-659.

United Nations Development Programme. (2018). Human Development Reports. Retrieved November 29, 2019, from http://hdr.undp.org/en/content/dashboard-4-environmental-sustainability-0.

Wang, Z. X., & Li, Q. (2019). Modelling the nonlinear relationship between CO2 emissions and economic growth using a PSO algorithm-based grey Verhulst model. Journal of cleaner production, 207, 214-224.

World Commission on Environment and Development. (1987). Our common future (Oxford paperbacks). Oxford: Oxford University Press.

Wu, X., Zhu, X., Wu, G. Q., & Ding, W. (2013). Data mining with big data. IEEE transactions on knowledge and data engineering, 26(1), 97-107.

Xu, R., & Wunsch, D. (2008). Clustering (Vol. 10). John Wiley & Sons.

Zhao, B., Wang, S., Wang, J., Fu, J., Liu, T., Xu, J., . . . Hao, J. (2013). Impact of national nox and so2 control policies on particulate matter pollution in china. Atmospheric Environment, 77, 453-463.

Zhao, S., Liu, S., Hou, X., Cheng, F., Wu, X., Dong, S., & Beazley, R. (2018). Temporal dynamics of so2 and nox pollution and contributions of driving forces in urban areas in china. Environmental Pollution (barking, Essex : 1987), 242, 239-248.