University of East London

RISK ASSESSMENT OF DEADLY ECONOMIC SOCIO-POLITICAL CRISIS WITH CORRELATIONAL NETWORK AND CONVOLUTIONAL NEURAL NETWORK

Jean Brice Ghislain TETKA

A thesis submitted in fulfilment of the
requirements of the University of East London
for the degree of Professional Doctorate in Data Science

School of Architecture, Computing & Engineering
University of East London
Dr Yang Li
Dr Bilyaminu Auwal Romo
London, United Kingdom
September 2022

Abstract:

From social analysis to biology to machine learning, graphs naturally occur in a wide range of applications. In contrast to studying data one at a time, graphs' unique capacity to capture structural relationships among data enables them to yield additional insights. Nevertheless, the capacity to learn from graphs can be difficult because meaningful connectivity should exist between data and the form of data such as text, numbers or categories should allow for building a graph from their relationships. Investigating hidden patterns in the variation of development indicators and severe socio-political crises that happened in low-income countries is an analytical approach that has been experimented with in this research. Evidence of a correlation between socio-political crises and development indicators suggests that a method to assess the risk of crisis should consider the context of each country, as well as the relative means of crisis. This research reviewed different risk assessment methods and proposed a novel method based on a weighted correlation network, and convolution neural network, to generate images representing the signature of development indicators correlating with a crisis. The convolution neural network trained to identify changes in indicators will be able to find countries with similar signatures and provide insights about important indicators that might reduce the number of deadly crises in a country. This research enhances the knowledge of developing a quantitative risk assessment for crisis prevention with development indicators.

Keywords—risk assessment, correlation network, convolution neural network, development indicators, crisis

Declaration

I declare that this thesis was composed by myself, that the work contained herein is my own except where explicitly stated otherwise in the text, and that this work has not been submitted for any other degree or processional qualification except as specified.

Table of Contents

List of Figures

List of Tables

Abbreviations

CAST	Conflict Assessment Tool System
CNN	Convolution Neural Network
DI	Development indicators
EVI	Economic Vulnerability Index
GRA	Gray relational analysis
HDI	Human Development Index
IMF	The International Monetary Fund
LEDC	Less Economically Developed Country
MEDC	More Economically Developed Country
MENA	Middle Eastern regions and Northern Africa
OECD	Organisation for Economic Co-operation and Development
QRA	Quantitative risk assessments
SDGs	The Sustainable Development Goals
SNA	Social Network Analysis
SPEED	Social, Political and Economic Event Database
UNDP	United Nations Development Programme
WB	The World Bank
WDI	The World Development Indicators

Acknowledgements

This thesis would not have been done without the support of the University of East London, in particular

My teacher and director of studies, Dr Yang Li

My supervisor, Dr Bilyaminu Romo

My teacher, Pr Allan Brimicombe

I'm deeply grateful to my colleagues from Transparency International, who have provided me advice and inspiration to conduct this research.

I express my sincere gratitude to Gareth, Jose, Juanita and Shaherah, who helped me review this research.

I express my sincere gratitude and appreciation to my wife, Anais Mbeng whose support and patience is profoundly appreciated.

Dedication

To honour the loving memory of my beloved father Emile Tetka.

For those who died during the February 2008 protest in Cameroon, and people who die because they express their anger during turbulent times.

I dedicate this work to:

my dear Mom Bénédicte and sisters (Marie, Catherine, Chantal and Claire)

my wife Anais

my kids Evan Rodney and Maia Julienne,

Thank you

Ofir to be an inspiration and a life mentor

Marc Ona to remain constant and loyal

Mr and Mrs Mbeng Ekorezock for being a second family

Dr Good and Dr Peart to teach with your heart

Dr Yang Li, for being always responsive

Gillian, Adam, Peter for your presence over difficult times

Arwa and Annette for being the little hands that change someone's life

Blessed be the Eternal who has allowed me to complete this research.

CHAPTER ONE: INTRODUCTION TO THE RESEARCH

1.1 Introduction

The World Development Indicators (WDI) is a compilation of relevant, high-quality and internationally comparable statistics about global development and the numerical measure of the quality of life in the world to fight against poverty. The benchmark is used to project a country's advance by gathering a range of economic, environmental and social aspects. Some indicators, like the environmental indicators, show the state of the planet and the use of natural resources and the observed impacts. They reflect how socio-economic circumstances influence people living in poor conditions to support violence. In the Sub-Saharan African context, low-income levels exacerbate political instability (Fagbemi and Fajingbesi, 2022). Spearheaded by pandemics, war and climate change, fast-rising oil and food prices are observed, and riots erupt on all continents, with the risk of violent protest and political instability in Africa (Gopaldas and Menzi, 2021). Poverty, unemployment, and poor social services worsened in an already precarious social context, where protests can move to a different level by combining the food price rise up to questions of political change or public reforms. “Development” means for a country, economic and social progress. Often it refers to the level of the economy, security, quality of life, education, or security. The significant primary aspects are economic, sociological, and communication. There is a standard benchmark level for assessing the development of the country's level, i.e., per capita gross domestic product, or income per capita, the industrialisation level, technological infrastructure amount, and the standard of living. Life expectancy, education and the real Gross Domestic Product (GDP) per capita are the three indicators for the Human Development Index (HDI), which the United Nations developed as a more comprehensive measure of "Development". The Human Development Index measures

longevity, knowledge and standard of living, and the final outcome is a score between be 0 and 1.

A significant number of variables can be used to describe the socio-economic and political situation of a country. Much research has investigated the correlation between the WDI and crisis. Demeke (2022) found evidence of a relationship between youth unemployment and political instability for countries member of the Intergovernmental Authority on Development, which includes countries from the Horn of Africa, the Nile Valley and the African Great Lakes. A classic method to examine such a relationship between variables is the regression analysis. One assumption of the regression analysis that often brings inaccurate data when dealing with development data is the assumption that the socio-political and economic data used to analyse political instability, for instance, are independent variables. If a correlation could exist between youth unemployment and political instability, youth unemployment could also correlate with education or poverty. Therefore, considering youth unemployment, education and poverty as independent variables might give false results.

The interaction between indicators representing poverty, unemployment, food or oil prices has been considered and can be expressed in a network model, where each node of the graph will represent an indicator and the edge of their relationship. The network model is a flexible way of representing the objects with their relationships. The challenge that comes with analysing the relationship between development indicators and the relationship with a crisis is to put in place a novel model. This requirement was amplified with a long list of variables to analyse and a high number of records.

The novel approach experimented within this study associate the correlation network representation with the power of image analysis. Like neurons in brains, a convolutional neural network (CNN) can serve to analyse images of a correlation network. Wang, Hu and Lyu (2020) acknowledge the challenge of extracting image features from 3D graphs where the shape can

change. A simplified visual representation of a correlation network is a centrality. Instead of providing a network with nodes and edges, the centrality measurement gives in one line the relative importance of a variable correlating with other variables. The image's shape is the same for all images, which suffices to train a CNN model to learn the features that will allow finding differences and similarities between images. Lastly, in the approach, because finding similarities does not tell what development indicators are vital for assessing a crisis in a country, another statistical method will analyse a few similar images to find what indicators play a significant role in a crisis. A CNN model trained from centrality images eases the identification and selection of development indicators assuming at the same time relationship among development indicators and the relation between development indicators and deadly crises.

Hu et al. (2019) proposed a security risk assessment method that integrates deep learning. The risk case is an image stored in a library, and the deep learning model will find a new case matching a risk image in the library. A similar approach is used in this research, where images representing a country's situation are compared with images in the library to find countries with a similar situation. However, the case reasoning used in similar risk searching implies that a solution to respond to the risk exists; socio-economic and political risks are highly versatile, and there is no perfect solution to a risk case. Many alternatives exist, and the decision-maker chooses one of the alternatives to implement a development policy. The deep learning method will present countries that faced the same challenges, and the statistical method will identify the indicators that have changed to reduce the deadly effects of the crises.

This research does not pretend to offer solutions to the deadly crises but will inform the decision-makers about indicators to consider in a development policy that could reduce the number of fatalities. Countries facing extreme tension when the number of deaths is exceptionally high are excluded from this analysis because the priority is not to bring development but to end, for instance, civil war or terrorism. Nevertheless, terrorism and civil

war can find their origins in poverty and unemployment. Other countries that are excluded from this list are those that have an extremely low number of fatalities due to socio-economic and political crises. Therefore, developed countries and countries in war are excluded from the analysis. Countries with missing data are also excluded from the analysis. These types of countries tend to be micro states and islands due to an unexpected data gap. Further research will investigate how to mitigate this gap and develop a risk model that can apply to countries with limited data.

This study focuses on deadly crises, analysing the relationship among development indicators to create a country database of images that describes the correlation between development indicators and deadly crises. The data used corresponds to the more recent and continuous data recorded for each country by year. Because of the required data quality, chapter five assessed the datasets used for this research.

Starting with the research introduction in Chapter one, Chapter two introduces the theoretical concept of development and crisis and correlation network with convolutional neural network, and risk assessment with economic vulnerability. This research claims to propose an innovative method; the systematic and literature review in Chapter three investigates if any similar research has been done and what are the gaps. Chapter four explored the limitations of the regression analysis. Chapter five presents the detailed methodology. Chapter six assesses the different datasets required for the research, including the social media data because of its tremendous role in the crisis, but with several limitations; due to political bots and disinformation, social media data are excluded from the analysis. On the data retained for the research, Chapter seven presents the results of the designed model, with the discussion and the conclusion in Chapter eight.

1.2 Why this research

With violence on the rise in many countries and the significant number of deaths that result from the violence, this research pursues the ideal of finding out how data science could help avoid or reduce the number of deaths resulting from clashes such as popular uprisings. The death of people protesting for their rights to better living conditions is unacceptable, no matter where it occurs, is the overall contribution this research want to make. The limitations of traditional crisis risk assessment methods that can integrate the uniqueness of each country with multiple time series variables require the development of a new method testing the innovative approaches offered by deep learning. A new method will identify the indicators that trigger a crisis or can help mitigate it.

Data mining reveals specific patterns in development data that correspond to states of socio-political crisis at the country level. The risk here is people being killed during a demonstration. For example, some indicators such as the Fragile States Index ((*Fragile States Index | The Fund for Peace*)) can provide the probability of imminent conflict in a country, but cannot isolate the development factors influencing the insight. This indicator informs about the state of fragility of a country, whereas this research seeks to inform about the causes of fragility. The preliminary concept map (Figure 1.2 1) of the proposed risk analysis method contains as input data the indicators of development in the world and the exit data from the deadly crisis. Analysing the correlation between input indicators and their variations will not imply causal analysis between development indicators and deadly crises; the hazard variables are excluded from the calculation only to consider predefined multivariate time series variables.

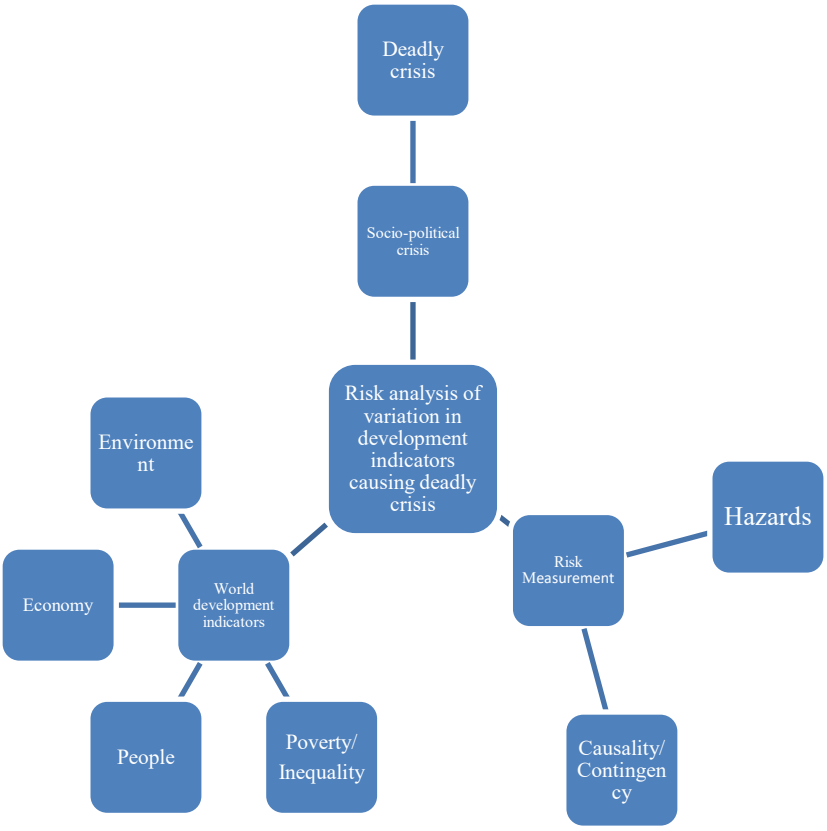


Figure 1.2-1 Preliminary concept map

This research will not look at what should be done to avoid protests but at the roots of economic and socio-political changes to which the population could react negatively and result in a crisis. It refers in this research to the discovery of hidden patterns between the variation in development indicators and the deadly socio-political crisis that has occurred in countries.

1.3 Research aims and objectives

This research aims to develop an innovative data model to assess the risk of violent crisis, using network analysis and graph convolution. Based on the assumption that development indicators are interconnected, the research will identify and analyse the networks of development indicators that preceded harmful crises and apply graph

convolution to find the change in development indicators that might lead to a violent situation. That is to improve the decision-making in an environment with a more significant number of variables, with non-dependencies and non-linearities. The Human Development Index (HDI) variables in Figure 1.3-1 link development with human welfare (UNDP, no date).

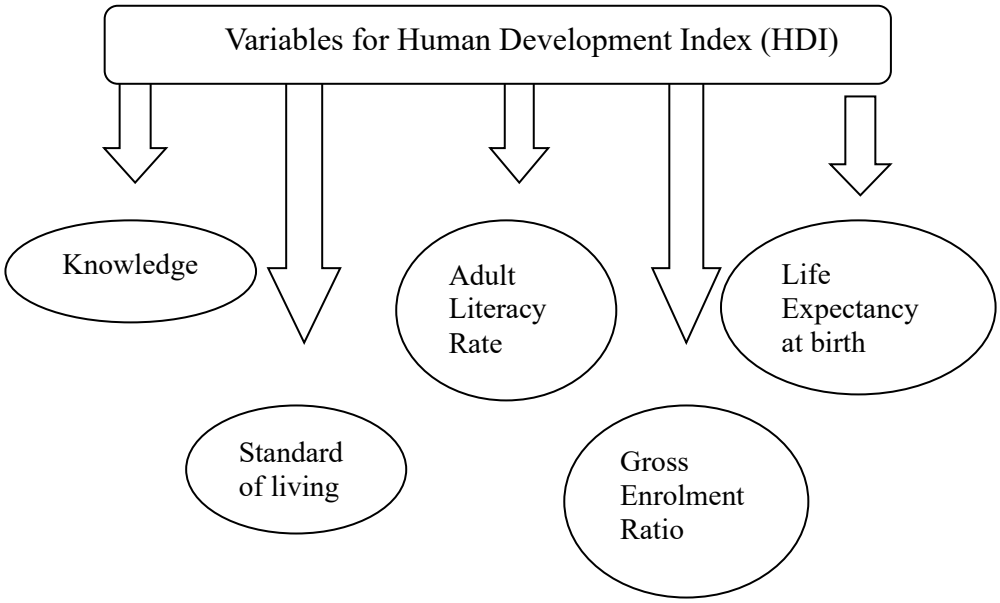


Figure 1.3-1 Variables of Human Development Indicator

The Human Development Index are basically of two types, i.e., social and economic development indicators. Social relates to the quality of life of the individuals and the wealth of individual are determined by the economy within a country. Education, health, employment and unemployment rates and gender equality are the major social indicators of development. When the major economic indicator of development is the Gross National Income per capita.

The development should be considered more than the citizens' well-being improvement. The development shown by Potter et al. (2012) improved our life's

conditions. The information on political, social and economic capacity is passed to provide the long-term basis function. The long-term basis for development requires breaking the circle of low Human Development and conflict, i.e., ending the "conflict trap" with low Human Development increasing the risk of crisis and conflict lowing development (Kim and Conceicao, 2010).

The objectives of this research are:

- To identify weighted correlation networks and patterns in the networks of World Indicators that might conduct to socio-political crises.
- To develop a risk assessment model based on the graph convolution network that can analyse the fingerprints of possible variation of network of World Indicators.

This research will be conducted based on the hypothesis that World Indicators represent reality from various perspectives such as people, poverty, environment or economy. With a time-series and cross-country analysis, there is a correlation between development and crisis indicators. The World Development Indicators as a source of data is associated in this research with human development as discussed by the UNDP (2015), as the three essentials of development which include the ability to lead a long and healthy life, to acquire knowledge, and to have a decent standard of life.

1.4 Thesis outline and contributions

The World Bank and the other institutions, like the Organisation for Economic Co-operation and Development (OECD), are active in international development, value the data and its role in improving their impact around the Globe. This research project will focus on data that could reduce the adverse effects of development initiatives in low-income

countries in particular, factors that upset the social equilibrium and conduct to death—considering that the development of nations is necessary. Finally, this research aims to reduce its negative impact on the population, based on data. Nagle and Guinness, (2014) describe that the human development data helps give a quality of life benchmark by including education and health. Hence, the potential for the country's future development is provided by the transparency of the country's wealth. An example is shown that a low HDI is obtained by spending low on education, and the government has a high GNP.

This research will enhance the knowledge in the development of quantitative risk assessment for crisis prevention with development indicators. A risk assessment that integrates multi-country risk assessment without assessing each particular country could help evaluate a specific region's risk. This research aims to propose a novel analysis method that, instead of taking development indicators as a separate indicator, will consider indicators connected with their correlation and compare a non-conflict indicator fingerprint with a conflict indicator fingerprint to find with network convolutional, indicators varying and assess the risk linked to the proximity with a "crisis fingerprint". Therefore, experimenting with a new data approach to classify networks based on their patterns or differences. This data approach will compare networks with graph convolution and the networks generated from the centrality measurement of the weighted correlation information network.

1.4.1 Risk in this research

When conducting risk assessments, it is common to utilise a combination of quantitative and qualitative methodologies. Qualitative rating or ranking approaches, for example, could be employed in conjunction containing more quantitatively oriented analytical tools that enable the weighing and prioritising of the criteria. Risk in this research

is expressed numerically using quantitative instruments. Quantitative risk assessments are more transparent and easier to verify as legitimate. Risk assessment quantitative techniques are the emphasis of this work and will assess the indicators that can cause dissatisfaction of the population rising until the point when deadly clashes with the authorities happen. It is used to quantify the risk to individuals expressed by the number of fatalities. In addition, it analyses the validity and robustness of quantitative results by identifying crucial assumptions and risk factors represented by the variation of development indicators that can trigger population dissatisfaction.

Risk in this research is more deterministic than probabilistic because the aim is not to use historical data to estimate the probability of an event to occur but more to explore the variation of developing indicators with a known output represented by the number of deaths. Risk in this research refers to population risk, which can be accessed via country data. The principal output variable used to assess risk is the number of fatalities, which can be clustered as a variable representing the importance of a crisis. The crisis might be economic, social or political; the risk is not that a crisis will arise because crisis is part of how a country works, rather the deaths it could cause. The input variables will be development and governance variables estimating socio-political and economic risks. Societal risk is part of the country's risk, together with political and economic risk. Death can be avoided when a crisis is unavoidable, and this research assessed the indicators that could trigger the deadly crisis.

1.4.2 Contribution of this research

A variety of risk assessment approaches that are mutually compatible can be applied together. This research identifies various methods that might be appropriate for assessing risk in the particular context of countries' historical macro data.

Novel country risk assessment method

The World Bank's systematic investigation of variables impacting a country's balance of trade and, consequently, its ability to repay external debt dates back to the late 1960s when national risk analysis first appeared (Avramovic, 1964). Debt servicing capability can be assessed using a combination of short-range and long metrics. They looked at the following short-term measures of such a country's ability to finance the external debt, which is linked to liquidity. This research proposed a country risk assessment method that encompasses the financial risk and assesses the effects of profound economic, social and political changes on society.

A data model that associates a CNN with a statistical method

The use of convolutional neural networks (CNNs) for image identification, including hyperspectral data classification and video classification, has become widespread. Building optimal networks that meet specific criteria is easier when using a statistical technique to optimise CNN models' performance. A wide variety of optimisation issues, including tuning perceptron neural networks' parameters, have been addressed through statistical experimental designs (Patel and Bhatt, 2018). The newly developed model will finetune a CNN model with a statistical method that allows a quantitative assessment to verify the accuracy of the outcome of the quantitative model. The accuracy of the data model will mainly be qualitative to verify if the risk identified by the model corresponds to what has been reported by the media or researchers.

Image-based CNN method applied to correlation neural network

Graph data, which contains much relational information on the relationships between parts, is required for many learning tasks. Predicting the protein interface and categorising the diseases requires a model that learns from the graph inputs when modelling physics systems or learning molecular fingerprints. The reasoning behind extracted features (like scene graphs in photographs) was a significant study issue in other fields, like learning from non-structural data such as images and texts. A correlation neural network represents a graph describing the relationship between indicators that can be transformed into an image using the Eigen centrality. The convolutional neural network will apply to find similarities between images representing the relationship between Development indicators from the perspective of crisis, created with the Eigen centrality measuring the influence of an indicator compared to the others.

CHAPTER TWO: INTRODUCTION TO DEVELOPMENT INDICATORS, CORRELATION NETWORK AND GRAPH CONVOLUTION

2. Theoretical background

2.1 Introduction

The Sustainable Development Goals Report 2018 identifies the interconnection between development goals and the interconnection of change problems like climate change, inequality, or conflict (Liu, 2018). The interlinked nature of development problems reflects a linkage between development indicators. In recent years, it has been evident that factors such as climate, food prices, and conflict rates are co-related to one another. An increase in food prices leads to a rise in the conflict rates and vice versa, as these two factors are related directly. Dry climates with very little rainfall have also led to an increase in the number of conflicts taking place, while it also has a profoundly negative effect on food prices, as less rainfall leads to lower crop yields, leading to increased food prices. In the World Development Indicators (WDI), data on precipitation has been missing for several countries and many years. Therefore, the results may change with new variables entered. This research intends to propose a novel analysis method that, instead of taking development indicators as a separate indicator, will consider indicators connected with their correlation and compare a non-conflict indicator fingerprint with a conflict indicator fingerprint; to find with network convolutional the indicators varying and assess the risk linked to the proximity with a "crisis fingerprint".

2.2 Crisis and socio-economic indicators

As per the WDI, it has been said that Sub-Saharan Africa is a region of the world that requires foreign help; the World Bank stated in 2018 that Sub-Saharan Africa is inhabited by people who suffer from extreme poverty. Goldsmith in 2001, made a worldwide observation

and discovered a correlation between foreign help, signs of democracy, and economic liberals. Goldsmith also quoted in 2001 that democracy impacts economic growth. This case study states that foreign aid and funds are directly related to the development of the economy. The variables of the tell-tale signs of economic growth, foreign help, and democracy which are a part of the WDI are not independent. Thereby, the relationship between the variables of WDI was examined and taken care of in this research. Just as one variable can influence another, a group of variables can influence another group of variables. For example, a significant amount of money for ensuring human protection is an excellent predictor of a good economy. Per capita income, life expectancy, health, infant mortality and literacy rates are all ways to measure the efficiency of a healthy economy.

2.2.1 Crisis assessment framework and methods

To evaluate the weaknesses of states during pre-conflict, active conflict, and post-conflict, the Conflict Assessment System Tool (CAST) is used, which the delicate state's index is based on (The Fund for Peace, 2014). The framework of the CAST system is based on 12 tell-tale signs that give states insight into the conflict. It also consists of characters that are a part of the WDI data, like the "Economic decline". This type of framework provides various opportunities for improvisation by the researchers.

For example, Liu et al. in 2018 had improvised on the CAST system by adding a different feature of climate to it. Liu et al. in 2018 had mixed the Gray Relational Analysis (GRA) with the entropy method to recalibrate the weight present in the indexes. As a result of this hybrid method, accurate results were achieved, keeping in mind the missing data and the variance in the heaviness of the diverse pointers. The output of the entropy technique of GRA was compared to the output of the standard GRA technique, and the results were astonishing. If it were to be compared to a different statistical method of its calibre, like the fuzzy set theory

of Zadeh discovered in 1965, its effectiveness would be visible. Berger-Schmitt developed an additional method used to measure social unity in 2002, which put more emphasis on the social signs instead of the economic ones. The social capital dimension and the inequality dimension are the two dimensions that are consisted of in this method.

2.2.2 The Social and economic indicators

The researchers found a strong association between social problems and economic fluctuations. Savun and Tirone in 2012 had found that social instabilities, such as civil wars, famine, and natural disasters, correlate with economic indicators related to people's income. They identified capita income of an individual as the very superlative instrument to predict a public war in low-income countries and based it on the article of Fearon (2007). The level of democracy, the ethnicity of the people, and the level of inequality are all factors related to the other indicators related to the per capita income. With the help of a survey that was conducted by the probers at the Pew Research Center and Vijaya et al. (2018), it was concluded that even if there was a correlation between lower socio-economic status, low-income, and violence at the macro level it was not prominent in the survey results when answered by individuals. The researchers investigated the factors contributing to individual economic status and the economy of the country as a whole, and their interaction with each other. They found out that factors such as unemployment and a low education level contribute to violence. So the research done by Vijaya et al. (2018) supported the investigations done before it that said that lower-income is directly related to violence.

2.2.3 Other indicators

In the previous year, per capita income was used as an essential indicator on the topic of violence prediction. However, it has been suspected that if combined with other hands, it would provide an even more pinpoint and accurate result. To find out why countries with the same income per capita could experience different levels of violence in their countries, Tebaldi and Alda in 2017, decided to research the contributing indicators that may have a part in this. They discovered that a contributing factor that plays a very significant role in the level of violence in countries with the same income per capita was the quality of the institutions present. With the help of the Corruption Perception Index of Transparency International, the quality of the national institutions can be measured. They also discovered that if the countries had already experienced violence in the past, it is very likely that they would experience it again. It is also to be noted that some indicators are regional; for example, the temperature in Africa is a determining factor. Van Weezel in 2019, found a strong correlation between the change in temperature and the amount of violence that takes place.

2.3 Correlation network

Analysing development indicators often requires a correlation analysis among indicators to assess their impact on the population or the economy. The need to investigate the correlation not only between indicators, but as a network, emerges. Without digging deep into the macro or micro-economics, the economic and socio-political data describing a country can be represented as networks connected with their correlations. Figure 2.3-1 illustrates the correlation of 12 development indicators.

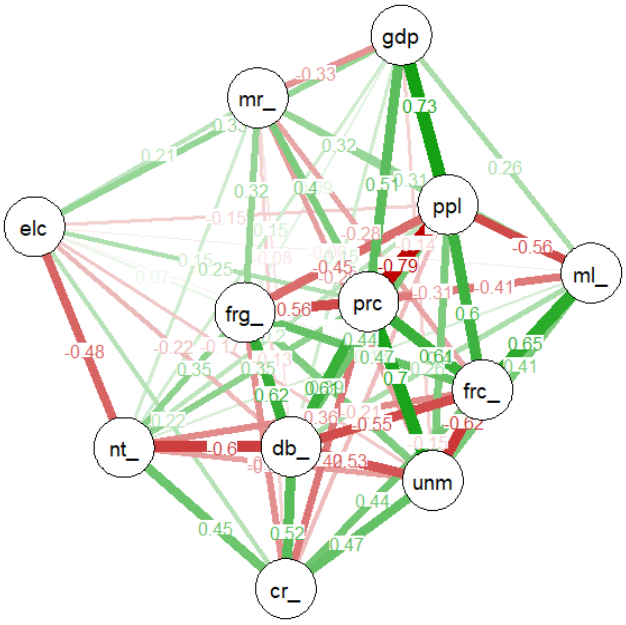


Figure 2.3-1 Correlation network of 12 development indicators from 2016

Statistics and the science of algorithms have been used with complex networks since the twentieth century to describe systems. Albert and Barabási (2002) review different categories of networks, the network of the internet and the World Wide Web, with a topology that involves routes, domain names, or computers. They also described a more specific network, like the network of human sexual contact, to estimate the propagation of sexual disease; or the citation network to represent the citation patterns of publications.

Mathematically, a network is a collection of edges joined by nodes. A graph with the same number of edges in every node is known as a regular network. A complex network is considered irregular because nodes do not have the same number of edges, and nodes or edges might not be equal. These are represented in Figure 2.3-1, where nodes have a different number of edges, and the edges representing their correlation have different weights. Newman (2003) describes specifically the categories of a network. The four types of networks that Newman (2003) organised into are technology networks, networks that are social, networks that are

biological and knowledge networks. The network that best describes the relationship between development indicators might fall under the information categories.

Development indicators are linked together in some ways. It might be, theoretically, by causality or by correlation. The data linking makes up the information network, where the edges connecting each indicator are their correlation. However, the development indicators network does not consider the causality, as it is mathematically hard to show (Mooij et al., 2016), and theoretically, similar indicators will most likely have an almost perfect correlation. In such an information network, where indicators are connected by their correlation, the weight of the edges varies between 0 and 100, represents the weight of the connection between two indicators. The network category for this research could be labelled "weighted correlation information network analysis".

2.3.1 Weighted correlation network

Complex network has been extensively used to represent networks of objects describing their logical relationship. This research enlarges this approach by representing a network of objects by how they all interact regarding a specific outcome. Analysing the development indicators from the prism of crisis gives a unique perspective of the interconnection of development indicators. Weighted correlation network analysis has been widely used to study biological networks. In bioinformatics, a node might represent a molecular, an edge might represent a gene, while the correlation coefficient will reflect the behaviour between nodes. That is described in detail in applying a correlation-based network to study cancer cell metabolism (Batushansky, Toubiana and Fait, 2016).

Langfelder and Horvath (2008), in developing a R package for gene network analysis, listed eight analysis goals of correlation networks. The first goal is finding clusters of interconnected nodes, with a correlation coefficient identifying nodes that tend to cluster

together. The second analysis goal is to reduce a group of connected nodes to a representative, for example, centrally located nodes, to mitigate the problem of multiple testing. The third analysis goal is to identify a significant group of nodes by measuring and selecting nodes with high importance to identify groups with high average node importance. The fourth analysis goal is to tag all nodes based on their proximity to clustered nodes. A fifth analysis goal is to identify neighbourhood nodes of a group of interconnected nodes, to find nodes interacting with a specified group of nodes. The sixth analysis goal is to search for nodes based on node selection criteria, which may be based on a measure of the significance of the node. The seventh analysis goal is to identify changes in different networks, changes in the connectivity, or the structure. Contrary to the previous goal, the eighth analysis goal is to find similar groups of nodes in two or several networks. In this research, the last two analysis goals, which compare several graphs to find patterns or differences, were boosted with graph convolution.

In his book, Networks: An Introduction, (Newman, 2010) identified the visualisation of a network as the first step in analysing its structure. Analyse visuals of networks that have over a hundred vertices is complex with a human eye; a measure of the centrality of networks will reduce the network to essential edges.

2.3.2 Centrality measurement

An essential and demanding research question is how to identify the most critical nodes in a complicated network. Various complicated network centrality measures have been devised to address the issue, but each has its own set of drawbacks and restrictions. A multi-attribute decision-making problem can be used to exploit the advantages of multiple centrality measures. The fundamental idea is to use the grey relational analysis approach to dynamically give the appropriate weight to each attribute (Yang, Liu and Xu, 2018). In more recent network analysis methods, the identification of patterns, key nodes, and network changes are aided by the

centrality analysis process. A recent example is the Eigenvector centrality metric's use to determine the principal pathways responsible for communication in biomolecules (Negre et al., 2018). It exists a different measure of the centrality to measure and compare networks.

Degree centrality, closeness centrality, and betweenness centrality are some criteria used to rank nodes in complex networks. These centrality measurements show different approaches to expressing the importance of a node. A node's importance is measured by how many neighbours it has, but the global network structure is not taken into account while computing the degree centrality value. The number of shortest paths via a node is used by betweenness centrality to identify the most influential nodes, although in most real networks, the information does not move along the shortest path. Non-centralised networks are not suited to closeness centrality's evaluation of node importance based on the ease of access to other nodes. Up to this point, several centrality measure apps have been developed to find significant nodes, but most of them only used a single indication to evaluate nodes' importance; thus, these methods have their own shortcomings or limits. (Kitsak *et al.*, 2010). Newman (2010) described several measures for quantifying networks, like the degree centrality measuring the connectivity of nodes in the network; or the betweenness centrality, measuring how information passes between nodes. Each centrality measurement will give different results for the same graph; it is crucial to identify the size that best reflects a development indicators network and a comparison with other networks. The measure that was principally considered for this research is the eigenvector centrality. For Newman (2010), the eigenvector centrality extends the degree centrality, while for Negre et al. (2018) is a mix of the degree and the betweenness centrality. The eigenvector measures the number of connections of a node in a network and its importance in the information flow. Measure the centrality is indispensable to identifying influential nodes in a network. It is on the centrality measurement that will apply the graph convolution.

Reviewing the different graphs described by Liu et al. (2018), this research contributes to drawing edge-informative diagrams, with development indicators as nodes and their correlations as edges. This standard informative graph will apply the propagation steps of the graph convolution to get the hidden states of networks.

2.4 Graph convolutional network

A perfect example of deep learning is the neural network called Convolutional, which prominently helps in image classification. A convolutional network uses filters to find and match patterns between images. The filters analysed the contours of pixels, combined them to represent the structure of ideas, and then proceeded with the comparison.

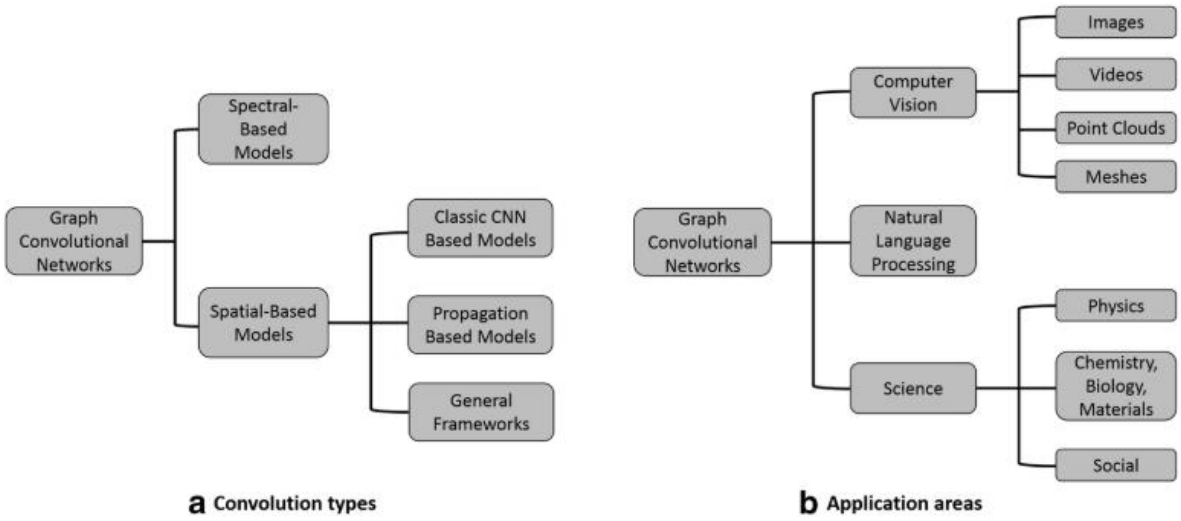


Figure 2.4-1 An overview of graph convolution networks (Zhang et al., 2019)

After several filter applications, the convolutional network will accumulate enough details to identify a form on any image. Zhang et al. (2019) identified two categories of graph convolution networks (Figure 2.4-1), the spectral-based and the spatial-based models.

Before describing the graph convolution, it is essential to understand what the convolution theorem is. In the Fourier transform, which transforms an image or audio into a signal or frequency, processing the call and converting it back to an image could give for

example sharpening image. The convolution in the Fourier transforms states that the product of two functions in the real space is the same as the product of the functions representing their frequencies. Taking the example of a continued image transformed into frequency by the Fourier transform, the convolution of two images taken at a different time is the convolution of their frequencies. The Spectral-based convolution defines the Fourier transform from a frequency point of view by the eigenvalues function. The method has a solid mathematical background, and often, the spatial-based is preferred to it. An extension of the Euclidean convolution, the spatial-based convolution is less complex than spectral-based convolution (Balcilar et al., 2020). The spatial graph convolution to define the property of a node aggregates information from neighbours' nodes. This framework applies to this research in its simple form, the Convolution Neural Network (CNN).

2.4.1 The Neural Network of Convolutions

A Convolutional Neural Network (CNN) used filters to break input image into smaller pieces before processing feature extraction. That could help reduce an image to recognise a nose or a mouth for image classification or object detection. The CNN is composed of three types of layers that can automatically and adaptively extract features from an image (Yamashita et al., 2018).

2.4.1.1 The convolution layers

The convolution layers perform feature extraction by scanning the input and performing convolution operations to produce as an output a feature map. Based on the characteristic of the filter (kernel), the feature map might have different outputs. For the case of an image, the filters will extract from the initial image several features map, representing a unique object or colour (Figure 2.4-2)

2.4.1.2 The pooling layers

The three-dimensional size of the featured chart is decreased with the help of the pooling layers. For example, let us take the feature Max pooling. Max pooling reduces a feature map by extracting from this map the maximum value to create another feature map and discarding the rest. The principle is the same for the Average pooling, which will use the average value of the feature map to create another map. This down-sizing is represented in Figure 2.4-2.

2.4.1.3 Layers that are fully connected

Layers that are fully connected serve for classification tasks. The feature maps created after repeated convolution and pooling are connected in a network of layers based on the weights of every layer. Last, the activation function will analyse the network of layers and will use it for the classification by assigning a probability for an object on the image.

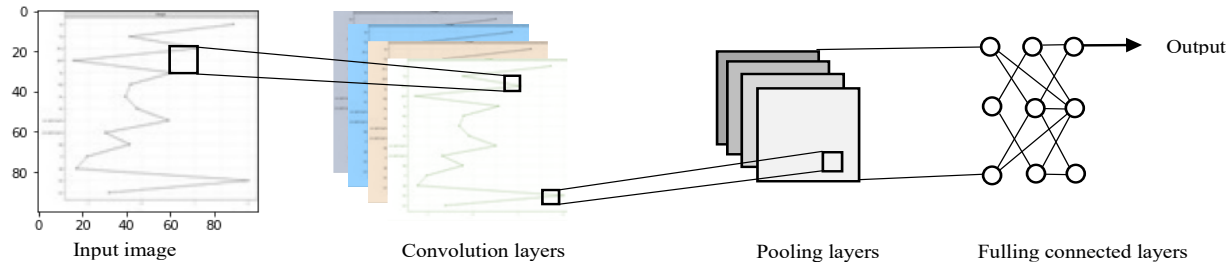


Figure 2.4-2 CNN architecture

The spatial graph convolution applies to this research, precisely the CNN, a successful model in classifying images representing networks (Zhang et al., 2019). A challenge to consider when using the CNN model is the analysis of indicators with missing data in the time-series, or for some countries, which might lead to a wrong interpretation of processed layers of the graph by the activation function.

2.5 Risk assessment

Based on the idea of risk matrices, the trio of Bao, Li and Wu (2018) have created the latest advancement in risk assessment methods. Their practices are based on situation-related analysis to assess the risks. Bao, Li and Wu added the fuzzy set theory to accumulate multiple individual risks to this risk assessment method. Firstly, the risk matrix will estimate and evaluate the personal risk rating with the axes being consequence and likelihood and then numerous individual risks will be accumulated with the help of the fuzzy set theory, and at last, each risk related to a specific scenario would be compared and analysed. It is also essential to include the time-series in the risk assessment method as, over the due course of time, the risk of crises occurring repeatedly may increase or decrease. It was showcased by Duan et al. in (2015) with the addition of the time-series to a regression model that the performance of an analysis that had included the time-series was higher than that of just machine learning. The graph convolution network can compare the signature of development indicators with the one from previous years, distinguishing between years when a crisis happened or not. It can also compare a country's signature with one from a neighbouring country or countries facing the same challenges.

The framework to assess the risk that shows some growth with the computation of risks to improve decision-making is the quantitative risk assessment (QRA). It has found itself a wide range of uses in many sectors such as project management, healthcare, and Information Technology. With each industry customising it to meet their own needs and to assess risks present in their respective industries.

2.5.1 Quantitative risk assessment

A QRA has been extensively used to assess the risk of an accident. This approach tests different scenarios and quantitatively estimates the likelihood and consequences of events as

risk. QRA can then help to predict the risk of a socio-political crisis based on development data. A decision in development is associated with risks of crisis that can involve death; the goal of the QRA coupled with graph convolution is to identify and describe the combination of indicators that can lead to crisis and analyse the probable effects.

United States Army Corps of Engineers (2018) described a list of methods for quantitatively assessing risk. Some methods might apply in this research and were experimenting. From their descriptions, four methods might be coupled with graph convolution to predict the socio-economic crisis.

- **Root Cause Analysis or Loss analysis**: This technique is used to identify the problems at the base level and the impact of an incident. It investigates the incident and how it can be prevented.
- **Sensitivity analysis**: This method examines how variation in the outcome of risk assessment can be distributed based on its quality or its quantity and also between various origins of unpredictability and natural variability.
- **Markov Analysis**: This method can analyse the reliability and availability of systems whose components highly depend on each other.
- **Bayesian Statistics and Bayes Nets**: The total and overall probability can be established by combining information that is already known and succeeding information with the help of this method.

Social scientists and economists, use different quantitative methods for assessing political risks; they are mainly based on country assessment.

2.5.2 Country risk assessment method

The task of assessing investment-related risks dates back many years. During the 1980s, the methods by which a country's risks were assessed were classified into five categories by Desta (1985). Some of the shortcomings and weaknesses that these groups have in them have been fixed by their writers, but some have been resolved by today's modern computers.

- **The ranking-order approach**: In the ranking method, the nations are ranked in numerical order. The signs and indicators are not universal among different ranking companies, which leads to it being impossible to be compared with other rankings. This is as the basis on which the countries score and are ranked is not revealed, thereby making it very hard to replicate. These are the flaws in the rank ordering approach.
- **The decision-tree approach**: The main intention behind the decision tree approach is to detect and predict the future occurrence of political events. It uses methods that are used to probate the possibility of future risks. However, its main drawback is the amount of time consumption required to accumulate and put together the necessary information.
- **The multiple regression analysis**: It is also used to predict political risks, but with the help of taking into account what has occurred in the past and probating the future accordingly. But its heft dependence on information that has to be constantly updated from the field leads to its downfall, as even slight complications in understanding the variables may lead to significant implications.
- **The discriminant analysis approach**: The way in which this method makes predictions is by using the discriminant coefficient with the variables that are available to it. This system has failed to predict most of the upcoming political risks

that were presented to it, so instead, it is being used as a system that gives early alerts in case there is a chance of future risks.

A connection between the amount of corruption and the amount of conflict taking place externally in relation to the Foreign Direct Investment flows was found by Al-Khouri and Khalik in 2013. In turn, this has an effect on the likeness of an investor willing to invest as the environment may lose its approval as it may be related to many political risks i.e corruption and external conflicts. This may deter away investors, as they would not be willing to invest in areas that are highly unreliable.

2.6 Economic Vulnerability Index in the least developed countries:

Economic instability can be described as the possibility of an unpredictable exogenous shock preventing economic growth in a country. For Development Policy's three-year assessments, the Economic Vulnerability Index (EVI) was among the three parameters of the UN Committee of the Least Developed Countries (LDC) list (UN-CDP) in the year 2020. This indicator substituted the 1991 economic diversification index, which represented less economic weakness. GDP per capita and the Human Assets Index (HAI) are also used as distinguishing criteria. For 2006 and 2012, analyses of the list of LDCs and the EVI calculation technique are updated (Cariolle, Goujon and Guillaumont, 2016). Economic weakness could be understood from 3 significant factors: exogenous (both foreign or natural) shock size and probability, shock exposure, and shock resilience (or reaction capacity). Whereas the two former predictors rely mainly on the systemic characteristics of the country (e.g., foreign market volatility, geographical position), resilience adapts to the existing economic policies of the nations, along with the levels of per capita income and human resources, that are taken into consideration independently for the classification of the LDCs.

The variations in the EVI of LDCs and non-LDCs are mainly due to the shock index differences. Similarly, exposure rates are dropping, in fact, for non-LDCs a little less than for LDCs, but owing to two factors that are not necessarily indicative of a related structural shift in LDCs: greater population growth; and a slower growth in the population in low elevations, with the 2012 architecture. This enhances the diagnosis that LDCs are less fragile in their systemic economies. Obviously, the median changes discussed here mask differences between LDCs and non-LDCs. The stronger EVI decreases in LDCs are caused by a reduction in the number of repeated shocks, not by changes in their structural exposure.

2.6.1 Vulnerability of entrepreneurs

Entrepreneurs are known for developing innovative projects that address deals with issues. It must be generalised in order to conceive of entrepreneurs as effective agents of societal shift capable of changing their communities. That does not mean taking time off work to fix social problems. It means deliberately integrating social objectives into the strategic planning of entrepreneurs as a means of optimising personal and mutual benefits. If that abolishes the concept that only governments and developed corporations are capable of discussing social needs, it also means that entrepreneurs are vulnerable to social and economic instability. Entrepreneurs who work closer to the people are in a great position to recognise and resolve crucial concerns, such as poor access to health and education.

The drastic reduction in spending amplifies public discontent, especially during crises. It is an opportunity for entrepreneurs who can capitalize on these grievances to create new domestic opportunities that help improve social well-being (Savun and Tirone, 2012). The positive impact of entrepreneurship beyond expectations is difficult to understand. However, it is explained by the improvement of the resources allocated to the business sector or its contribution to reducing poverty.

2.7 Conclusion

The study highlights the hidden patterns found between socio-economic and political crises that happened in developed countries. The study prefers the convolution neural network to make the classification and identify the patterns of socio-economic development indicators. The study focused on crisis and socio-economic indicators such as democracy, economic liberals and foreign help. It is significant to determine the study's crisis assessment framework and methods. Assessing the development indicators needs a correlation analysis over the indicators to determine the impact on the economy. Correlation network, weighted correlation network, and centrality measurement are the risk assessment methods and techniques. The correlation analysis is similar to the multiple regression analysis used in the country risk assessment method described above. The regression analysis experimented in Chapter seven shows the limit of this method, therefore, justifying the need for a novel method. Many studies revealed that graph convolution would be used to identify and describe the combination of indicators that can cause a crisis and evaluate the probable effects. Graph convolution associated with correlation analysis gives information about quantitative risk assessment methods used to enhance decision-making. To determine the political risks, economists and social scientists used different quantitative methods, principally statistical, to identify what can be the most indicators at risk. This research experimented with the P-Value to determine the relevant socio-economic and political indicators.

CHAPTER THREE: SYSTEMATIC AND LITERATURE REVIEW

3.1 Systematic review

3.1.1 Questions for review

Suppose development, social and political crisis research seems to be a highly dense and different hypothesis that has been tested based on evidence. In that case, the literature on complex network and graph convolution is more recent and specific to some sectors. The literature review gathered literature from scholarly journals reviewed by peers who addressed three questions.

i. What is the selection of development indicators that might conduct to socio-political crises?
ii. What are the patterns in the correlation network of development indicators?
iii. Are the risks of brutal social crisis assessable with graph convolution?

The questions presented above have most of their focus on low-income countries, as they are the most susceptible to long-term damage when a large-scale crisis occurs. The outcome of this research includes a system of indicators and a risk assessment procedure.

3.1.2 Review method

This part gives a basic understanding of how it would be reviewed, describing the criteria according to which the literature selection process takes place and the search strategy executed in the selected journals. The first review identified relevant literature that matched the full inclusion criteria by reading the title and the abstract. Reading the entire paper assessed its relevancy for literature not fully matching the inclusion criteria.

The systematic review of the literature used the PICO (Population, Intervention, Comparison, and Outcomes) framework.

3.1.2.1 Inclusion criteria

A wide range of literature involving social crisis or development indicators is available. The selection criteria refined this broad literature to confirm the most proper and apt study/research on this matter.

The population of studies: Countries that are in the process of development and upcoming are all the main targets of this study, and the people of these countries have the following characteristics in common:

- Require external assistance to develop and have a low amount of income
- They were categorised as states that were delicate and uncertain
- Where a part of a socio-political crisis that we related to the deceased

Intervention: The intervention of the systematic review is on the DI data refined by population, which includes articles about DI linked to the deadly crisis.

Comparators: The comparator of this systematic review is the literature available for developing countries compared to the literature available for developed countries or countries in an exceptional situation like war. Another comparator is the type of crisis described in the literature. Some crises are socio-political while others are economical; the deadly aspect of a crisis is not always addressed in research, which will also be a comparator.

Outcomes: This research would have found a relation between dataset and time-series to find out about the issue within the WDI that has led to problems with assessing risk related to the social crisis based on DI. Thereby time-series is an essential aspect of this study. This analysis specifically reviewed the following literature:

- Association/relation between progress indicator variables and reporting of events, and data, over time or by region

- The repeated series of indicators for growth before the crisis
- Moreover, the social crisis risk assessment focused on growth indicator research.

Type of publication: Generally well-critiqued articles from academic journals were the main focal point of the systematic review. The names of these journals were mentioned in the E-journal's list of the UEL (University of East London)

The date of publication: The research was not bounded by articles from a specific period. But most of them were articles that had been published between 2015 to 2020.

Language: There was no barrier to the languages that could be used. The research contained articles written in many different languages.

Location: The main focus was put on developing and low-income countries. The research, although worldwide, had focused much of its attention on areas experiencing conflict and crisis in recent years, such as sub-Saharan Africa, the middle east, and South American countries.

3.1.2.2 Strategy for search

The terms of the search request inserted as keywords in the research databases are made up of any research query translated to a bag of words.

Question No 1: Combination; Selection; Indicators of World Development; Fragile; Social; Political; Crisis

Question No 2: Patterns; Correlation network; Indicators of World Development; Socio; political; crisis

Question No 3: Fatal; Social crisis; Assess; risk; Indicators of World Development; graph convolution

Some keywords were combined with synonyms to increase the accuracy of the search.

- Association: correlation, coefficient, linear
- Crisis: Critical, conflict, Change, pressure,
- Development: official development assistance, low income, least developed
- Correlation network: graph theory, graph network
- Fragile: weak, delicate, uncertain
- Cohesion: Cooperation, participation,
- Deadly: violent, fatal, harmful, dangerous, brutal,
- Assess: Evaluate, determine, estimate
- Predict: Forecast, anticipate, determine
- Graph convolution

3.1.2.3 Journals

The main focus of this systemic review was well-critiqued articles that provided helpful insight on indicators of development such as data mining, assessment of risks, and social movement.

Social/Political crisis – Social movement

- SAGE - The journal of conflict resolution
- Taylor & Francis Journals Complete

Development indicators and computing

- EBSCOhost - Business Source Complete
- ABI/INFORM Collection

Data mining and risk assessment

- Wiley Online Library

- Science Direct
- IEEE Xplore
- ACM digital library

3.1.3 Systematic review results

3.1.3.1 Search results

In the eight online journals, 251 pieces of literature are found by entering the requests reflecting each research issue. Appendix A summarised the results obtained for each journal from the question entered with the refinements.

A final 206 journal articles were obtained when all the results were combined and were freed of wrong and duplicate entries. An increasing amount of interest was shown on the graph convolution according to the analysed and studied results (Figure 3.1-1). Due to the uprising of increased interest in the development of data, the quality and quantity of data being provided have drastically improved and increased to keep pace with the current socio-economic goals. Along with this, there also has been an increase in the interest in risk predictability and the amount of data received from the WDI (Table 3.1-1). There are a few articles related to the patterns in correlation network and development data related to the crisis.

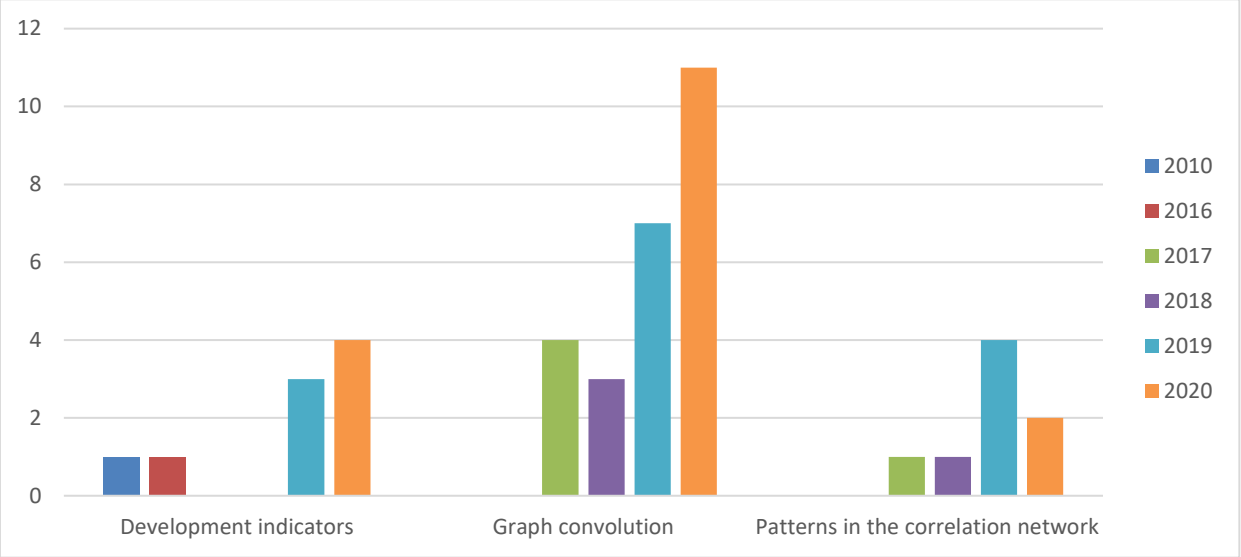


Figure 3.1-1 Number of articles refined by interest

Interest	**2010**	**2016**	**2017**	**2018**	**2019**	**2020**
Development indicators	1	1			3	4
Graph convolution			4	3	7	11
Patterns in the correlation network			1	1	4	2
Total	**1**	**1**	**5**	**4**	**14**	**17**

Table 3.1-1Number of articles per year

3.1.3.2 Selected articles

A deep and detailed review of the 42 articles has been provided by a first hand view of the 206 articles. The main and primary topic that encircles every article is referred to by the interest. In order to answer the questions related to the study about the indicators that are linked to the social crisis, aid is received from articles in the interest of "Development indicators".

When the articles related to the "Graph convolution" are linked to the question on the graph convolution and crisis assessment; and the last question patterns in the correlation network of development indicators. Articles that provide a good amount of intel and contribute directly to answer the research question are known as articles of high relevancy; articles that do not contribute directly to answering the research question while still providing a decent amount of information are known as articles of average relevancy and articles that do not provide any amount of information related directly to the research question but still are overall somewhat valuable to the research are known as articles of low relevancy. Nevertheless, no articles left in the initial review have a high enough relevancy to aid the research if the amount of interest/fascination related to the data and prediction of risks grows exponentially over the course of time. In this case, 6 of the articles contribute directly to the research questions, and the remaining 4 have information that may not contribute directly to the research question but the research as a whole.

3.1.4 Conclusion

Aside from the World Bank data, countless amounts of data from other institutions are also collected by the WDI, creating valuable data that plays along with the WDI's themes. With the help of the time series, the data collected is also universally comparable to data from other countries. A tremendous amount of interest has been invigorated worldwide on the matter of sustainable development or crisis prevention by the WDI. A very tiny amount of research has previously been done with development data on the subject of quantitative risk analysis. It is contrary to the approach of the method taken by most of the researchers investigating to understand the variability and value of the data found in the WDI. The systematic review had proven this true as it was found that it did not directly respond to the research question if the risks of a social crisis were already predictable with the help of the data from WDI. Therefore,

this study's main focus is creating a quantitative risk assessment for the prevention of crisis with the help of development data. This study was done assuming that the WDI directly correlates to the interests and situation of the state, looking at it from different angles such as poverty, the people, the environment, and the economy. The help of time-series and cross-country observations is also to be taken in this study. It is also to be speculated that crisis indicators and the WDI are linked to one another; depending upon the study and observation of the socio-economic variables, the impendent rise of a violent social crisis can be predicted and pre-judged. The cross-impact matrix method is the aptest method to be used as it predicts the probability of an event occurring by judging and calculating the other events. It is a well-known forecasting method that assesses the upcoming risk.

3.2 Literature review

This research has examined hidden trends between the difference in development indicators and the dangerous socio-political crisis that has arisen in countries in an attempt to discover out whether the fusion of development indicators contributes to a deadly crisis. It is to understand what shifts in development indicators have triggered intense tensions in countries; to consider, based on evidence, the combination of development indicators that should be tackled as a priority. National and foreign organisations would be involved to help implement growth through lessons from previous encounters.

Several internal and external factors could cause social death. Some economists believe that social disturbances result from economic changes. This research investigates crises causing direct death and puts human life as the limit a crisis should not exceed.

3.2.1 Selection of development indicators

3.2.1.1 Level of economy

The International Monetary Fund (IMF) gives loans to countries in difficult situations, including countries facing terrorism. In issuing the loan, the IMF imposed several conditions that negatively affected the economy of debtor countries (Hunter and Biglaiser, 2020). The Poisson time-series regression to assess the impact of IMF in 115 countries between 1980 and 2013. Hunter and Biglaiser (2020) used economic, political, and social indicators as independent variables to analyse how they correlate with incidents of terrorism. They found that IMF loan has lower adverse effects on domestic terrorism and contributes to reducing it. Associating the economy level with the number of terrorism events, the economy level might be represented by trade, foreign investment, income, or demography. Social scientists agree and disagree on the existence of a correlation between the level of economy and terrorism or its effect. Hunter and Biglaiser (2020), when demonstrating the effect of IMF loans on terrorism, did not explain the indicators or group of indicators that influence terrorism. Suggesting that IMF loan conditions are not the same for all countries, the indicators that have an effect in a country might not have the same effect in another country; because of an external factor like a pandemic, the effects of the IMF loan might differ from one year to another. Analysing the data for a democracy like Hunter and Biglaiser (2020) might show a negative correlation when analysing the indicators for countries experiencing high inequality might give a positive correlation between the level of economy and terrorism.

3.2.1.2 Resources abundance and effects

Natural resource abundance is assimilated into an economic curse because of all the problems it causes, especially in countries with low and middle income. Aljarallah and Angus (2020) investigated in their study the resource curse in a resource-rich countries from the

Middle Eastern regions and Northern Africa (MENA) region from the economic, political and social aspects. The authors used various econometrics functions and models to demonstrate that resource rent has an impact on country growth. In Per capita GDP; Real capital stock per worker; Total factor productivity; education; Institutional quality. Analysing the data from one country (Kuwait), Aljarallah and Angus (2020) give precise effects of natural resources on Kuwait's economic, political, and social aspects. The effects are different in other countries from the MENA region that experienced war methods used to investigate the indicators in relation to the natural resources cannot apply to many countries at the same time as the regression analysis does in the previous section. In the same way, the regression analysis will not provide a detailed analysis for each country.

3.2.1.3 Conflict history

ViEWS, a method for forecasting an armed conflict, has utilised data from Uppsala Conflict Data Program (UCDP) to deter armed conflict (Hegre et al., 2019). The UCDP aggregates data from 123 countries collected between 1898 and 2017. The ViEWS assessed data from African countries to forecast 36 months ahead of various types of state violence. It used random logistic regression and random forest models to identify the conflict's precise location. This early warning system can contribute to identifying risks and alerting the public opinion to prevent them, but it cannot identify the causes of the conflict, so it does not provide sufficient information to prevent the conflicts. Like the models described above, the ViEWS used data related to education, natural resources, democracy, low economy, the demography, including a new indicator, the conflict history, which provides a record of conflict events.

3.2.1.4 -Post-conflict indicators

Changes in development indicators to follow a crisis are also primordial for crisis risk assessment. In a post-conflict situation, people are vulnerable, socially and economically, affected by the conflict. In their literature review, Donaubauer, Herzer, and Nunnenkamp (2019) cited The World Bank in identifying military expenditure as a significant indicator of conflict. The World Bank found an association between military expenditure and social and economic infrastructure in relation to the concept of 'development in reverse.' The social and economic infrastructure includes every four indicators (Table 3.2-1), which give with the military expenditure, nine indicators identified by Donaubauer, Herzer, and Nunnenkamp (2019) correlating with conflict-related data.

The social infrastructure	Education: primary enrolment ratio
	Health: maternal mortality ratio
	Government &civil society: corruption free
	Water &sanitation: Safe water accessibility
The economic infrastructure	Transport: air carrier departures
	Communication: mobile cellular payments
	Energy: consumption of electricity
	Banking &finance: domestic private sector credit
Military	Military expenditure

Table 3.2-1 Post-conflict indicators

In analysing post-conflict data, the authors did not make precise the number of countries concerned by this research nor the size of the datasets used to calculate the correlations; the indicators matched with the indicators previously listed, in addition to military expenditure.

This section of the literature review contributes to the identification of indicators that can be associated with the crisis. Some indicators are shared by several researchers, like education, natural resources, democracy, demography, and the Per capita GDP; other indicators are more specific to some studies like trade, foreign investment, conflict history and military expenditure. A minimal number of studies applied a graph theory to analyse the correlation among development indicators. Graph theory can visually represent economic, social, and political problems. The variation of the development indicators gives a static figure of the probability of a risk in a country. The causes, volatility, origin, and nature of the crisis were identified and analysed in the next section of the research in order to understand it better.

3.2.2 Socio-political risk

Two categories can be differentiated from artificial types of uncertainty: social and political risk and economic risk. Al-Khouri and Khalik (2013) suggested a question, 'Does Political Risk Affect the Flow of Foreign Direct Investment Into the Middle East North African Region?'. Socio-political danger encompasses potential adverse acts or influences aimed at foreign companies arising from any host country's social community, political authority, or government body. Subsequently, there could be more distinctions between social risk, the risk from government policy, and politics. Aljarallah and Angus (2020) provide a case study about the dilemma of Natural Resources Abundance in Kuwait. The communal risk is a joint activity of organisations such as trade unions, Non-Government Organisations, or other individuals who are involved in informal groups, which, calmly or not, constitutionally or not, lobbies the authorities from local or international firms to affect their policies or actions relates to social

risk or the societal relevant risk directly. These demonstrations can be inward, as was the case during the ripening of the McDonald's restaurant in the southern region of France by Jose Bov and the leader of the French farm workers' union. He and other activists demolished what he considered a 'symbol of global imperialism' for more than 100,000 dollars. That can be a reasonably mild form of action. In the worst-case situation, however, the social danger of foreign workers or even the abduction can go all the way to the physical assault.

3.2.2.1 Country economic peril

The Country's economic peril is divided between macro risk (all foreign companies) and micro-risk (focused toward a particular section of action or certain firms). Bao, Li and Wu (2018) proposed a fuzzy mapping framework for risk aggregation based on the risk matrices. Economic risks can result from poor political management, but they should not be an overt consequence of a political decision, as opposed to the socio-political risk mentioned above. The risk involved in macroeconomics denotes the global economic environment instability, such as production, prices, interest rates, exchange rates, and trade conditions. For example, a wave of hyperinflation between the 1980s and the early 1990s has occupied many Latin American countries. Brazil encountered an inflation rate of 50.75 percent monthly in the riskiest scenario in June 1994, which equates to more than 13 000 percent per year. Desta (1985) did a review on assessing the Political Risk in Less Developed Countries. A detailed analysis of country risks disturbs business daily and can be very expensive.

The risk involved in microeconomics includes all the adverse events that can occur at the industrial and business levels. It covers both the risk of resource needs of the organisation to conduct its business (raw materials, work, capital) and the risk of production and marketing uncertainty. Most of these micro-economic unforeseen scenarios can be categorised as risks of 'glocalisation,' i.e. when local features are needed in a global strategic plan. This category

concerns the market environment in the day-to-day activities of a foreign company. All the risks that affect the transactions in the company and its management are specific to the host country: growth, marketing, finance, procurement and distribution, human resources, technology, and organisational structure. It also incorporates unique cultural elements of risk evaluation and the "safety culture" This definition is especially relevant as multinationals export "industrial risks" to developing countries that may be less stringent in terms of this type of risk in the context of local infrastructure, regulations, safety procedures or environmental standards.

3.2.2.2 Socio-economic factors forecasting through machine learning

Perhaps researchers have been enabled by allowing sharing of the parameter with adjacent nodes in a graph; instead of the following repetition, the convolution that is widely used in images is now feasible on graphs. Amidi and Amidi (2020) presented the Convolutional Neural Networks cheatsheet. Convolution is a two-function mathematical function (f and g), resulting in a third function expressing how one type is changed to the other. Instinctively, it is possible to picture the convolution as passing one function along with the other, and the overlapping region will be the resulting convoluted function.

Convolutional procedures can be carried out in spatial (Euclidean) or spectral (frequency) domains. The knowledge of the local barrio graphs in the spatial domain cannot easily be represented; the graphical convergence takes place in the spectral field instead. Balcilar et al., (2020) ‘Bridging the Gap Between Spectral and Spatial Domains in Graph Neural Networks’. The convolution theorem shows that complete graphic convolution requires the structure of the graph Laplacian matrix (L), and Fourier's transformation is then carried out on the graph:

$H^{(i+1)} = f(H^{(i)}, A)=\sigma(AH^{(i)} W^{(i)})$

When a forward pass function f is used to represent graph neighbourhood information using attributes from the preceding layer and graph adjacency matrix A (usually standardised or transformed), then the non-linear $\sigma(\cdot)$ activation function (e.g. ReLU) is used for propagating towards the next layer.

3.2.3 Risk assessment with deep learning techniques

The risk assessment represents the process and methods used to identify and analyse hazards and risk factors that can potentially harm. For the context of this research, it was identifying political, economic, and social factors that can lead to harmful crises. Recently, researchers have developed QRA that integrates machine learning and deep learning techniques.

3.2.3.1 Deep learning techniques and supervised learning

Hu et al. (2019) presented a security risk assessment method that integrates deep learning combined with case reasoning for risk assessment. The authors used a triple CNN to extract features from security risk images for the deep learning method. The features constitute a dataset that is refined with features coming from images of similar security domains and removed features that are not relevant for the security assessment. For the case reasoning, the authors used the nearest neighbour algorithm to calculate case similarity. The process for the case reasoning is described in Figure 3.2-1.

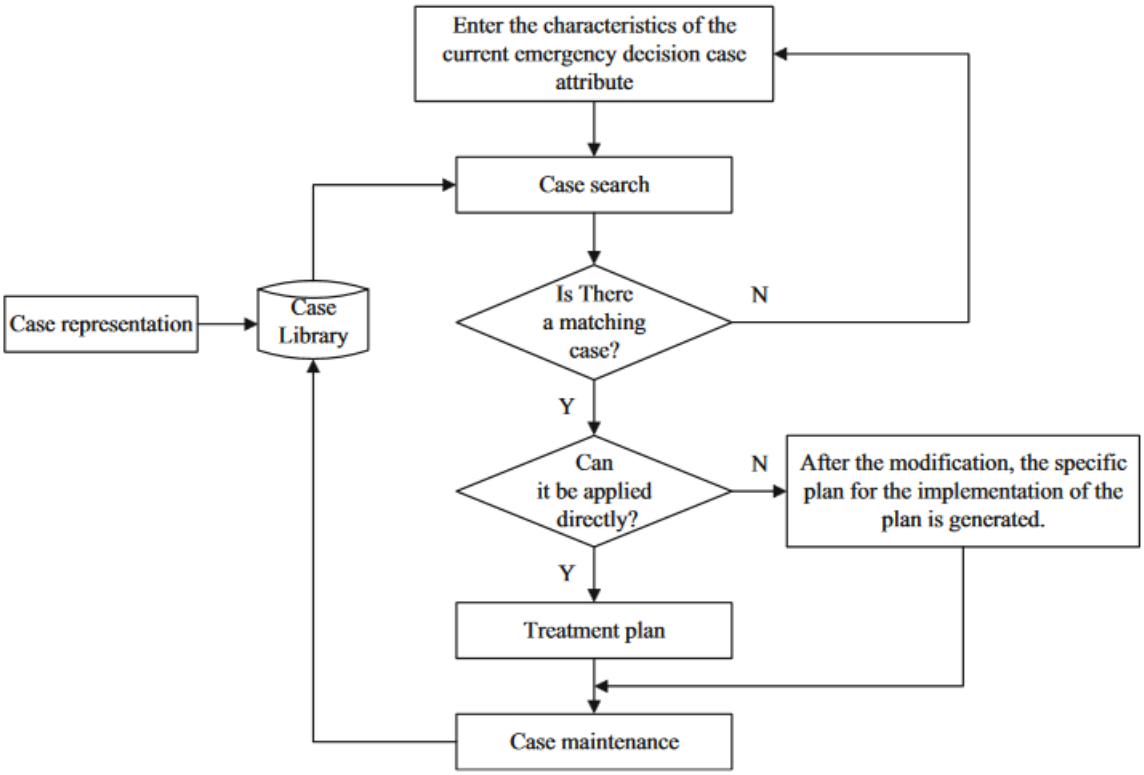


Figure 3.2-1 Case-based reasoning (Hu et al., 2019)

The case representation is the database of risk cases that are stored in the library in the form of images. For a new risk, the model searches in the library to find if the new case's image matches any in the library. It then returns cases of similar risks with their solution. If the new risk is not precisely similar to a case stored, the model suggests an amendment of the risk solution closer to the new case to better respond to the new risk. The new case is then recorded in the library as a new case.

Case reasoning is a method that can quickly and accurately find similar risks and solutions to address them. In the case of risks that can occur in a different environment and at a different time, this method requires using a vast library to store each risk and its solution. This method is also resource-consuming to analyse and store every case. An alternative might be to use unsupervised learning that does not require training and testing datasets.

3.2.3.2 Deep learning and unsupervised learning

Jena et al.(2020) developed a model based on various cluster analysis methods to estimate the risk of earthquakes in Palu city in Indonesia. The authors used Euclidian distance in hierarchical clustering to show the clustering association with a dendrogram. With the hierarchical clustering, the authors used locational clustering to organise event based on geographic coordinates. Lastly, the authors used a Silhouette clustering analysis to study the separation between each cluster. The power of the cluster analysis was combined with a CNN model, trained with earthquake inventory images to extract features related to the earthquake and features not related to the earthquake. With the cluster analysis and the CNN, the Analytic Hierarchy Processing, a multi-criteria decision making (MCDM), was applied for pairwise comparison and weight calculation of the vulnerability.

The Cluster analysis does not require the class to be identified in advance. Instead, the various clustering methods and the criteria of the multi-criteria decision-making attribute a class and a weight to a vulnerability. The CNN will still have to be trained, but the entire model does not need to know in advance each vulnerability. Instead, the MCDM will need a predefined list of criteria, and the authors weighted seven criteria without assessing the relationship between the criteria. The Analytic Hierarchy Processing is ineffective when the criteria are independent (Kumaraswamy and Ramaswamy, 2016).

This research shows a correlation between political, economic, and social criteria that can be used to assess the crisis risk. A model that includes a decision-making process can use the Technique for Order of Preference by Similarity to Ideal Solution (TOPSIS), with the Criteria Importance through Inter-Criteria Correlation (CRITIC) technique to include the correlation between the criteria in the computation (Mohamad, Liyana and Halim, 2020).

3.2.3.3 TOPSIS

Hwang and Yoon (1981) advanced the method for choice of command by contrast with the optimal solution (TOPSIS). The central idea of this technique is that the selected alternatives should be as far away from the positive optimal situation as possible and as far from the ideal negative response as possible. TOPSIS assumes that each parameter continues to increase and decrease monotonically. The perfect positive and negative solutions are therefore easy to explain. In order to determine the relative similarity of alternatives to the optimal solution, the Euclidean distance method was suggested. A sequence of comparisons of these relative distances can deduce the desired order of the alternatives.

One of the critical goals of the effective structure for the industrial strategic planning assessment is the approach chosen for MCA. Chances, such as pre-specified method, to achieve this objective. Crucial points in developing a performance criteria framework are the goals of market penetration, market development- assignment of goods diversification requirements, which are typically converted into acceptable performance for any criterion by management - as well as how values are compared to operating and development planning. The multi-criteria approach used for the comparative paper in this representative measure is based on the principle of the shifted assessment of the ideal relative market position.

The core principle of this theory is a set of companies that produce similar goods and the optimal solution that shows how efficient the corresponding pursuit of the highest performance is. That is why it is sometimes used as a benchmark business operation in a basic mathematical form. In addition, it is a justification for the construction of corporate league tables. In addition, it is claimed that the human desires of the consumer are best articulated in vague terms with regard to such anchor points. These criteria depend on the specific characteristics of profitability and the overall performance of the issue under consideration and may be moved by the enterprise. Several objections to this kind of in- or lower value based on

solid comparisons are linked to the contrast between historical alternatives and the specific performances obtained in the financial statements. In this way, the criteria needed for the assessment process are cost-based. Despite this limitation, the regular source of knowledge accessible to the analyst to continue improving the business's performance is described as motivating the continuously published accounts while taking the current into account a comparative assessment of companies.

3.2.4 Conclusion

This chapter discusses the literature discussing socio-political crisis from a data perspective, publications describing the use of system analyses and diagrams, and articles covering quantitative risk analysis methodology for the socio-political crisis. In the area of risk assessment, graph convolution is usually applied to spatial data. This research is a novel contribution to experiment graph convolution on images generated from a correlation network. Before experimenting with this novel approach, preliminary research investigated with regression analysis the relationship between development indicators as input data and fatalities as output data to assess risk.

CHAPTER FOUR: REGRESSION ANALYSIS

4 Regression analysis of the relationship between development indicators and crisis events in developing countries

4.1 Introduction

Regression is among the most often employed statistical techniques for detecting and analysing the association among dependent variables and one or more independent variables, also called predictor and explanatory variables, as in behavioural and social sciences and physical sciences. Linear regression analysis is used only when one continuous dependent variable and discrete independent variables are included in the model. This study tested how the variables are linked linearly. It investigates the link between multiple independent or predictor factors and a criterion variable.

For finding out the relationship between different variables, the technique of Regression Analysis has been used. The independent variables are constant and depend on how an independent variable is affected by the dependent variable. Shyti and Valera (2018) propose the validity of the regression analysis method on economic performance. Regression analysis is used for prediction and forecasting. The predictive modelling technique is Regression analysis; this looks into the relationship between an independent (predictor) and dependent (target) variable.

Utility of regression

- Useful in Economic & Business Research
- Prediction
- Degree & Nature of relationship
- Estimation of relationship

Types of regression analysis

- Partial and Total Regression
- Simple and Multiple Regression
- Linear and -Non-Linear Regression

Difference between correlation & regression

Degree & Nature of Relationship

- The relationship degree between X & Y is measured by Correlation.
- The relationship between the nature of variables is studied by regression; hence, the prediction is possible on the basis of one variable on another.

Cause & Effect Relationship

- The correlation does not assume the relationship between the two variables.
- Regression clearly expresses the cause and effect. The regression clearly expresses the relationship between two variables of cause and effect.

Prediction

- For making predictions, correlation will not help.
- By using the regression line, predictions are made.

Symmetric

- Regression coefficients are not symmetrical, i.e. the relationship between two variables is not equal and can represented with an equation.
- Correlation coefficients are symmetrical, i.e. the coefficient equally represents the relatioship between two variables.

Origin & Scale

- In Correlation, the change of scale and origin is independent

- In the Regression coefficient, the change of origin is independent but not of scale.

Before introducing a new method of analysing development data with deep learning, this research tested the use of the classical statistical model to verify the pertinence of developing a new method to estimate the risk of crisis with development indicators. The statistical model used here is a regression as a predictive model to assess the risk of social crisis.

As stated above in section 2.5.2, the regression analysis is a classic country risk assessment method that is used to predict political risks. The multilinear regression experiment is a supervised learning method where the fatalities (output data) are known, and the Development indicators (input data) can be used to predict the output variable. As a supervised method, regression has been preferred to classification methods because the research objective is to investigate the relationship between the output and input variables, which is the main objective of this research.

As a supervised method, multilinear regression is a unique method that allows an investigation into the relationship between variables. Multilinear regression is a mathematical description of this process and is given in the equation below.

$$Y = a + bX1 + cX2 + dX3 + \epsilon$$

In which

ϵ – Error (Residual)

b, c, d – Slopes

X1, X2, X3 – Explanatory (Independent) variable

Y – Dependent variable

There should be a minimum correlation between independent variables. In a situation where the independent variables are highly correlated, it will be challenging to determine the genuine relationship between the dependent and independent variables. Regression is the most suitable method that could apply to development indicators where a correlation between independent and dependent variables exists.

The dataset used to test the regression analysis is the initial WDI of the World Bank data, used as the predictors. The dependent variable is the prepared crisis data generated by the ACLED project. The WDI used to test the validity of a regression model to predict the risk of crisis is the initial WDI, identified in the literature review with other indicators that might be relevant. In analysing development data, some researchers identified a strong correlation between socio-political crises and economic changes. This section aims to investigate the relationship between development data and socio-political crisis events in low-income countries from a data perspective.

4.2 Methodology

The regression analysis describes the relationship between the dependent and independent variables. Applying to more than one independent variable, this analysis was exploratory instead of predictive to understand the effects of WDI on the crisis in low-income countries. The regression models experimented within this research were a Poisson regression and the multilinear regression. The results of the two regression models were compared to assess the relationship of variables instead of emphasising the predictability of the models. The linear relationship assumes a normally distributed error, and the value of the predicted value is continuous. It could be negative when the Poisson regression fits better to predict discrete and only positive values (Grace-Martin, no date). For Ramcharan (2006), like many researchers, a correlation among variables does not imply causality. Correlations are common in economics

and quantify the relationship among variables. This methodology did not include a theoretical model that is crucial to analyse the causality among variables but instead focused on the statistical results of the regression analysis to compare with the findings of other researchers described in the previous section.

The methodology of this analysis started by (1) exploring the predictors, analysing each indicator separately, using descriptive statistic techniques to inspect each variable, analysing the normality, and detecting outliers and missing data. After analysing each indicator separately, (2) all the independent variables were analysed together, with multivariate techniques to review the multilinearity among independent variables and the linearity with the dependent variable. The results of the multilinearity were conducted to (3) a factor analysis to avoid using independent variables with high correlation. Lastly, (4) the regression model described the relationship between DI selected and crisis, giving the weight of each DI and the margin of error.

Multiple regression has been increasingly used in research over time. Plonsky and Ghanbar (2018) listed two issues in using a regression model. A low sample size weakens the model's generalizability, and the outliers affect the model's accuracy. To avoid these issues, the ratio of cases to predictor variables should not be lower than 5:1, and the outliers should be removed before proceeding with the analysis. Also, based on the statistical assumptions of a regression model, Plonsky and Ghanbar (2018) recommended that multicollinearity expressed as a high correlation among predictor variables should not be present because it can reduce the predictive power of the variables; for normality, all residuals should scatter around predicted values; for linearity, they should be a linear relationship between the predicted values of the dependent variable. A review of the WDI listed statistically significant and relevant variables for the regression analysis.

4.3 Data description

4.3.1 Independent variables

In addition to the indicators prepared in the data evaluation, other indicators were included to test their correlation with crisis data. Each of the six themes of the WDI gives a large set of indicators. Only the most representative indicators were selected, the one with enough and recurrent data, excluding refined indicators, indicators lacking data, or with high similarity with other indicators. Table 4.3-1 gives the list of the fifteen indicators selected grouped by theme.

Poverty and inequality
1. *Poverty headcount ratio at national poverty lines (% of the population)*
2. *GINI index*
3. *Growth rate in per capita, total population (%)*
People
4. *Population, total*
5. *Government expenditure on education, total (% of GDP)*
6. *Unemployment, total (% of total labor force)*
Environment
7. *Access to electricity (% of the population)*
8. *Total natural resources rents (% of GDP)*
Economy
9. *GDP (current US$)*
10. *GNI per capita, PPP (current international $)*
11. *Foreign direct investment, net inflows (% of GDP)*

States and market
12. Military expenditure (% of GDP)
13. *Armed forces personnel, total*
Global links
14. External debt stocks, total (DOD, current US$)
15. *Merchandise trade (% of GDP)*

Table 4.3-1 List of independents variables for the regression analysis

Poverty and Inequality

This theme provides three leading indicators used by several economists and researchers in analysing country data. The poverty indicator represents in percentage, the number of people per country living below the poverty line; The Gini index is a coefficient that measures the economic inequality in a country; its value varies between 0 and 1; The growth rate per capita represents the average income of the population.

People

This theme includes indicators that concern the general population. The first key indicator is the population, which gives the human population per country. The second key indication related to the population is the government expenditure on education, expressed as a percentage of the GDP. This indicator does not include private investment in education but reflects a comparable indicator among countries of public investment in education. The following key indicator is the unemployment rate, which reflects the available labour force without seeking employment.

Environment

This theme has several empty indicators. Just two indicators provide historical data: the percentage of people that can access electricity. It reflects the country's energy poverty status—the natural resources rent is the contribution of the natural resources to the GDP.

Economy

This theme contains the most popular and overly used indicators. The first one is the gross domestic product (GDP), or the value of services and goods produced in a country in an annual period; it reflects the country's progress. Another key indicator is the gross national income (income per capita), the income earned by the population and companies in a country; it includes foreign investment and development aid. The last indicator is the foreign investment indicator, representing international investments in a country as net inflows or the contribution of foreign investors to the national economy.

States and Markets

In this theme, two crucial indicators reflect the fragility of a country and its development; the military expenditure, or the percentage of the GDP that was spent on armed forces in an annual period—the armed forces personnel, which is the number of active military personnel.

Global Links

On this theme, indicators reflect the economic relationship between a state and external actors, like the external debt stocks; the debt owed by a state to non-residents entities like the International Monetary Fund. The merchandise trade is the GDP percentage representing exportation and importation.

Income rate

The income rate categories countries per income, it helps to refine from the complete list of countries, only countries that need development aid. That includes low- and middle-income countries, which often face governance crises and political instability (UN-OHRLLS, 2020).

4.3.2 Dependent variable: Crisis data

The dependent variable is the crisis event list per country generated by the ACLED Project. The ACLED project collects information about fatalities and incidents related to political violence and protests across Africa, South Asia, South East Asia, the Middle East, Europe, and Latin America. As assessed in previous research, the ACLED data are not timely but are frequently updated compared to other similar projects. Considering the constraint of the resources required to capture and update such information, the ACLED project is performing well regarding timeliness. The initial dataset was transformed to create a dataset that provides for each country the number of events that happened per country in a year. The data gathered between 2015 and 2018 is the denser period with the highest number of crises event reported per country. The period where most of the crisis data and the DI are available is between 2015 and 2016.

2016 was selected as a year where the relationship was analysed because it has the most significant data. To verify the validity of the data analysed in 2016, the results were compared with the results from 2015 to verify if the relationship between the DI and crisis event does not change in consecutive years.

4.4 Regression

The dataset created in the previous section was analysed using the R language and environment, free software for statistical computing and graphics (*R: What is R?*, no date). As stated in the methodology section, the objective of the study is to analyse the relationship between WDI (15 independent variables) and crisis events (one dependent variable). The relationship was analysed using the regression method. An overview of the entire dataset shows that all the variables are numeric, with discrete values such as the population or the crisis event

and continuous values like the GDP (Appendix H-1). With 137 observations of the 16 variables, the dataset has some missing values that a univariate analysis identified.

4.4.1 Univariate analysis

The regression analysis of the designated data started with an exploratory analysis of the variables. The first step was to remove all the observations with missing values in the dependent variable to create a sub-dataset of independent variables for further analysis. Univariate analysis of the dataset gives a descriptive statistic of the independent variables. The summary() function gives an overview of all the variables (Appendix H-2). The results of the summary() function show that they are some variables with several empty fields; the significant difference between the mean and the median for several variables indicates a not normal distribution; the dependent variable has some extreme outliers, visible with the gap between the third quartile and the maximum.

Empty values and outliers

All independent variables with more than 50% of empty values were removed. That includes four key indicators, "poverty_ratio", "Gini," "growth_rate", "education", which reduced the number of independent variables to eleven. After removing unusable columns, empty rows were removed by selecting rows from empty the crisis data (69 rows). The crisis data contains several outliers (Figure 4.4-1). These outliers were deleted and concerned countries were facing extreme situations like war. When the value of the third quartile of "crisis_event" is 858, the countries that have outliers are India (13,305 events), Iraq (10,470 events), Yemen (8,761 events) Pakistan (4,469 events) Turkey (4,152 events), Philippines (3,195 events) and Somalia (2,664 events). Two independent variables, the GDP and the

external debt, have several outliers (Figure 4.4-2), but only one extreme value of the GDP was removed.

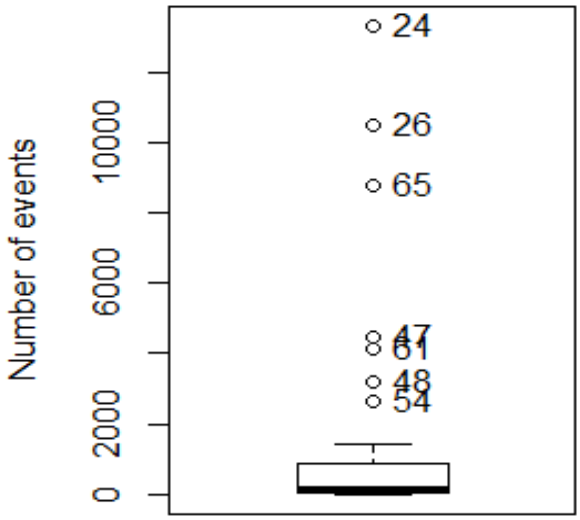


Figure 4.4-1 Boxplot of crisis_event

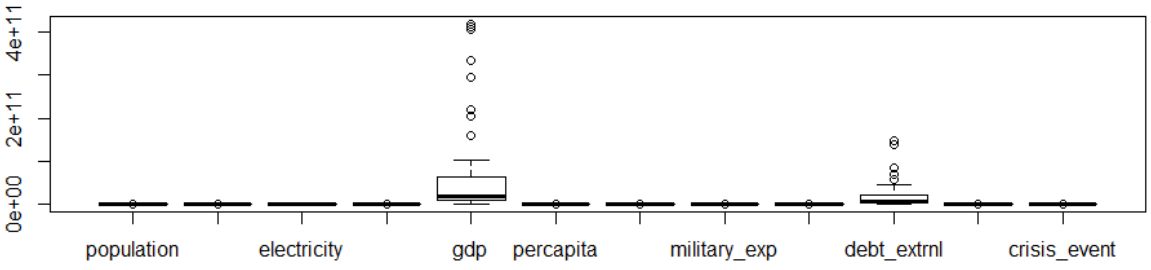


Figure 4.4-2 Boxplot of selected variables

Normal distribution of variables

The distribution of the variables was analysed with the Shapiro-Wilk test; it rejects the null hypothesis about the normal distribution of the variables if the P-Value is larger than 0.05. The results of the test on variables show that none of the variables has a normal distribution,

including the crisis data (Appendix H-3). That will not affect the regression analysis because it is not a requirement to have the dependent variables normally distributed. Instead, the regression modelling needs the dependent variable to be normally distributed, this requires the normalisation of the dependent variable of crisis event count.

The "crisis_event" was first standardised with the scale() function because the country data event was recorded as they occurred, off-scale. The function best normaliser (), from the R package with the same name, selected the best method to normalise the vector (Appendix H-5). This function applied the Ordered Quantile normalising transformation, which gives a uniform distribution of the variables with the Shapiro test, P-value=1.

4.4.2 Multivariate analysis

The Poisson and the linear regression both agree that the independent variables should not have high multi-collinearity, and the independent and dependent variables should have a linear relationship.

Linear relationship

This assumption required the relationship between the independent variables and the dependent variable to be linear. The scatterplot shows that the "crisis_event" variables have a moderate linear relationship with five variables and a low relationship with six variables (Table 4.4-1).

Independent variables	**Correlation Coef.**	**Rate**
population	0.615881	Moderate
unemployment	-0.00106	Low
electricity	0.216587	Low

natural_resources	-0.07534	Low
gdp	0.641132	Moderate
percapita	0.105894	Low
foreign_investment	-0.11182	Low
military_exp	-0.00599	Low
forces_prsnl	0.62466	Moderate
debt_extrnl	0.576629	Moderate
merchandise_trade	-0.40406	Moderate

Table 4.4-1 Correlation coefficients of the independent variables with the crisis event

The correlation coefficient was calculated using the Spearman rank-order correlation, which evaluates the strength of the relationship between the DI and the crisis data. The Spearman method is appropriate as it does not carry any assumptions regarding the distribution of the variables.

Multicollinearity

Independent variables do not influence each other, meaning they should be little or no correlation among the independent variables. Appendix H-4 shows that there are several independent variables with a strong correlation. The strongest collinearity is between external debt and GDP (0.89). This strong collinearity among the dependent variables did not satisfy an assumption of the regression A factor clustering grouped variables with strong collinearity as factors to create a new list of dependent variables with low collinearity.

4.4.3 Factor analysis

To address the issue of DI with a strong correlation, the factor analysis identified the most important variables and reduced the number of independent variables by creating new

factors that cluster the variables with a strong correlation. The factor analysis method used here is exploratory. This method associate variable based on their correlations; the associated variables were reduced to factors representing the new variables.

The Eigenvalue: Kaiser-Meyer-Olkin

The Kaiser-Meyer-Olkin tested how the dependent variables suited for the factor analysis. To suit a factor analysis, the measure of sampling adequacy (MSA) of all variables should be higher than 0.6. The overall KMO value of the dependent data is 0.67, with three variables with an MSA value inferior to 0.62 (Appendix H-6). The eigenvalue calculated the variance of the variables to factorise and gave new factors with the number of variables they associated. The factors with a value higher than one were retained, as they associated a minimum of one variable. That gives four factors for further analysis (Figure 4.4-3).

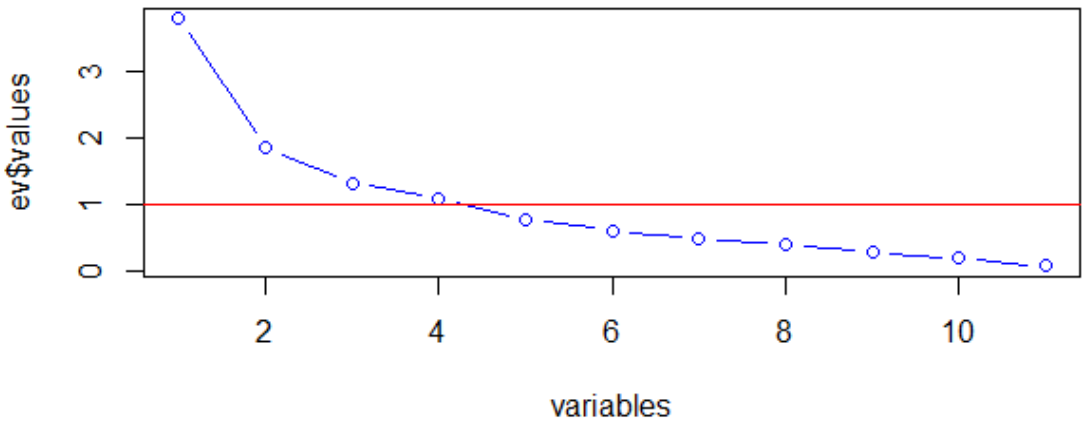


Figure 4.4-3 Eigenvalue

The varimax rotation

The varimax or maximum variance rotated dependent variables into factors to create a factor loading matrix that gives the correlation between the factors and each variable (Appendix

H-7). This matrix associates each dependent variable to one factor, Figure 4.4-4 below give a visual representation of this association.

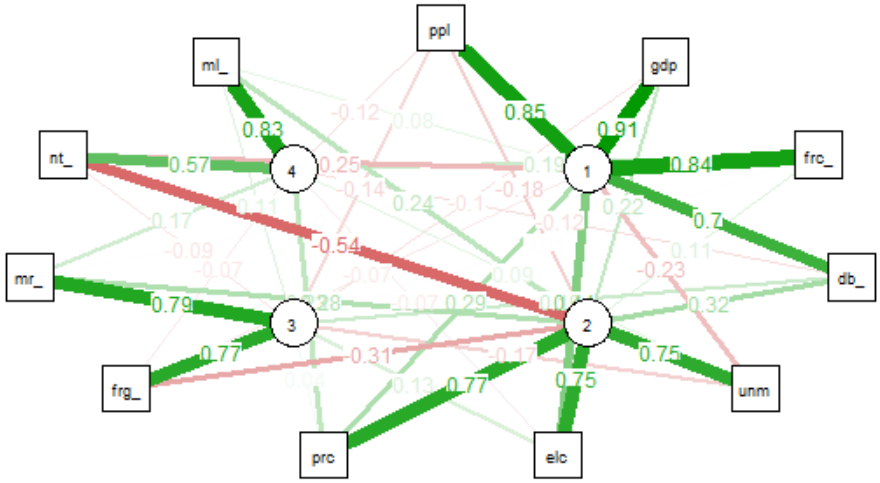


Figure 4.4-4 New factors visualisation

New factor analysis

The new factors statistically represent the associated variables. Table 4.4-2 summarises the factor analysis results and describes the correlation between factors and variables.

New factors	**Variables**	**Correlation coefficient**
RC1	population	0.85
	gdp	0.91
	forces_prsnl	0.84
	debt_extrnl	0.7
RC2	unemployment	0.75
	electricity	0.75
	percapita	0.77

RC3	foreign_investment	0.77
	merchandise_trade	0.79
RC4	natural_resources	0.57
	military_exp	0.83

Table 4.4-2 Factor analysis results' summary

RC1:

All variables associated with the factor RC1 have a statistical and theoretical relationship. The GDP, as the contribution of national producers to the economy, has the strongest correlation with other variables; the correlation with the population (0.79) shows that as the population grows, the GDP also grows. The military force personnel, which quantitatively represents the manpower of a country, is also strongly linked with the GDP (0.82); lastly, the external debt for developing countries also strongly correlates with the GDP (0.88). If the statistically most representative variable is the GDP, the most meaningful indicator is the population. The population is a demographic data that impact naturally in other indicators. The RC1 was renamed "population_featured."

RC2:

The income per capita is the statistically most important indicator for the second factor of this analysis. Both the unemployment indicator and the access to electricity represent a percentage of the population. If the Income per capita is not a percentage of the population, it reflects the income generated by resident producers. The RC2 will be renamed the "population_featured2" as it reflects the population from the economic and environmental perspectives.

RC3:

This factor is associates with "foreign_investment" and "merchandise_trade" indicators. If these two indicators represent both a percentage of the GDP, their correlation with

the GDP is weak. Coming from two different themes, the foreign direct investment and the merchandise trade indicators do not have a relation explained by their definition and are not statistically linear. This factor was renamed "random_factor."

RC4:

Natural resources rent and military expenditure are both associated with the factor RC4. As for the previous factor, the correlation between the two indicators as well as other indicators is low. By definition, there is no similarity between the two indicators; they are independent of each other. This factor was renamed "random_factor2."

The four factors created were analysed before applying the regression analysis. Univariate and multivariate analysis shows with the Shapiro-Wilk test that none of the factors is normally distributed (Appendix H-8); the linear relationship between the factors and the dependant variable is moderate with the first and the third factor and low with the second and the fourth factor (Appendix H-9). Lastly, regarding multicollinearity, there is no collinearity among the factors (Figure 4.4-5); the factor analysis fully fulfilled its role.

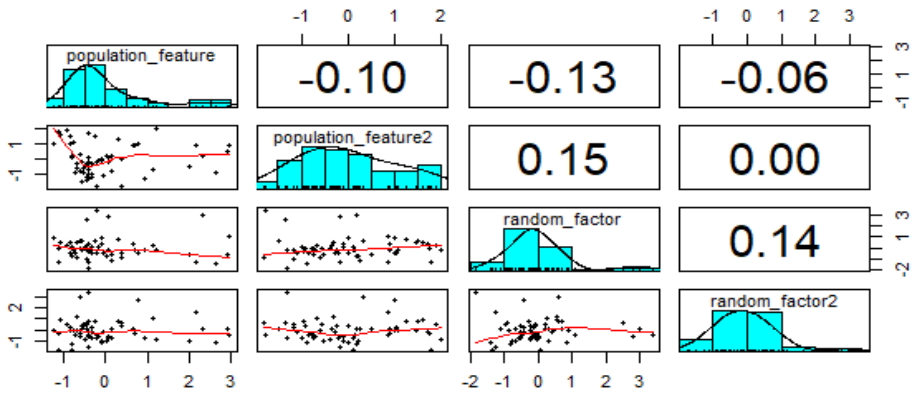


Figure 4.4-5 Factors Scatterplot

The factor analysis gave some interesting insights regarding the relationship between independent variables. Just two of the four factors have a linear relationship with the dependent.

The regression analysis investigated the relationship between the factors and the crisis event and the relationship between independent variables with a linear relationship with the crisis events data.

4.4.4 Regression analysis

Two different datasets experimented with the regression analysis. The univariate and multivariate analysis results followed by the factor analysis demonstrated that some indicators are key when others are less relevant for this regression analysis. The first dataset (dataset 1) contains all the five DI that have a linear relationship with the crisis event variable; the second dataset (dataset 2) resulting from the factor analysis, is composed of the two factors having a linear relationship with the crisis event variable. The results of multilinear regression analysis with a Poisson regression were applied to the two datasets. To find if the results of the two methods give different results. The analysis did not emphasise the results of the residuals nor the performance of the model because the interest is only on the coefficient of the regression as they describe the relationship between the DI and the crisis event.

Regression lines

- There should be two regression lines if there is two variables X & Y given,
    - Regression Line of X on Y
    - Regression Line of Y on X
- The average relationship between the two variables is shown and is known as Line of Best Fit.

Nature of Regression Lines

- The two regression lines become identical if $r = \pm 1$.
- The two regression lines intersect at 90° if $r = 0$.
- The greater the degree of correlation is, when nearer regression lines are to each other.

- The correlation is positive when the regression lines rise upward from left to right.

4.4.4.1 Multilinear regression

The model of the relationship between the response variable and explanatory variables is attempted by Multiple linear regression. The independent variable x is combined with the value of the dependent variable y for every value. A linear equation is fitted to the observed data. The units should observe a random sample from the well-defined population. Uyanık and Güler (2013) analysing the five independent variables are in the standard model or not. To measure the dependent variable on a continuous and interval scale, the interval scales should be measuring the independent variables. All the variable distributions are normal. There should be linear relationships between the independent variable and the dependent variable. Correlating the independent variables, there should be no near-perfect (or perfect) appearance, multicollinearity is a situation called for this. In the Anova sense, having no interactions between the independent variables, for testing b coefficients, a rule of thumb is N >= 104 + m, here m = number of independent variables.

The purpose of this analysis is not to predict future outcomes. The predictive performance of the model is less relevant in this research, more important is the description of the relationship between variables. The results of the multilinear regression analysis on dataset 1, representing the five DI that have a linear relationship to the crisis event variable, show that three variables have a meaningful weight (P-value smaller than 0.05); the population, the external debt, and the merchandise trade (Table 4.4-3).

- If the population increases by one unit, the estimated increase of a crisis are 1.159e-08units
- If the external debt increases by one unit, the estimated value of the crisis increases by 1.266e-11units

- If the merchandise trade increase by one unit, the estimated value of a crisis decrease by -1.206e-02 units

An increase in the population and the external debt increase the risk of a socio-political crisis, when the increase of the exportation and importation in a country decreases the risk.

```
      Coefficients:
                                  Estimate Std. Error t value Pr(>|t|)
      (Intercept)                 7.997e-02  2.412e-01   0.332  0.74166
      indicator2$population       1.159e-08  4.256e-09   2.724  0.00902
**
      indicator2$gdp             -2.181e-12  2.093e-12  -1.042  0.30284
       indicator2$forces_prsnl    7.725e-07  8.700e-07   0.888  0.37907
      indicator2$debt_extrnl      1.266e-11  4.720e-12   2.682  0.01007
```

Table 4.4-3 Coefficient summary of the multilinear regression of dataset 1

Similar to the results above, the summary of the multilinear regression on the factors in dataset 2 shows that:

- The "population_feature" factor that associates the population, the GDP, the military forces personnel, and the external debt increase the crisis risk by 0.3621 unit, when this factor increase by one unit
- On the opposite, the "random_factor" that includes the foreign investment and the merchandise trade in a country decreases the crisis risk by -0.6298 unit when this factor increases by one unit.

When the "population" and its associates' variables increase, the probability of a socio-political crisis increases as well. When the variables related to business activities and investment increase, that reduces the risk for countries with low and middle incomes (Table 4.4-4). The measurement of the importance of each factor shows that the population factor contributes 80.3% in the regression while the business factor contributes about 19.6% (Appendix H-13).

```
Coefficients:
                          Estimate Std. Error t value Pr(>|t|)
 (Intercept)                 5.5485     0.1699  32.657  < 2e-16 ***
fit.data$population_feature  0.3621     0.1091   3.318  0.00158 **
fit.data$random_factor      -0.6298     0.1889  -3.334  0.00151 **
```

Table 4.4-4 Coefficient summary of the multilinear regression of dataset 2

4.4.4.2 Poisson regression

Poisson Distribution is the separate prospect of a total of the events that come about aimlessly in a given period. Nelder, Chatterjee, and Price (1979) employed the square-root transformation for eliminating the mean-variance identity. Moksony and Hegedűs (2014) adopted the dependent variable transformations. Cupal, Deev, and Linnertova (2015) proposed Poisson regression on flood occurrence as the dependent variable. The trials should be very much in the Poisson distribution. Under the observation, the probability of occurrence outcome is negligible. In addition to the requirement, consistency of probability and the independence of series from the row to row property are needed. The following conditions are satisfied by the Poisson random variable:

- There is an independent should be in the number of successes within the two disjoint period intervals.
- During a small interval of time, the entire length of the time interval is proportional to the probability of success.

To model, the count data Poisson regression, which is in the form of regression analysis is used. The Poisson regression applied to (non-transformed) count crisis data identified the same key variables to predict crisis risk as in the multilinear regression, but with different weights for each variable and a higher margin of error. The over-dispersion in the data caused

by the variance of the dependent variable is larger than the mean. Therefore, the quasi-Poisson regression was applied and found that the population, the external debt, and the merchandise trade are confirmed not to have a relationship with crisis due to chance. The estimated value of the crisis is affected by the positive weight of the population (1.153e-08) the external debt (1.710e-11); and the negative weight of the merchandise trade (-2.302e-02) (Appendix H-14).

The generalised linear models are Poisson regression models with the logarithm canonical link function. In a Poisson distribution, assume the response variable Y, The unknown parameters of linear combination modelled the expected value logarithm. When using the model contingency tables, a log-linear model is used, which is only a categorical variable. Multiple regression analysis is permitted by Poisson regression with cohort data having a dichotomous outcome and continuous or categorical variables usually used when the outcome is a rate or rate ratio especially useful for rare diseases in large population models of an exponential function rates magnitude with the linear combination of unknown and covariates parameters are referred to as a "log-linear" model because it is a log transformation of an outcome variable (e.g. a rate) related to a linear equation of predictors.

The quasi-Poisson regression applied to the factors with a linear relationship with the crisis data gives similar results to the multilinear regression. The "population_feature" has a positive weight on the regression (0.3621) and the "random_factor" a negative weight (-0.6298) (Appendix H-15). With the quasi-Poisson regression, the population factor also increases the risk of a socio-political crisis, while the business factor reduces it.

The regression analysis results from the multilinear and the Poisson regression give three DI and two sets of factors from the list of DI that have a relationship with the crisis event, not due to chance but with a significant margin of error. Analysing the performance of the multilinear regression, the R-squared of the first model indicates that the DI explains 48.3% of crisis data. This is less than half of the collected data; the other half of the crisis data are

explained by other factors. The second model generated from the factor analysis has a lower R-square (0.3102), but both have a P-value smaller than 0.01; the multilinear regression models are significant, then the models are not performed by chance. Same for the Poisson regression, the low Fisher scoring iterations show that the models are not performing by chance and can be used to predict future crises.

4.5 Conclusion

Regression can apply to predict the future, correct an error in thinking or provide a new perspective. This section investigated the relationship between development indicators and socio-political crises. From a statistical perspective, it demonstrated that some DI is quantitatively linked to the crisis. It does not contradict the findings of some economists or social scientists but gives slightly different results regarding the importance of each DI. Social instability correlating with the economic indicators like the income per capita, as stated by several researchers, does not correspond with the regression analysis findings, giving a low linear relationship (0.10) between the income per capita and the crisis event for developing countries. Some indicators theoretically associated with the income per capita correlate with crisis and can contribute to predicting it. The merchandise trade indicator alone or associated with foreign investment indicators as factors can contribute to predicting the crisis. When the merchandise trade indicator decreases, the risk of a crisis increases, as initially expected from the income per capita. Analysing the impact of individual economic factors and the economic growth of a country, some researchers found that unemployment and education have an impact on the likelihood of supporting violence. Like the income per capita, the unemployment data do not correlate with the crisis, and the education indicator did not provide enough data for the analysis. Instead, the number of the population as a single factor correlates with the social

instability. It can contribute to predicting it and associated with other indicators (the GDP, the military forces personnel, and the external debt).

If the relationship between the economic and people indicators with the socio-political crisis was demonstrated, the indicators under the poverty theme were excluded because of the low quantity of data available. Therefore, this analysis will not exclude nor include the poverty indicators as relevant for crisis risk prediction. An alternative could be to use the data from the OECD describing the level of income to represent poverty as categorical data. Other indicators like access to electricity, with a not-so-low correlation with crisis (0.21) might give different results for a country or a time-series data analysis.

The regression analysis applied to the 2016 data was also experimented with the 2015 data to find if the results were significantly different. A similar process obtained almost the same results with some differences; for 2015 data, the military forces personnel replaced the external debt as predicting variable (Appendix H-16). It demonstrated that the importance of DI in relation to a crisis can vary from one year to another. Further analysis like a time-series analysis of the relationship between DI and crisis events over several years and per country could generate interesting findings in predicting crises. Keeping in mind that the correlation does not necessarily imply causation, and the relationship between the DI and crisis event might be reversible, it is possible that it is the crisis event that can influence DI, like in a country facing instability causing the merchandise trade to decrease.

CHAPTER FIVE: THE RESEARCH METHODOLOGY

5 The risk assessment framework

5.1 Methodology

This study uses a quantitative research method. This method supports decision-making and planning, but no prediction can be entirely confident; prediction is difficult if there is no existing model, and making a prediction could influence the result (Castle, Clements, and Hendry, 2016). The forecasting method that is applied to this research is the cross-impact matrix method, a multivariate forecasting technique where the probability of an event is affected by other events. Each development indicator aggregates different other variables and could represent a system. The cross-impact matrix method analyses how each indicator's variation can help estimate the likelihood of a social crisis. The cross-impact matrix method can be compared with the simulation methods to measure its performance. It is ideal for models with the relationship between variables or the trend extrapolation, which takes trends in historical data over time to predict the future.

5.2 Research Philosophy

The philosophy that is applied to this research is **postpositivist**. The postpositivist is an improvement of positivism, which is described by Uusitalo (2014) as a scientific philosophy that does not interfere with the phenomena of study, which is repeatedly observed. This research philosophy includes methodology as forecasting to predict likely future events. The postpositivist emphasises creating new knowledge that can contribute to social justice (Ryan, 2006). The researcher is not only an observer of the phenomena but contributes to resolving the problem. This research provides additional knowledge about the correlation between

development indicators and social crisis data and contributes to calculating the risk of changes in indicators that can harm society.

5.3 Method

The postpositivism philosophy is the root that underpins the methodology used for this research; **the deductive methodology** refers to an evidence-based reality (Casula, Rangarajan and Shields, 2021). Opposed to the inductive methodology, which aims to develop a theory, a deductive methodology is an exploratory approach that starts with an existing theory - formulates a hypothesis - collects, analyse and tests the data - to accept or reject the null hypothesis. Often used in medical research, this method tests the risk assessment theory against the new dataset.

The literature review suggests developing a novel theory to assess quantitative risk when all risks are not known in advance and where a collinearity relationship exists between the dependent variables. Considering that the CNN and TOPSIS data model can assess the risk of a deadly crisis based on World data and ACLED data, the deductive methodology will test the Convolutional neural network and a MCDM to assess the risk of deadly crisis on a given comparable dataset that describes countries from a social, political and economic perspective.

To the question "Can a data model be capable of assessing the risk of a violent crisis, using network analysis and graph convolution?", the null hypothesis is "assessing the risk of a violent crisis is not feasible with a data model developed with neural network and graph convolution". This research will reject the null hypothesis by describing a data model that can identify the risk of crisis and statistically identify the indicators that could play a critical role in mitigating the risk of crisis.

5.4 Risk identification

The risk factors are represented in this research by the variation of development indicators. The number of fatalities represents the consequence of crises, and the number of fatalities materialises the level of risk for the population; this variable is clustered to represent the relative level of a crisis.

The crisis level is associated with the risk level represented by the developing indicators changes. The risk of crises can be considerably mitigated by population development and the economy (Clarke and Dercon, 2019). Changes due to climate, pandemic or economic shocks can be represented with the variation of development indicators; about their effects on loss of life, Clarke and Dercon (2019) proposed in the International Development Association 2019 to invest in the preparedness for response to changes during and after a crisis. The risk to be identified in this research is the change in development indicators in relation to a previous crisis which has been mitigated. The level of risk is expressed numerically using quantitative instruments. Additional detail might require a better comprehensive risk analysis of a country than could be accomplished using qualitative techniques. Furthermore, if the risk analysis results turn out to be incorrect, it may be necessary to conduct additional, more in-depth investigations.

Risk identification with imagery analysis is an area where deep learning methods have been used extensively. It might be for an earthquake or landslide, where imagery analysis with convolution contributes to identifying a feature in an image that represents a risk, mainly on satellite images. The experimentation of imagery analysis for risk identification on correlation graphs is novel. The novelty is principally due to the fact that the visual representation of the centrality of various indicators has not been used as the data source for imagery analysis. In this research, the proposed method will create a database of images representing several countries' social-political and economic states according to different crisis levels. The CNN

will learn the patterns that can identify similarities and later inform about the changes that can reduce the number of fatalities due to crisis.

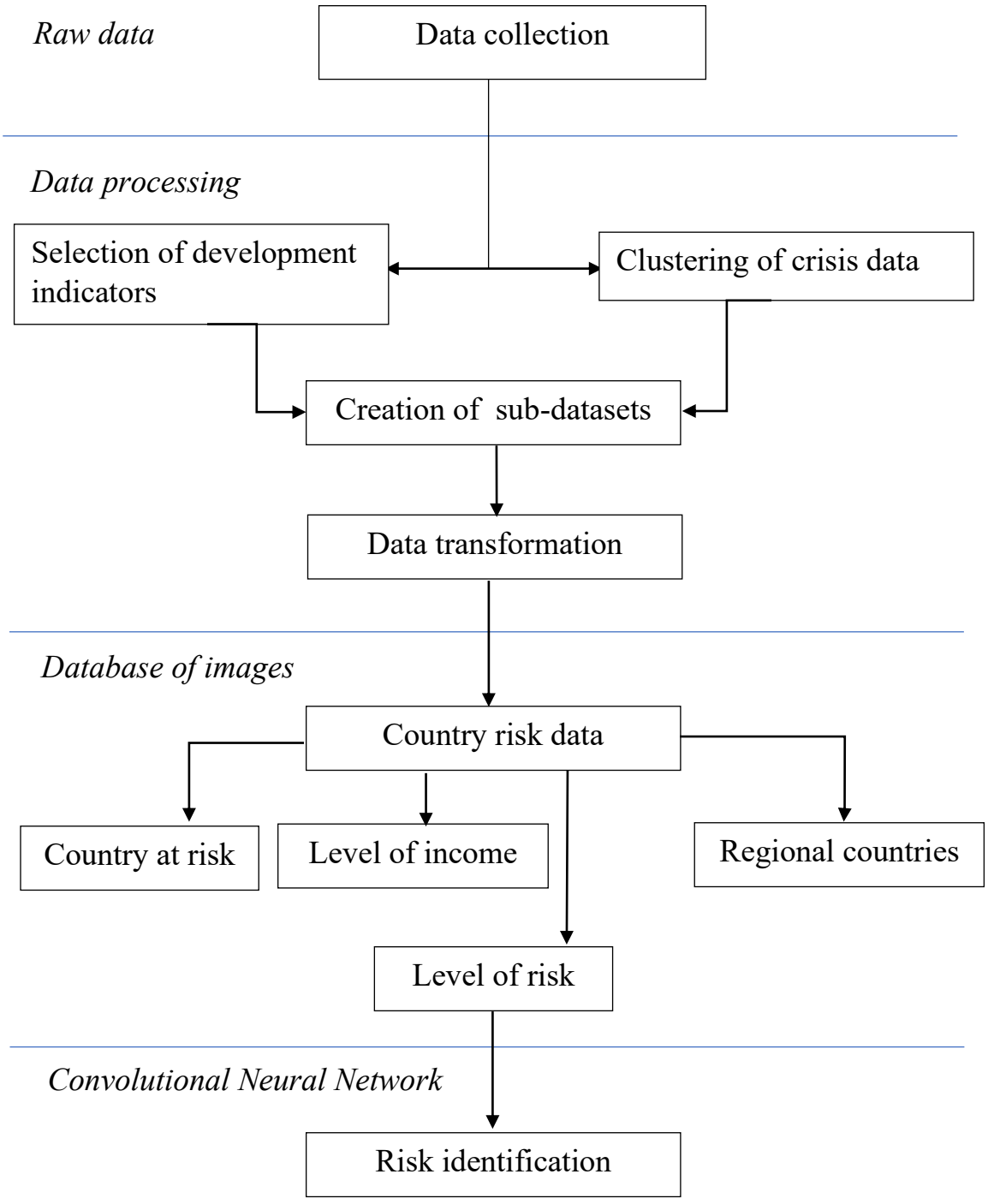


Figure 5.4-1Risk identification using CNN

The risk identification in this study is based on the landslide identification model developed by Wang et al. (2020). In that method, deep learning techniques are used to identify landslides. The method is highly based on five steps. For this research, this method was updated to integrate steps that transform development data into images (Figure 5.4-1).

5.4.1 Raw data

The raw data of this research consists of two sets of data: the development indicators data and crisis event data. Both are historical data available on many countries that allow a comparative analysis annually or by country. The literature review identified a set of predictors that are grouped into political, social, economic, and governance indicators, natural resources, and military expenditure (Table 5.4-1).

Predictors
Economic Indicators
Total factor productivity (TFP) Banking and finance: domestic credit to the private sector Capital stock Energy: electric power consumption GDP per capita, excluding oil rents GDP per capita, oil rents only Growth in GDP per capita, excluding oil rents Growth in GDP per capita, oil rents only Informal economy Per capita GDP Trade Openness
Governance indicators
Corruption

Democracy
Government &civil society: freedom from corruption
Law and order
The proportion of population excluded from power
Semi-democracy
Time since independence
Time since the pre-independence war
Time since the regime change
Political indicators
Civil Liberties
Political Rights
Polity
State Capacity
Voting US
Social indicators
Communication: mobile cellular subscriptions
Education
Education: primary enrolment ratio
Health: maternal mortality ratio
Population size
The proportion of the population between 15 and 24 with at least lower secondary education
The proportion of population living in urban areas
Transport: air carrier departure Transport: air carrier departure
Unemployment
Water and sanitation: access to safe water
Other indicators
Military expenditure
Resource rents

Table 5.4-1 List of predictors identified in the literature review

The outcome variable is the crisis data generated from the Armed Conflict Location and Event Data Project (ACLED). This data source is crucial for this research and was subject to a data assessment. After collecting the raw data, the next phase of the methodology will be to process the data and create a database of images. The images will represent the centrality of the weighted correlation that connects each indicator.

5.4.2 Data processing

The literature review identified a set of data that can serve as indicators to evaluate the crisis risk. Because this method is based on a correlation relationship between indicators, consistent and regular data are required to create an identical set of images. A distinct set of indicators in 2 images will interpret incorrectly because the convolution network cannot make a difference between two indicators but will just assess the differences between the graphs representing the indicators.

Data selection

The first step of data processing is data selection. It will comprise identifying indicators regularly entered for the countries to assess. Historical data covers different periods of low and high or no death crises in a country. The successive years covering the analysis period should contain enough data to allow a correlation analysis. Indicators with empty data will be removed from the list.

Clustering of crisis data

The second step of the data processing assesses the crisis level corresponding to each year. The level of a crisis in a country can be low, medium, or high. The level of crisis linked

to death in a country can be estimated by the number of deaths. To a country unfamiliar with the deadly crisis, one death during a crisis could be assessed as a high crisis; in countries where violence is recurrent, a significant increase of deaths in a year will represent a high level of crisis. It can vary from one country to another because of the frequent level of violence but also because of the demography.

The number of deaths will be transformed into categories that reflect the level of risk. The clustering algorithm of the partitioning around medoids (PAM) is suitable for categorising integer numbers. The PAM clustering algorithm will cluster the set of N number of deaths into K clusters. The K-number of clusters is defined in advance where K=3, with:

[1]=Low crisis

[2]=Medium crisis

[3]=High crisis

The clusters do not reflect the actual level of crisis in a country but a relative level from a data view. Presenting the years where the number of deaths was relatively low compared to other countries—using the example of the number of deaths reported by the ACLED project for Benin, a West African country (Table 5.4-2). The PAM clustering method performed well in classifying the data. Once the crisis variable is transformed into categories, the next step in the method is to group successive years with the same crisis level in a table.

Years	Number of deaths	Level of crisis
2012	14	1
2013	10	1
2014	7	1
2015	8	1
2016	17	1
2017	35	2
2018	32	2
2019	60	3

Table 5.4-2 Number of deaths recorded in Benin because of political crisis

Create sub-dataset

Once the indicators and the crisis level are set, the third step is to split the data into the respective successive period of the crisis level. The split of the records is done by successive grouping of years corresponding to each class of the crisis level. Using the example of the data in Table 5.4-2:

DS1: Dataset 1 represents the data stored from 2012-2016. The list of indicators selected in the previous step will present for these years the state of the country from economic, social, and political perspectives, where the level of the crisis was low.

DS2: Dataset 2 represents data covering the period when the country had a medium level of crisis

And DS3: Dataset 3 represents data covering the year when the country had a high level of crisis.

Transform datasets into images

The correlation network of every dataset gives a representation of how various indicators are connected. This step will consist of transforming the correlation network of each dataset into an image that gives the eigenvector centrality of the dataset. Eigenvector centrality measures the influence of an indicator in the dataset. The image of the centrality will represent the entire dataset, giving how vital are each indicator in the selected period. That will allow, in the risk assessment phase, a comparison of images having the same indicators to assess how the importance of each indicator varies with different levels of crisis.

An example of the transformation with the datasets created above will give images representing the centrality of the correlation of development indicators in Benin when the country encountered low and medium crises. The centrality measure used in this research is the

strength centrality, which measures the centrality of an indicator by calculating the number and weights of each linked indicator. The image generated is a coordinate graph. The value of the centrality measure is the x-axis, and the y-axis gives the list of all indicators varying according to the crisis level.

Figure 5.4-2 shows the centrality plot of Benin in a circumstance of low crisis and in a circumstance of medium crisis. Figure 5.4-3 compares the two graphs and shows significant changes in the importance of some indicators between the two states.

The indicator trade 'trd' in low crisis has several positive correlations with other indicators; when in medium crisis, it decreases to be close to zero. Another indicator that changes significantly is the education 'edc'. Education in a low crisis context has several negative correlations, and in a medium crisis has several positive correlations. Correlation analysis of the prominent indicators will provide more insight about their interaction with other indicators.

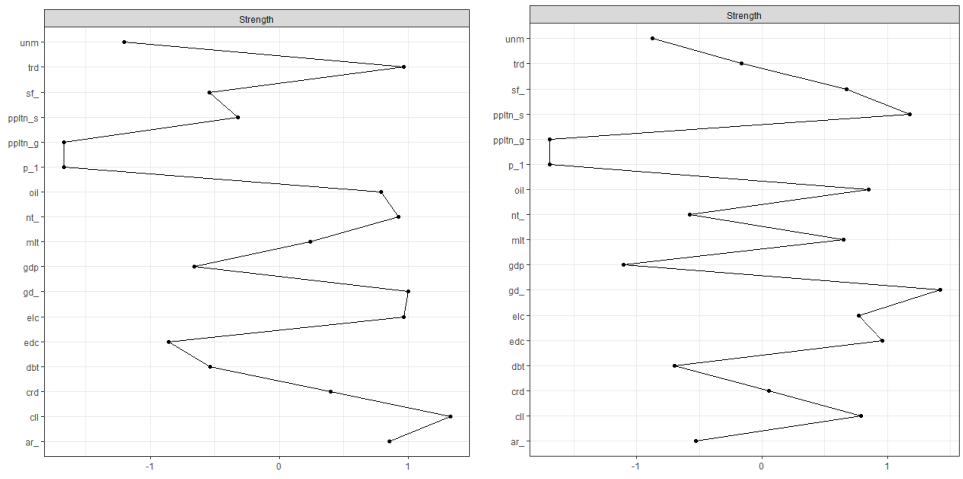


Figure 5.4-2 Benin: Low crisis (left) and medium crisis (right) centrality plot

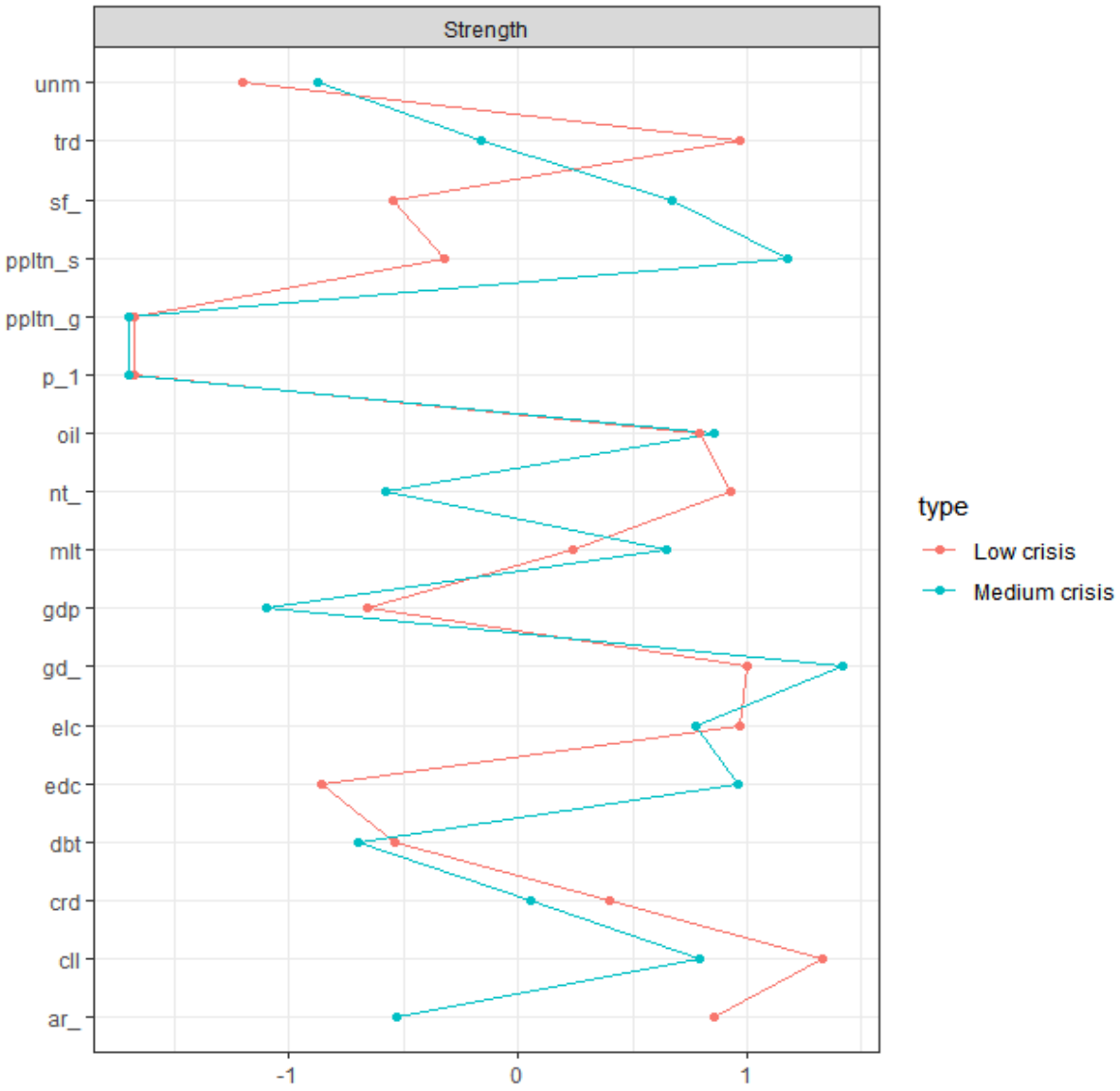


Figure 5.4-3 Benin: comparing of centrality plot of low and medium crisis

The identification of primordial indicators in crisis can be identified with the above steps to guide the assessors into what indicators require particular attention. The CNN will identify key indicators faster and more precisely by comparing centrality plots from a given period to plots from other periods or plots from other countries. It could allow, for instance, a regional analysis of a crisis to identify important indicators to reduce the risk of a deadly crisis. Before automating risk assessment with CNN, the database of images should be established.

5.4.3 Database of images of countries

A database of images will support the assessment of risk. This database will contain images of centrality plots of many countries' low, medium, and high crises. That will allow the CNN to extract features representing indicators varying in the context of crisis. The CNN will compare a new centrality plot with the plots stored in the database. The comparison will be performed from three perspectives; comparing the plot of a country with the plot of the same country but in a different time slot; comparing the plot of the country with the plot of countries from the same region; lastly, comparing a country plot with the plot of countries with the same level of income.

Taking the example of Benin, the database of neighbouring countries will include first the immediate neighbouring countries, then the countries from the local region, and lastly, the country from the global region. The location of a country in relation to other countries will be a new variable for the database of images. The new database will include the list of variables in Table 5.4-3. The new database will also include the level of income as a variable to allow a comparison of countries with the same level of economy.

Level of crisis	Low, Medium, High
Country	Name of the country
Image	Name of the image in the image folder
Country positioning (Neighbour)	Positioning of the country in relation to the country of study (Yes, No)
Region	The region where the country is located
Continent	Continent where the country is located
Level of income	Low, Lower-middle, Upper-middle, High

Table 5.4-3 Variables of the created database

The database was created from the risk identification of this methodology and reduced the initial raw data. The raw data initially contained hundreds of rows describing countries from political, social, or economic indicators. The new data will reduce all the indicators to a visual representation of the importance of each indicator in relation to others with the collinearity. The CNN will read in one image all the development indicators of a country corresponding to the level of crisis, extract the features and compare them with other images.

5.4.4 Risk identification with CNN

Wang et al. (2020) tested different algorithms in their method to find the best. They measured the performance of eight models and found the Deep Convolutional Neural Network (DCNN-11) as the most suitable to assess landslide risk.

From social analysis to biology to computer vision, graphs naturally occur in different application disciplines. This research exploited the graphs' unique capacity to capture structural relationships between economic and socio-political data with crisis data. That enables them to yield more insights than simply studying data independently. Nevertheless, the capacity to learn from graphs can be difficult because meaningful connectivity should exist between data, and the form of data such as text, numbers or categories should allow for building a graph from their relationships. On the other hand, representation learning has achieved remarkable accomplishments in many fields. The graph properties can be retained by learning the representation of graphs in a low-dimensional Euclidean space. Despite all the work done to address the graph representation learning challenge, many solutions still have insufficient learning mechanisms to handle the issue. Deep learning models using graphs (e.g., graph neural networks) have recently evolved in machine learning and other related domains and showed higher performance in many problems (Zhang *et al.*, 2019b). These days, convolutional

modules are being used to train network data representations, inspired by the dramatic improvement in image recognition of CNN.

In this methodology, the CNN model will be trained with the database created to learn features that describe the different crisis levels from all countries and all years. The CNN model with the learnt patterns will be able for an image representing a country with the risk to identify, to find the most similar images in the database. The similar images identified might correspond to a low, medium, or crisis level for a country at a specific year; a manual review of the most similar country that has been in the same situation in the past and reduced the risk of deadly crisis will provide the basis for statistical analysis to identify the most varying indicators. The analysis of countries experiencing similar deadly crises is done through the country's historical data, and it can be from a different perspective. From the perspective of countries experiencing the same type of crisis, countries with the same level of income, and from the perspective of countries from the same the region.

To identify the risk, once the country that experienced a similar deadly crisis in the past and improved is identified, the set of images from that country will give a progression by the year from the lowest level of crisis to the highest level of crisis. That will make up the alternatives set of indicators to select to reduce the level of risk. The CNN will identify, the countries with a matching risk level from an analysis of the indicators. The matching analysis is conducted from the perspective of income and geographic location. Using Benin as an example, for every country matching with the level of risk of Benin, the lowest and highest level of risk for the countries matching will contribute to calculating the best plot reflecting the changes to implement in selected development indicators to reduce the risk of crisis.

The process of deciding on the best alternative is based on the indicators' weighted-TOPSIS method. The best alternative corresponds to the changes in the development indicators that can reduce the level of risk of crisis in a country.

	electricity	cellular	education	oil	gdp_ growth	military	population_s lums	nat_ resources	trade	safe_ water
Low	3.09	3.33	1.88	2.98	3.12	2.61	2.24	3.07	3.09	2.09
Medium	3.29	3.30	3.42	3.35	3.74	3.20	3.57	2.35	2.64	3.22

Table 5.4-4 weight of indicators in a correlation network

Benin's low and medium centrality plot shows that the indicator with the higher weight is the "cellular" for the low level of crisis and "population_slums" for the medium level of crisis. Table 5.44 gives the indicators with the highest weight for a low and medium crisis level in Benin. Some indicators remain essential regardless of the crisis level when some indicators change in relation to the crisis level.

5.5 Indicator weighted-TOPSIS

The TOPSIS method developed by Hwang and Yoon (1981) is an MCDM method that ranks alternatives based on favourable or unfavourable criteria. This method will use an improved TOPSIS method that will rank the set of development indicators in order of choice, based on the importance of indicators. The correlation network between indicators reflects the importance of the indicators in this research. The importance of an indicator or the indicator weight compared to other indicators is the indicator with the higher number of weights from correlated indicators.

The TOPSIS method in this research will follow a few steps to select the set of indicators to target to reduce the risk of a violent crisis in a country. The positive ideal set of indicators (PISI) is the set of indicators that correspond to the lowest level of risk of crisis. In contrast, the negative ideal set of indicators (NISI) is the set of indicators that corresponds to

the highest level of risk of crisis. The ideal solution to reduce the crisis level is the set of indicators close to the PISI and far from the NISI.

The matrix of the decision model is like Table 5.4-4, where the alternatives are the level of crisis, the criteria are the development indicators, and the value of the criteria is the node weight of the correlation network. Further, this research will define the positive criteria or the criteria that can contribute to reducing the risk of crisis; and the negative criteria that can contribute to increasing the risk of crisis.

The decision matrix

The decision matrix with m alternatives and n criteria is a matrix represented as follows:

$$DM = \begin{array}{c} \\ A_1 \\ A_2 \\ \vdots \\ A_m \end{array} \begin{array}{c} \begin{array}{cccc} C_1 & C_2 & \dots & C_n \end{array} \\ \begin{pmatrix} x_{11} & x_{12} & & x_{1n} \\ x_{21} & x_{22} & & x_{2n} \\ & & & \\ x_{m1} & x_{m2} & & x_{mn} \end{pmatrix} \end{array}$$

Where:

$(A_i)_m$ is the list of alternatives. A set of indicators corresponding to the centrality plots resulting from the risk identification process.

$(C_j)_n$ is the list of criteria. It represents the indicators used in the methodology.

And $(x_{ij})_{m \times n}$ is the values matrix representing the indicator weight in the correlation network.

With $i = \{1 \dots m\}$ and $j = \{1 \dots n\}$.

Normalisation of the decision matrix

This step standardises the original decision matrix. Each number in the column of the original matrix is divided by the square root of the sum of the squares of the numbers in the same column of the matrix.

The normalised matrix, $(r_{ij})_{m\times n}$ each value of the matrix is calculated with the following formula $r_{ij} = \frac{x_{ij}}{\sqrt{\sum_{k=1}^{m} x_{kj}^2}}$

Weighting matrix

The new matrix contains normalised values of the initial matrix. The weight associated with each criterion corresponds to the importance of the indicator in the country of study at risk. Each criterion will be multiplied by the weight associated.

The new matrix is $(t_{ij})_{m\times n} = (r_{ij})_{m\times n} \times (w_j)_n$ with the vector $(w_j)_n = (w_1, w_2, \dots, w_n$

PISI and NISI

PISI is the positive ideal alternative (A^+), when NISI is the ideal negative alternative (A^-). The positive ideal alternative is the alternative that corresponds to the lowest level of risk, while the ideal negative alternative is the highest level of risk.

The positive ideal vector is $A^+ = (v_1^+, v_2^+, \dots, v_n^+)$

The ideal negative vector is $A^- = (v_1^-, v_2^-, \dots, v_n^-)$

With n the number of criteria.

Euclidian distance to PISI and NISI

This step calculates the Euclidean distances to the positive and negative ideal for each alternative. The distance of the target alternative to the positive ideal is d_j^+,and the distance to the negative ideal is d_j^-.

$$d_j^+ = \sqrt{\sum_{i=1}^{n} (v_{ij} - v_i^+)^2}$$

$$d_j^- = \sqrt{\sum_{i=1}^{n} (v_{ij} - v_i^-)^2}$$

Proximity to the positive ideal solution

This step measures the relative proximity of each alternative to the positive ideal alternative. It is calculated by dividing the distance to the negative ideal by the sum of the distances from the positive ideal and the negative ideal of the same line.

$$D_j^+ = \frac{d_j^-}{d_j^+ + d_j^-}$$

Ranking of the alternatives

The last step of the TOPSIS method will rank the alternatives. The best alternatives are the ones nearest to the positive ideal alternatives and distant from the negative ideal solution. The alternatives are ranked by order of preference.

Following the risk identification with CNN, the TOPSIS method used a matrix of weighted development indicators to identify the best set of indicators to reduce the risk of crisis for a selected country. This method identifies not only the best ideal set of indicators to reduce the risk of crisis, but the best alternative that considers the importance of the indicators of the country of study.

5.6 Conclusion

This chapter deals with the methodology, research philosophy, and risk assessment framework used for the study. This study evaluates a set of data that can play as indicators to identify the crisis risk. The method is based on a correlation relationship between indicators; regular and consistent data are needed to develop an identical set of images. A database of images will determine the assessment of risk. It constitutes images of a centrality plot of the low, medium, and high crisis of many countries. The risk identification can be evaluated by using CNN. In this research, the TOPSIS method follows some steps to choose the set of indicators to minimise the risk of violent crisis in the country. Followed by the risk identification with CNN, TOPSIS prefers a matrix of weighted development indicators to determine the set of indicators which can minimise the risk of crisis for the chosen country. Among all the methods, the best alternative has been taken into consideration for the significance of indicators of the country of study.

CHAPTER SIX: RESEARCH DATA

6 Data assessment

6.1 Development and crisis data assessment

6.1.1 Introduction

The main focus of attraction is the data quality from the past few years among the various companies and organisations (Silvola et al., 2016). The volume of databases and data stored in the last decade has grown exponentially. This trend will continue without any reduction in intensity or strength due to the 'cloud' data and low-cost ‘torag’ (Hug, Zachariassen and van Liempd, 2011). For expensive business software, more investments are made to access these databases. In technology, solutions mostly have to be sought. It should be ignored entirely when there is an issue of insufficient quality data being obtained.

Alpar and Winkelsträter (2014) promote many aspects of data quality operations. When erroneous data is obtained, it is like a contagious virus, as the incorrect inputs negatively impacts the outputs. To adapt to other databases, it should spread through the organisation, and this adaptation served as reference data, and management information reports formed. The business areas and operations have some impact on poor data quality (Puurunen, Majava and Kess, 2014). While the database size is increasing, inaccurate data volume is increasing alongside.

Loshin (2010) analyses the business processes through small components and the lost values. For poor business decisions, the result obtained likely to be inaccurate, which will guide to ineffective and inefficient operations in the performance; hence the organisations and companies suffer millions of euros’ damage or loss for their productions. Batini et al. (2009) analysed the data quality with the state, measurement, and improvement. Many companies are seeking to enhance their returns and save costs, and even there, the economy is

stagnating. Therefore, enhancing quality data is vital. In the coming years, the control and management of data quality will become the major themes in the field of Information Technology. There are some business rules like:

- Attributes values restricted
- Relational integrity rules
- Historical data rules
- Time value rules restriction
- Rules of dependence attributes

The data assessment methodology used in this research to assess the raw data is a data-based evaluation: The Process Integrity analysis, with the Cross-Industry Standard Process for Data Mining known as the CRISP-DM methodology, and is a hierarchical process model, which analyses a dataset from four levels of abstraction (Chapman *et al.*, 2000). Designing the evaluation methodology involves two levels, defining the phases of the assessment and, from each phase, listing generic tasks that have to take place. The second part, which includes the other two levels, refers to the implementation of the model with specialised tasks and process instances. To assess the data of this research, three levels of abstraction will be used and adapted in order to respond to the specificity of this research project. It starts by clearly distinguishing the different phases of the assessment from each phase, identifying assessment tasks that have to take place, and lastly, an evaluation of tasks. Finally, the results of the evaluation will present the outcomes of the assessment.

6.1.2 The CRISP-DM methodology

The evaluation of the raw data will assess if deep learning models could apply to the data required for this research. Based on the CRISP-DM model, the methodology to assess the data of this research project will go through three (3) phases. The project understands the data understanding and the model assessment. The model assessment includes the model quality and bias to assess the level of accuracy and reliability of the information, and the preparation model will identify the sample data to which the machine learning model could apply.

1. Project understanding: The starting point of this data assessment describes the objectives. This part does not aim to review the background of the project data but will describe the objectives of the raw data and the data requirements related to the deep learning modelling.

2. Data understanding: The primary outcome of the data assessment is to generate raw data that can be used for the risk assessment. Understanding the data and becoming familiar with them is the objective of the second part of this methodology. This step represents the core part of this evaluation. It has four tasks:

- The data quality analysis
- The data ethic analysis
- The data bias
- The data preparation to construct the data that be used with deep learning#

3. **Model assessment:** The third phase is the model assessment. It is adapted from the UK National Audit Office, Frameworks to review Models and will proceed with a machine learning model assessment to determine which machine learning model could apply to the data. The model assessment will also include:

   - **Model quality and bias**: To assess the quality of the data and to find whether there is any bias caused by;

- *Human error*: Did the data come from a system prone to human error?
- *Background technology*: What technology facilitated the collection of the data?
- *Data collectors*: Was participation of the data subjects voluntary?
- *Overall context*: Does the context of the collection match the context of use?

➢ **And a preparation model**: To determine the sample data that can be used for machine learning, considering the data's quantity and quality.

6.1.3 Machine Learning Techniques

In all the fields, Machine learning finds its application. Developing machine learning-based is quite a challenging diagnostic tool for developing. Moreover, the proper decision-making chosen is essential for achieving an accuracy of better diagnostic. Also, it is crucial to choose the convenient feature—Muhamedyev (2015) analysed the data for big data processing using machine learning techniques. Kotsiantis, Zaharakis, and Pintelas (2006) provide an overview of Machine learning, and a labelled data function is inferred. Marsland (2014) discusses how machines learn themselves to solve a specific problem. The significant artificial intelligence is the Machine Learning application which makes the system for learning automatically from its observed data. Tzanis, Katakis, and Vlahavas (2006) present the modern applications of machine learning techniques. The machine learning aim is to learn the system by themselves with the development of computer programs.

The emerging techniques of machine learning are most preferably the classification tasks. The artificial intelligence subfield is Machine learning, and the machine learning subfield is the deep neural network that performs complex calculations on input like an image and can make predictions on learnt data. These algorithms work by building a model through the example input to make decisions and data-driven prediction programs that follow the

instructions strictly. The main objective is learning the model of the pattern to make a decision on the basis of the data observations.

The steps involved in supervised learning are:

- Based on the input and the output data, a classifier is built by a human.
- A training data set is trained by the classifier
- A test data set is tested by the classifier
- Satisfied when the output is deployed

6.1.4 Deep Learning Techniques

In all fields, Deep learning also finds its application. Similar to Machine Learning, Developing deep learning-based is also a quite challenging diagnostic tool. Furthermore, to achieve better diagnostic accuracy, the proper decision-making is chosen. Also, it is essential to choose a suitable layer to extract the features and classify the outputs to achieve better diagnosis accuracy. Benuwa et al. (2016) review the deep learning technique with state-of-art techniques. The machine learning subset is Deep Learning. For regression and classification, neural networks are used by Deep learning to learn the features directly from the data. It provides pre-trained models, algorithms, and apps to train, create, simulate and visualise, cluster, regression, perform classification, time-series forecasting, control, system modelling, and dynamic and dimensionality reduction.

The patterns like feature learning, regression and image classification, auto-encoding, Convolutional Neural Networks, and the topology network of Directed Acyclic graphs are included in Deep learning networks. The toolbox provided by the deep learning networks of long short-term memory (LSTM) for regression, time-series classification, the activation functions, and the immediate layers is visualised by modifying the network architecture and, therefore, by monitoring the training progress. The deep learning network of pre-trained

models performs transfer learning for small training data sets (the pre-trained models used are VGG-19, VGG-16, AlexNet, ResNet-101, ResNet-50, Inception-v3, GoogLeNet). The layers of the Deep Neural Network find the correct mathematical operation for turning input data into output data. For training the model, DNN considers the various parameters, namely, the initial weights, the learning rate, the number of units per layer, and the number of layers.

6.1.5 Comparison of Machine Learning and Deep Learning Techniques

Artificial Intelligence are neural network, and it consists of several subsets. The artificial intelligence field makes machines do tasks. By the experience and acquired skills, the machines can learn without human involvement. The subset of artificial intelligence is machine learning, and the subset of machine learning is deep learning. The human brain inspires the algorithms used for these techniques. When the data is diverse, interconnected, and unstructured deep learning allows machines to solve complex tasks. Artificial Intelligence is a computer program that acts like a human brain, i.e., it replicates the human brain's thoughts and functions.

Deep learning is learned through artificial intelligence. Therefore, Artificial Intelligence is a computer system having interconnected neurons that process the information acquired; it responds to external inputs. The machine learning evolution is deep learning. Using the machines proposes a programmable neural network for making accurate decisions. This will be performed without the help of human brains. Deep learning analyses the data with a logical structure. Some layered algorithms are proposed in DL and are called artificial neural networks. This ANN design was inspired by the human brain's BNN (i.e., Biological Neural Network). Having more capacities for learning the process than machine learning. Getting the learning process correctly requires lots of training data, to be able to make intelligent decisions

on its own. Many different levels of interpretation algorithms are present in deep learning to convey the data.

Deep Learning does not need structured image data for classification. Hence, the input image is passed to the different levels of layers in the classification process for better performance. Deep Learning can learn the input data, whether it is unlabelled or unstructured, and can be successfully applied to big data for discovery and knowledge-based predictions. Convolutional Neural Network is the most popular Deep Learning algorithm for classification tasks. In Machine Learning, structured data is required, while deep learning needs not. It is designed to learn by the labelled data to produce new results. If incorrect results occur, there is a need to teach them correctly. In Deep Learning, multilevel layers are used for training, and it learns the own mistakes that occurred and rectifies it. The performance in Deep Learning depends upon the quality of the data given to the input. Machine Learning is not suited for complex tasks. There is a need for the development of computer programs to learn more data in machine learning. Deep learning requires GPU to train correctly a large amount of data. It does not learn on an incremental basis. It can be tuned in various ways. Machine learning requires less data to train. It also requires GPU to train appropriately and learn on an incremental basis.

Deep Learning can create new features, but the machine earning needs accurately identified features by human intervention. Deep Learning needs high-performance hardware, but machine learning needs not. Artificial Intelligence makes interconnected neurons, each and every neuron are classified by its activation functions, weights, and bias. In the input layer, the input is fed, and linear transformation is performed by the neurons on the input by bias and weights. The input values have specific functions. Each neuron has output values. The input values are classified by a bias vector and weights through the functions applied. CNN applies

mainly on image classification and image recognition. Face recognition, and object detection, are some kinds of areas in which CNNs are used widely.

6.1.6 Development data

The development data has been extensively described in the literature review. This section assessed the data that should be used in this research. The World Bank is a data supplier of development data. It provides detailed country data, informing about the population, the economy, the environment, and health situation. World Bank data is used as a primary source for several aggregated indexes, like the Corruption perception Index or the Fragile States Index. The World Bank compiled comparable cross-country data on development called the World Development Indicators (WDI) (The World Bank, no date). The WDI is a compilation of comparable statistics that contains 1,600-time series indicators for 217 economies and above 40 country groups, with many indicators going back more than 50 years. The WDI has been widely used to interpret and analyse development worldwide.

The definition of a country in this research is based on the definition used by the World Bank. It refers to a country as a territory where authorities report social or economic statistics (The World Bank, 2022). The World Bank collects data from countries that are World Bank members. In the WDI, the data are not only from countries but also aggregated countries by geographical location, level of income, or different membership. All the aggregated countries were not relevant for this research to only investigate self-administered sovereign states. Some countries with partial autonomy because of their dependency on "Parent State", or some microstates were also not removed from the analysis due to the absence of data.

The literature review identified a set of development indicators (Tab. 4), which will be used as predictors. This section will assess the predictor's data to extract the data that will

constitute the raw data of this research. The WDI, as the primary source of data, will provide data for 19 indicators from the 37 indicators identified in the literature review.

Governance and political indicators are not available in the WDI. The WDI informed mainly about the economic and social situation of a country. Historical governance and political data will come from another data source, the Worldwide Governance Indicators (WGI). The WGI is a source of data provided by the World Bank that covers 213 countries, with data collected between 1996 and 2019 (Kaufmann, Kraay and Mastruzzi, 2011). The WGI is an aggregated data source from various other sources produced by companies, governmental and non-governmental organisations and public institutions. The WGI dataset does not only include governance indicators but also political indicators.

The final set of data to evaluate is a list of development indicators that merge the WDI with the WGI. Taking one indicator for each of the six dimensions will give 25 development indicators to assess in this section (Appendix C). The WDI and the WGI will be assessed separately but following the same methodology.

6.1.6.1 The World Development Indicators

Development indicators are the main data source of this research and provide one-third of the predictors. The nineteen indicators in the WDI list correspond to the social and economic indicators in the literature review. The WDI provides data from 217 countries and 47 aggregated data collected between 1960 and 2020. Some indicators are not generated by the World Bank but by other international institutions and national institutions. The primary sources of data from the WDI used in the research are listed in Appendix D.

The OECD and the World Bank are the leading providers of Economic data, especially related to the GDP; when the UN agencies provide human-related data, like the population total or the school enrolment. Other indicators are provided by specialised institutions:

- The IMF for monetary data
- The ITU for telecommunication data
- The ICAO for air transport
- The ILO for labour data
- The WHO for health data
- The IEA for energy data

Lastly, the Stockholm International Peace Research Institute provides information on military expenditure. These international institutions collect non-human data from national agencies or companies, while most of the human data are collected via survey. The data quality assessment will identify the quantity of the data that can be used for this research.

6.1.6.1.1 Data quality

The Statistical Office of the European Communities (Eurostat) used four components to measure the quality of economic data. The WDI data are updated annually, and the data collected covers the period between 1960 and 2020. Every 217 countries have 19 indicators, making 4,123 rows for 60 columns. Many fields are empty; the completeness review will identify the years that do not provide enough data for this research.

Completeness

The completeness of the WDI dataset is calculated by getting the percentage of missing records. One hundred ten thousand two hundred four records are empty, which represents 44,54% of missing records. The completeness of the data will be analysed from three perspectives. The first perspective is the historical data analysis to identify years that do not contain enough data; the second is country data analysis to identify countries with several missing data; and the third is the indicator data analysis for those with empty values.

Historic data analysis

The WDI selected for this research does not cover complete data over the years. In Figure 6.1.6-1 below, between 1960 and 1990, 40% to almost 80% of the indicators are empty. The successive period that has low data missing starts from the year 2000.

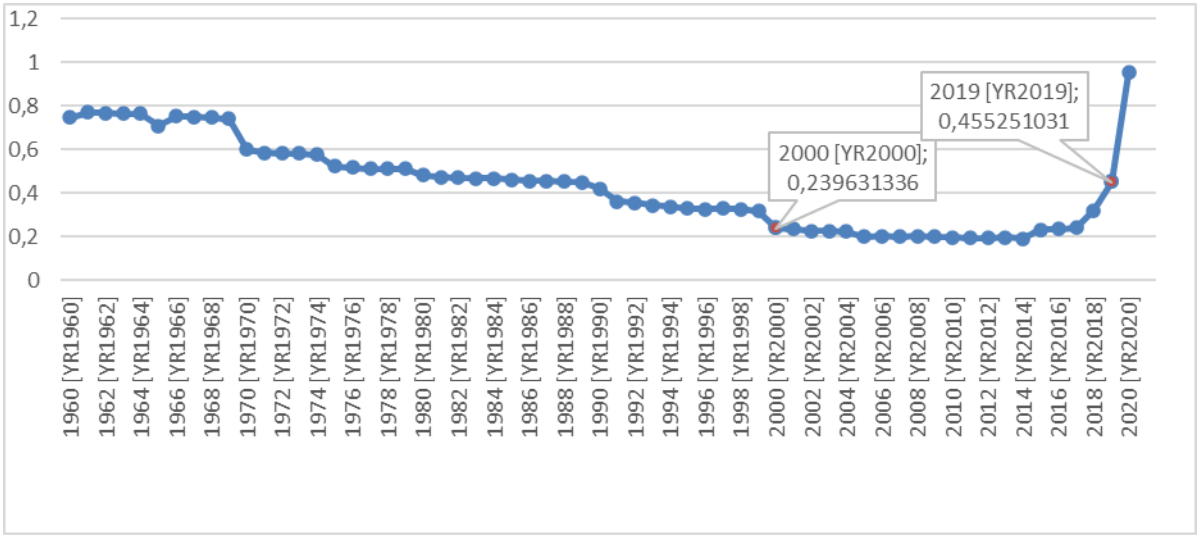


Figure 6.1.6-1 Historic WDI data

Country data analysis

Deleting records entered before the year 2000 and after 2019 significantly dropped the percentage of empty fields (Appendix E). Most countries with more than 50% of missing data are countries under high tension like Somalia, South Sudan, or ISLANDs (Table 6.1.6-1).

Country Name	**Empty fields 1960-2020**	**Empty fields 2000-2019**
St. Martin (French part)	95%	95%
British Virgin Islands	82%	80%

Sint Maarten (Dutch part)	89%	79%
French Polynesia	68%	77%
Channel Islands	84%	76%
Gibraltar	78%	74%
New Caledonia	66%	72%
Somalia	61%	69%
Isle of Man	79%	69%
Curacao	79%	64%
Turks and Caicos Islands	78%	64%
Andorra	70%	62%
Nauru	78%	62%
Liechtenstein	71%	61%
Virgin Islands (U.S.)	77%	61%
Cayman Islands	76%	60%
Korea, Dem. People's Rep.	72%	60%
Faroe Islands	76%	59%
Monaco	68%	59%
South Sudan	79%	58%
San Marino	74%	58%
Northern Mariana Islands	77%	58%
Tuvalu	72%	56%
Palau	79%	56%
Bermuda	61%	54%
American Samoa	74%	53%

Marshall Islands	67%	51%

Table 6.1.6-1 WDI Data: Countries with more than 50% empty data

It is interesting to find that several islands have many missing data. The majority belong to rich countries, like New Caledonia, a French territory, or the Turks and Caicos Islands, a British territory. Since this research is only about some indicators, and the ones related to the socio-politic and economic crisis, excluding countries encountering an exceptional level of crisis like war or countries barely experiencing crisis, it will not affect the results of this research. Nevertheless, the fact that many countries with many missing data are islands requires further investigation, which unfortunately is out of the scope of this research.

Indicator data analysis

Removing the years with missing data and countries that have more than 50% missing data allows seeing that two indicators have more than half of their data empty (Figure 6.1.6-2). The indicator related to corruption, and the indicator related to access to clean water. These two indicators will be removed from the dataset, but also the indicator giving the national estimate of the total unemployment. The national unemployment estimate will be removed to keep only the unemployment indicator calculated with the International Labor Organisation (ILO) model.

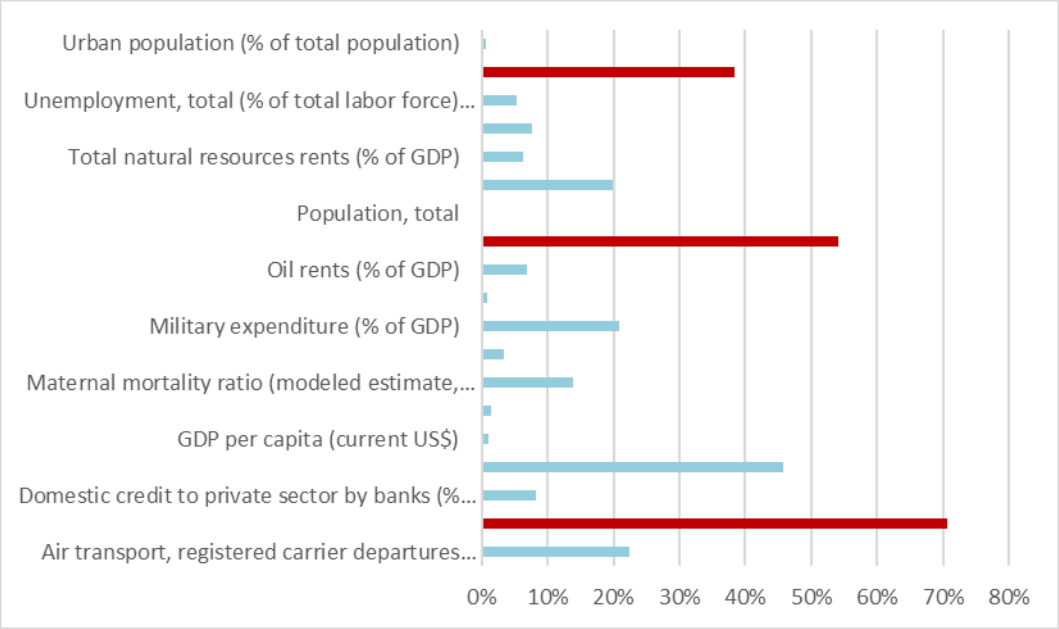


Figure 6.1.6-2 Indicator Data Analysis

The WGI include an indicator related to corruption, which might replace the corruption-related indicator of the WDI. The indicator representing people using safely managed drinking water has been replaced by another indicator that provides similar information; The People using at least basic drinking water services (% of the population). These latest indicators have 0% of empty collected between the years 2000 and 2017, then it will provide sufficient data for the analysis.

Accuracy

Indisputably, the WDI represents the most extensive compilation of global development data and the most accurate. The accuracy of each indicator is variable, as they have different data sources and different ways of collection. But each indicator was also provided by reliable international institutions, and the limitation of each indicator is openly shared by the World Bank.

The quality of the WDI data reflects the quality of primary data collected by the different institutions. The World Bank acknowledged that nationally relevant data might not be

suitable for international standard use due to different methodologies or lack of documentation. Almost every singular indicator of the WDI was generated by a unique source of data, some data sources were national, so following their country standards, whereas other data are generated by large institutions like the IMF or the UN that follow their standards to collect and aggregate data. Regardless of the different methodologies used, the quality might also be compromised by the different surveys and incomplete data coverage. Some data are estimated like the total population; others are generated from a statistical model, like the maternity mortality ratio, generated by a regression model; and others are provided directly by the households or the company, like the credit to private institutions, which financial institutions generate. Another factor that significantly affects the quality of the WDI is the delay in reporting data. The synchronicity between the data source and the WDI is very lengthy, as the data needs to go through the entire process validation by the institution primarily in charge of the data. Up to January 2021, 95,4% of the indicators collected in 2020 are missing data. With only less than 10% of data recorded in 2020, it is impossible to exclude that COVID-19 slowed the data collection process in many countries. The current downloaded might be different from the dataset exported in two more recent years due to the delay in updating the WDI.

6.1.6.1.2 Data ethic

All the indicators in the WDI are under the Creative Commons Attribution 4.0 (CC-BY 4.0). It is the default license of World Bank data, authorising all users to access, modify, share and use data for any purpose, including commercial use. This is what makes the WDI unique because it offers a reliable source of data that is openly accessible to any user. With its Open License, the WDI is anonymised to protect the identity of the users.

The WDI does not allow a user to identify the source of the data, the identity of individuals or businesses behind the data is not traceable. That is the main ethical concern of

the General Data Protection Regulation (GDPR) implemented by the Council of the European Union in 2018. With respect to data privacy, the World Bank is offering complete transparency of its data.

The World Bank for each indicator gives the data source, the method used to generate the data, the relevancy of the data, and the limitation. This allows a transparent assessment of the WDI, but does not prevent the data from a bias that could occur when the data is collected.

6.1.6.1.3 Data bias

National data are often provided by public institutions, which might not be willing to share some data like military expenditure. Military expenditure is the single indicator that is collected by a research institute, Stockholm International Peace Research Institute (SIPRI). The SIPRI estimates the military expenditure from data openly shared by countries, they will not assess the accuracy of the data. It is the same for the Air transport data created by the International Civil Aviation Organization (ICAO). The Air transport data are submitted by countries and might not include all carriers.

Bias in the data significantly affects the quality. Because of the large number of country data to assess, with the various sources of data, this research will consider that the bias in the WDI will not affect the result of this research. This research will aggregate all indicators in one graph, and the assessment method will only look at the variation of indicators. The value of an indicator is less of a problem, more important is its variation.

6.1.6.2 The World Governance Indicators

Another set of indicators used by the World Bank is the World Governance Indicators (WGI). The WGI covers over 200 countries with governance data in six areas. One of the areas includes the data on corruption. That will replace the data-poor indicator data on the corruption

of the WDI. Governance is the way that a country is managed at the highest level and the systems for doing so (Cambridge Academic Content Dictionary, 2020). This definition suggests that governance indicators are higher-level indicators compared to development indicators; and might be unique for each country, as no country is managed precisely the same as another. Evaluating the WGI data will not come without the challenge of comparing the cross-country data entered over time.

In a paper describing their methodology, Kaufmann, Kraay and Mastruzzi (2011) highlighted the subjectivity and the perception-based measures of governance data. They collected data from institutions specialised in governance, business firms, and public data provider. The data are aggregated to represent six dimensions of governance.

The indicators provided by the Worldwide Governance Indicators (WGI) are:

- Control of Corruption- For gaining the public power Control of Corruption is used
- Rule of Law- The agent follows and trust the society rules
- Regulatory Quality-To implement and develop the regulations and policies of the government
- Government Effectiveness- Public service quality represented
- Absence of Violence and Political Stability- government likes are to be destabilised in terms of mean violent
- Voice and Accountability- In government selection citizens have participated.

The six dimensions described above are each represented by one indicator. The authors of the data acknowledged an inter-relationship between the six indicators that fully represent a governance map of a country. With the limited number of indicators, they will be all integrated into the analysis, which will contribute to getting accurate information from the analysis of the

WGI. The various sources of data, with inconstancies in how data are collected, negatively affect the data quality.

6.1.6.2.1 Data quality

The WGI dataset contains data from 214 countries. When selecting the percentile rank indicator of each of the six themes, that makes six records for each country, giving a total of 1,284 records saved in the database. The database covers the years between 1996 and 2019, with some data missing.

Completeness

A quick overlook of the dataset shows that data from the years 1997, 1999 and 2001 are missing. For the years available, very little data seems missing. The year with the highest number of records missing is the starting year 1996, which is missing 11% of the data available for that year (Figure 6.1.6-3). The number of missing data is small enough not to reduce the dataset by year.

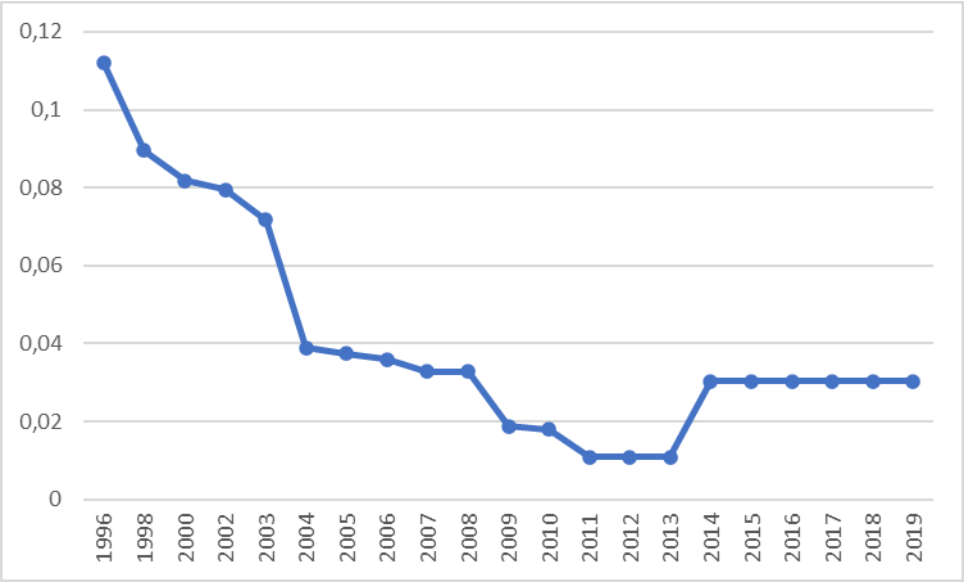


Figure 6.1.6-3 Historic WGI data

Reviewing the WGI data from the perspective of countries gives a little number of countries with more than 50% of data missing. Interestingly, like for the WDI, the countries with a higher number of data missing are countries in a war like South Sudan or Somalia, but also islands (Table 6.1.6-2). Removing the seven countries with many missing values will leave the dataset with data from 207 countries. With less than 5% of records missing in data organised by indicators, the WGI is a reasonably complete dataset with very little data missing.

Country Name	**Empty fields**
Niue	79%
South Sudan	60%
Monaco	58%
San Marino	58%
Jersey, Channel Islands	57%
Cook Islands	56%
Netherlands Antilles	52%

Table 6.1.6-2 WGI data: List of countries with more than 50% empty records

Accuracy

The WGI data are aggregated from different sources, and their accuracy is reflected by the accuracy of the sources. The accuracy of this dataset can be estimated by the standard error calculation, available for each indicator. The standard errors are calculated with the number of sources available for each indicator and the extent to which the sources agree with each other. The more the number of sources is higher and agree with each other, the higher the accuracy of the indicator is. The accuracy cannot be estimated here by comparing the data with the real object, but by comparing indicators. That will give the relative accuracy of an indicator compared to other indicators.

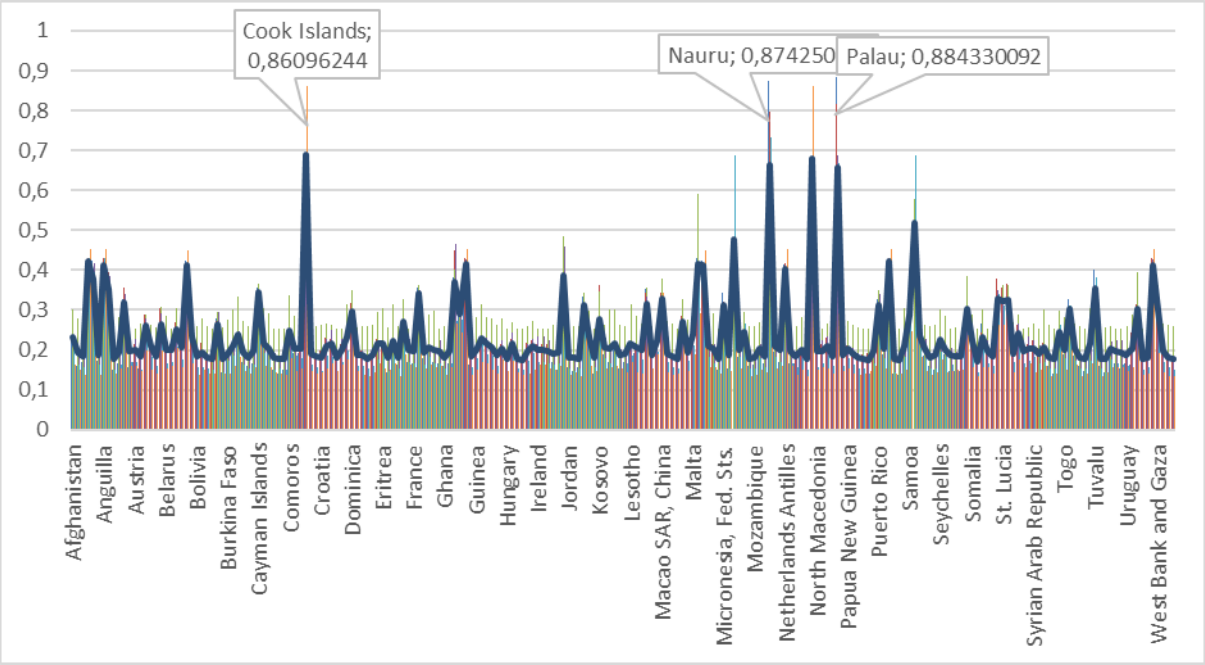


Figure 6.1.6-4 WGI Data: Standard error of countries by indicators

Figure 6.1.6-4 shows that Palau is the country with the highest average standard error of countries by indicator. This average was calculated for a country by dividing the sum of standard errors each year by the number of non-empty indicators. The standard error can highly variate from one country to another. On total average, the indicator with the highest standard error is 'political stability and Absence of violence', while the most accurate indicator or the indicator with highest number or sources and agreeing with each other is 'voice and accountability' (Table 6.1.6-3).

Indicators	**Average standard error**
Control of Corruption: Standard Error	0,217004
Government Effectiveness: Standard Error	0,243589
Political Stability and Absence of Violence/Terrorism: Standard Error	0,294199
Regulatory Quality: Standard Error	0,24103
Rule of Law: Standard Error	0,211894

Voice and Accountability: Standard Error	0,186111

Table 6.1.6-3 Total average standard error of WGI indicators

About the remaining quality of the WGI data, the WGI data are calculated from the data generated by different sources. In the most recent years, the WGI data was able to provide data for the majority of the 214 countries available and their indicators. The data are generated every year, at the date when this analysis was done (January 2021), the 2020 data was not available.

6.1.6.2.2 Data ethic

The WGI data is available for the public to download and use. The license of use is not specified by the authors, but presumably, this dataset has the same license as the WDI data, CC-BY 4.0 license.

All the sources of the WGI data are available, with the contribution of every variable to the governance indicator. The authors are transparent about the data collected and how they were aggregated. The level of abstraction of the information related to governance might be subject to different interpretations that can affect a country's reputation. Taking the example of the 'control of corruption' indicator, the country that has the lowest control of corruption in 2019 is South Sudan (0), followed by Equatorial Guinea (0,48). Using that analysis to rank South Sudan and Equatorial Guinea are the most corrupted countries in 2019 can affect the economy of these countries, especially a country like South Sudan already encountering the challenges of war. This analytical issue has been reported by the authors, who recommended to more focus on the analysis of value for each country instead of a cross-country. The authors also recommend being cautious with time-series analysis as the number of sources can vary from one year to another. The level of abstraction of the governance data can contribute to a data bias.

6.1.6.2.3 Data bias

Most of the data sources used to create the WGI are fully available to the public, except the World Bank's Country Policy and Institutional Assessment (CPIA) which is not fully public. The data are collected from surveys and expert assessments, their authors identified them as subjective or perceptions-based. If perception could shape reality, this perception can be influenced by the environment, media and social media.

The impact of disinformation on governance data is important, which is further described in the Chapter on social media. The following dataset to assess principally used data from media to create a dataset on the crisis. The bias created by information from media will be extensively discussed there.

6.1.7 Crisis Event Data: The ACLED project

In 2018, 24,707 incidents of civilian fatalities were reported in 65 countries by The Armed Conflict Location & Event Data Project (ACLED). The ACLED project is an initiative that started in 2014, to collect data on political violence and protests across the world and conduct analysis to describe, explore, and test conflict scenarios (Raleigh et al., 2010). The ACLED project currently has data from 1997 from six (6) regions. This represents a large amount of information to exploit for future projects related to social or political crises. Applying machine learning models to such data requires assessing it first. This thesis will present an evaluation of the data collected by the ACLED project, as well as the processes and resources used to collect the data. The methodology to evaluate this project will combine techniques for external evaluation with data analysis methods. The objectives of the evaluation are to assess whether it is possible to apply machine learning predictive models to the data collected by the project and identify the limitations.

The project started formally in 2014, and the dataset has been revised several times since 2005 when the idea was initiated. The project is currently receiving financial support from the Bureau of Conflict and Stabilisation Operations (CSO) at the U.S. Department of State, the Dutch Ministry of Foreign Affairs, the Tableau Foundation, the International Organization for Migration (IOM) and the University of Texas at Austin.

The ACLED project collects information about fatalities and incidents related to political violence and protests across Africa, South Asia, South East Asia, the Middle East, Europe and Latin America. The actors, types of violence, locations, fatalities and dates are collected and made publicly available via API and exportable in standard .csv. ACLED data have been used by academicians, researchers and journalists for a real-time follow-up of ongoing conflict dynamics. It has also been used by decision-makers working in the humanitarian and development sector. The ACLED project claimed to have the highest quality, a real-time dataset on political violence and protests around the world. The data remains subject to some limitations identified by the project owner. Such as before conducting an analysis, the user should be aware that the data is gathered based on publicly available secondary reports. Particularly fatality data which are vulnerable to bias and inaccurate reporting.

If the ACLED project is principally targeting political conflicts, various conflict-related datasets exist. Eck (2005), presented a non-exhaustive list of conflict data projects. This list includes a variety of projects like the Arbeitsgemeinschaft Kriegsursachenforschung (AKUF) offering a database of 218 wars and violent conflicts since 1945; the Computer Aided System for Analysis of Conflicts (CASCON), offering a database of conflicts from the post-World War II to May 2000; the Center for International Development and Conflict Management (CIDCM) providing data about conflict and conflict resolution.

Among those various projects, the ACLED project is claiming to offer real-time collection, verification and publication of data. While most of the projects update their

databases on an annual or bi-annual basis. The timeliness of information is a key indicator in assessing the quality of the ACLED data.

6.1.7.1 Data quality

The ACLED data is a record of data on political incidents that happen in several countries. It aggregates information on incidents that have happened in 93 countries on a daily basis, involving state actors. The source of information could be national, regional or international media, which reports incidents related to a political crisis. The information related to political crises is saved by time, location and providing details about the incidents. ACLED data record information on a daily basis, incidents reported in media news or by credible institutions. If the project recorded 22 years of information, some countries would miss annual information. The project started in 1997 with African countries, updated in 2010 with Middle Eastern and Asian countries, followed by European countries in 2018. Up to 15th April 2019, the dataset contained 501,485 rows with 31 columns. That represents information collected from 93 countries.

Timeliness

The ACLED data are captured from a primary source of information and released following the project agenda. In the 2019 ACLED data release schedule, the next release of information is for the spring. Data is planned to be released on a real-time basis for 15 Eastern European Countries and 48 Latin American and Caribbean countries.

Starting with an analysis of the data entered over time, the visualisation (Figure 6.1.7-1) of daily data entered shows a decrease of data entered in the latest days. This most likely shows that not all incidents have been entered; they might be under verification.

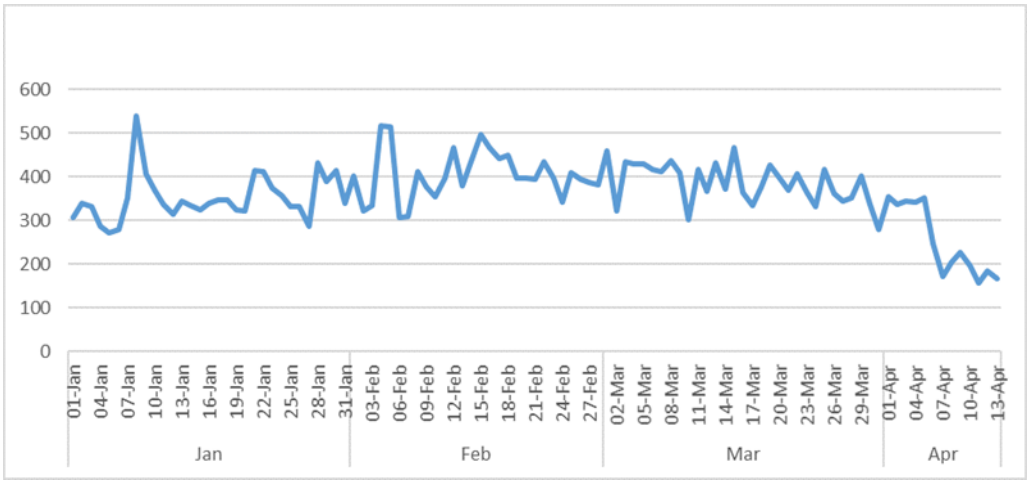


Figure 6.1.7-1 Number of incidents entered per day

The calculation of the standard deviation of monthly data shows that in April the number of the information entered varies significantly (Figure 6.1.7-2). This confirmed that the process of entering the data is ongoing and data are entered by block.

Figure 6.1.7-2 Monthly Standard deviation ACLED data

The annual standard deviation analysis shows several fluctuations in the number of data entered (Figure 6.1.7-3). First, 2010 corresponds to the years when the project added new countries. Next is 2016, the project did not specify any significant change in the data collected, but the comparison of 2015 and 2016 data reveals that in 2016 the project started to collect information from the Middle East (12 countries) and data from the Philippines and India also started in 2016. The project did not provide an explanation for the reason why some country data records started only in 2016.

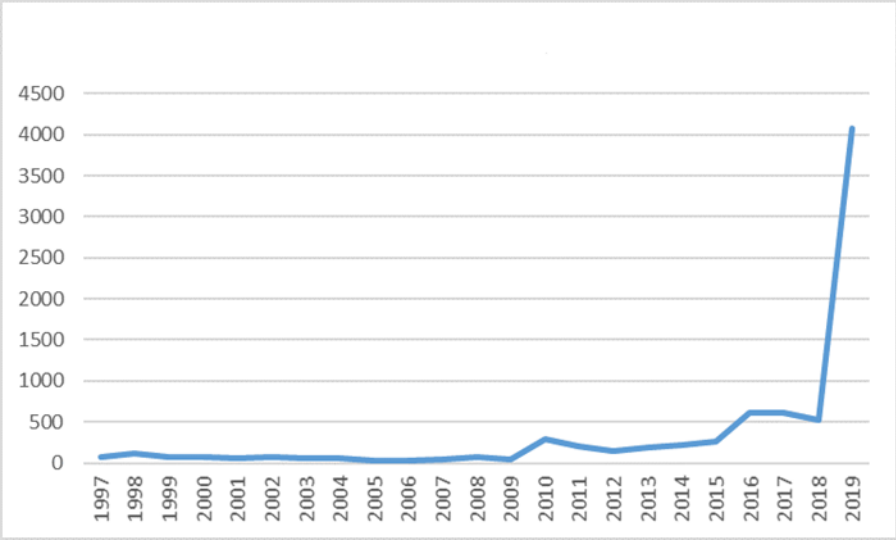


Figure 6.1.7-3 Annual Standard deviation ACLED data

The ACLED data saved up to 15th April 2019 are not timely. The analysis of February and March information shows that data are more accurate after a month. In addition, this figure shows that the only period we have the maximum and most timely information is between 2016 and 2018.

Completeness

The ACLED data to capture disorder that occurs within states requires updated information that accurately represents reality. The project team planned a review of the entire

data set twice a year and review extensively some information. The review of the initial source or findings from a different media source could conduct an update of data recorded. The variable "Timestamp" is a code for data entered simultaneously. It does not say if data have been updated, it aggregates information entered at the same time regardless of the event date. Presumably, the timestamp that shows records for different years includes updated records, at least the records with an earlier date. That corresponds to approximately 30, 789 records with a different (earlier) date of the incident, which is small considering the large number of incidents recorded daily. In the figure below (Figure 6.1.7-4), it corresponds to points that are scattered not in a line.

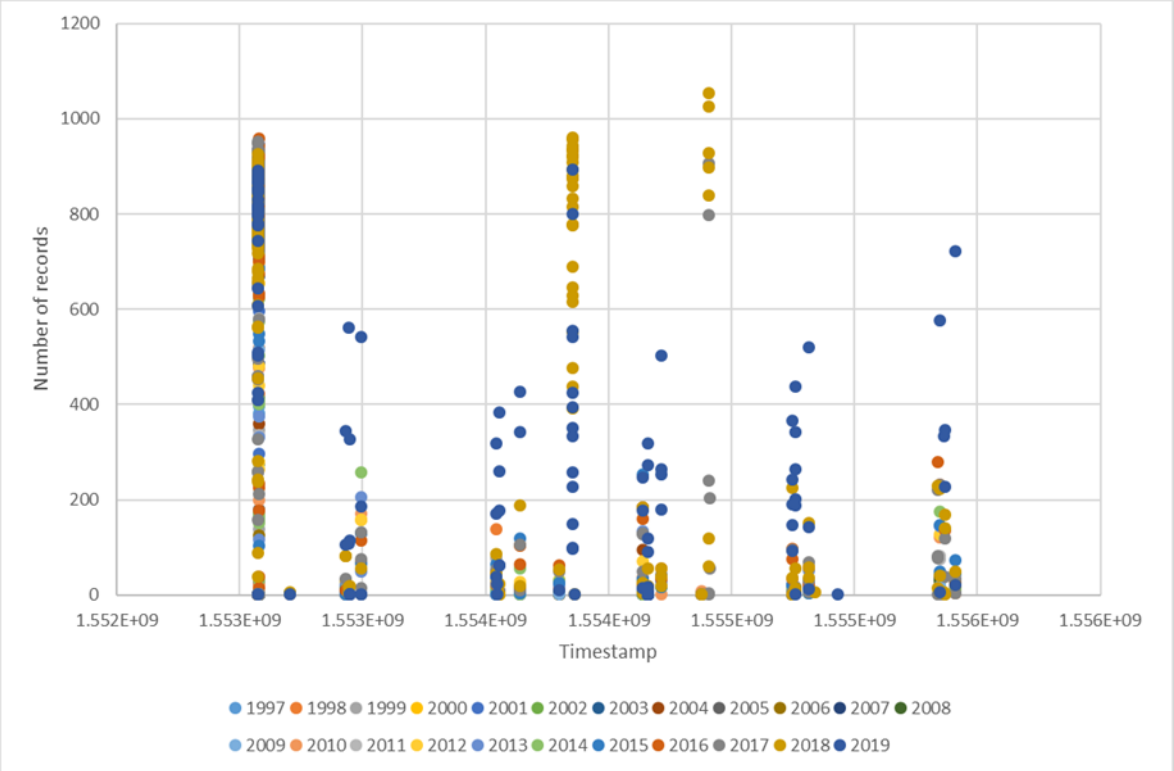


Figure 6.1.7-4 Scatter plot timestamp by year

The timestamp seems to follow an ascending order, from" 1552576388" the initial timestamp (in 2010 when the project started) until "1555409436" (Data downloaded April 15, 2019). The analysis of the timestamp order shows that more than two million timestamps are missing (precisely 2,832,169 timestamps). This period corresponds to a period where records

were entered or updated for different years (Appendix G). The data seem to be reviewed or updated intensively at specific points of time, but not every year.

Reliability

The data are generated from a media source and seem to be recorded in the dataset manually or semi-manually; The project does indicate if the records are coded by a machine, by a human or both, but they are certainly scraped from news websites and reports. The data might be reliable if these are collected consistently, but the verification and validation process described in the methodology presents several layers of cross-checking by researchers (human). Therefore, the data might be subject to errors caused by various interpretations.

Integrity

With the issue of reliability, the integrity of the information might also be compromised by the political or financial interest of the source that reported the incidents. Media favouring the government seems to minimise incidents and use a soft vocabulary, while media against a government might do the opposite. A source of information might be trusted but remains exposed to bias or manipulation for political or personal reasons.

Accuracy

The ACLED project does not provide a link to verify the source of information. Without the link of sources used to generate the records, it is difficult to verify the accuracy of information. ACLED source materials were evaluated as trusted sources, but an estimation of the level of accuracy is not possible. A random selection of records to verify the accuracy reveals that some information, even those concerning a large number of fatalities, is not available on Google Search. One reason might be the language of the initial information, if not

in English; another reason could be that the information is not available anymore on the website, they might be deleted or lost.

Confidentiality

The data are available to the public with no restrictions. No specific security measure or confidentiality agreement seems to be in place to protect the data. With respect to the privacy of information, the data set does not provide personal information. The different layers of verification also exist to prevent the disclosure of confidential information. If a data policy does not exist in many countries, the project incorporates some ethical considerations in collecting the data to protect the anonymity of individual victims or perpetrators.

6.1.7.2 Data ethic

The terms of use and attribution policy of the ACLED data present several limitations to the use of the data (ACLED Project, 2019). It stipulates that the use of data to harm, target, oppress or defame a group or population is not permitted. The purpose of this restriction is to protect the population from any wrongdoing resulting from the data analysis. This policy also restricts the use of the data to develop a similar platform or in competition with the ACLED project. This latest restriction contradicts one key principle of open data, related to the re-use and distribution of the data "the data must be provided under terms that permit re-use and redistribution including the intermixing with other datasets."(*What is Open Data?*, no date). It also does not align with the principle of publication and distribution of the Open Science approach, described by Gema Bueno de la Fuente (no date), which includes open licence and open evaluation to apply no or minimal restrictions to the whole research cycle. The ACLED project does not fully contribute to the diffusion of knowledge, especially regarding the development of a solution that could address social crises using Machine Learning. The main

target of this policy seems to be competitors when some research warns about the use of ACLED data.

6.1.7.3 Data bias

The data quality analysis highlighted quite a few biases in the ACLED data. The human intervention in the verification and validation process for increasing the accuracy might also cause some errors due to the wrong interpretation of the information. An error could occur in the transcription of the primary source of information due to the language difference or the different understanding of the words that can have other meanings in different contexts. The error might also come from the source of information itself.

The data are collected by aggregating local and international news sources. News information is not generated automatically but written by journalists. An author of an article might be neutral but might also be influenced by political or social (internal or external) factors. The review of media sources shows, for example, that the media "Daily Star", used as a source for 5,078 records, required further investigation when obtaining information from this source (MBFC News, no date). The ACLED project assessed their records as trusted and coming from a credible source when the verification of hundreds of information published and cross-checking of various sources is not systematic, and the update of information does not happen very often.

6.1.8 Assessment results

6.1.8.1 Quality of development and governance data

The WDI and the WGI constitute the two components of the development data used for this research. The data evaluation of these data reveals that their quality is their Achilles' heel. The WDI report (2000) stated that the WB is not primary data collection for most areas other

than the living standard survey and debt. That also applies to the WGI, which is not the primary data collection for all the six indicators it provides. Data from secondary providers are not always made available on time. That conduct to get incomplete data, mainly missing for the most recent years. The integrity of the data might be affected when data moves from the primary data host to the WB databank. An Application Programming Interface (API) might exist to transfer data automatically, but that does not exclude a breach of integrity caused by a human or system error.

A human error could also happen when data are collected. For the World Bank, many countries do not have enough resources to employ skilled staff and equip them with the material needed. That does not suffice to explain why the data quality from islands has a higher number of missing data. Taking the example of the Channel Islands, with 84% missing WDI and 57% missing WGI, this country has a high-income level. Several other islands with many missing data have high-income levels. What could be the reason for the data gap for islands? When there are clearly not experiencing conflict or have enough capacity to conduct statistical analysis. This question deserves an investigation to understand the reason for the data gap for islands with high-income levels. Despite the questions arising from the evaluation of the World Bank data, the WDI and the WGI are reliable datasets that can be used for this research.

6.1.8.2 Accuracy and reliability of information

The ACLED is essentially based on data collected by the media. The most critical part of the assessment of this dataset is its accuracy and reliability of the data. Previous assessment of the ACLED data exists, Eck in 2012, compared the ACLED dataset with the Uppsala Conflict Data Program Geo-referenced Events Dataset (UCDP GED) based at the Uppsala University in Sweden. Eck (2012) recommended being wary of the ACLED's data because of the bias resulting from the uneven quality control. When she sees the value of the ACLED data

in providing data on non-violent events, the two data initiatives have to be used carefully due to their dependence on media sources.

The review of ACLED's documents does not provide detailed information about the data collection and verification process. The constraints of collecting the data, country by country, are not described and the web link referring to the media source providing information about the incidents is not available. From an external perspective, the data cannot be analysed as a whole block, it needs to be refined.

The analysis of the timestamp variable shows that the data is fully uploaded over a month. In addition, the ACLED data does not differentiate between the date of the incidents and the date when the data was entered or updated. It is difficult to determine if the information provided is complete. However, from the analysis of the timestamp variable, some records were entered in a different year when the incidents occurred. That might correspond to an update of the record. The data are also subject to bias, the coding of information from primary sources is manual, so subject to human error.

The data rely on a primary source that does not offer deep detail. In addition, translating information from another language to English might also cause the loss of certain information.

The term of use and attribution policy of the ACLED project restrict the use of their data, especially regarding the development of a similar project. That includes solutions that could address social crises using Machine Learning. However, alternatives exist, providing a fully open dataset on conflicts, with a human-machine collection process but not always timely. The Event Data on Armed Conflict and Security project (EDACS) used human and machine-assisted coding to generate conflict event data (Chojnacki *et al.*, 2012). The EDACS project is similar to the supervised-learning approach used in the Social, Political and Economic Event Database (SPEED) (Nardulli, Althaus and Hayes, 2015).

6.1.9 Data preparation

6.1.9.1 Development data preparation

The WDI and the WGI data are time-series data. Data are inconstantly recorded with missing values for some countries. The countries or indicators with missing values will be removed to prepare the data for deep learning. The data quality assessment of the WDI and the WGI prepared the data for the risk identification process in the methodology of this research. The new datasets of WDI and WGI will be merged into one dataset and will exclude countries and indicators with more than 50% of empty data.

The matrix of development data is a table with years as columns and indicators in a row. The same group of indicators is repeated for each country. The WDI and the WGI datasets were downloaded from the same source and then have the same format, making their merge easy. The WDI data started in 1960, while the WGI started in 1996.

The first step in the data preparation is aligning the two datasets by removing the years that are not covered by the two datasets. All the WDI between 1960 and 1995 were removed. But 1997, 1999, and 2001 are also unavailable in the WGI. The new dataset of World Development and Governance Indicators (WDGI), includes some countries that have data only for governance indicators while other countries have data only for the development indicators. Data from these countries (Table 6.1.9-1) will also be removed to keep only data from countries that have both development and governance indicators. The countries with 6 indicators are countries that have only the governance indicators, and the countries with 19 indicators are the countries that have data only on development indicators. Removing the 21 countries from the WDGI gives a list of 205 country data.

Countries	Number of indicators
Anguilla	6
British Virgin Islands	19
Channel Islands	19
Cook Islands	6
Curacao	19
Faroe Islands	19
French Guiana	6
French Polynesia	19
Gibraltar	19
Isle of Man	19
Jersey, Channel Islands	6
Martinique	6
Netherlands Antilles	6
New Caledonia	19
Niue	6
Northern Mariana Islands	19
Reunion	6
Sint Maarten (Dutch part)	19
St. Martin (French part)	19
Taiwan, China	6
Turks and Caicos Islands	19

Table 6.1.9-1 List of countries to exclude from the WDGI dataset

Except for the year 2020, almost empty, all the years have less than 50% missing data. That is due to the late update of the data as mentioned in the evaluation of the data quality of the WDI and the WGI. The next step of the preparation will consist of removing countries and indicators with more than 50% empty data. The total average of the percentage of missing values for each country shows that a small number of countries have to be removed (Table 6.1.9-2).

Country Name	**Empty values**
Somalia	65%
Monaco	59%
San Marino	59%
South Sudan	58%
Nauru	53%
Virgin Islands (U.S.)	52%
Palau	51%

Table 6.1.9-2 Total Average values for each country

After the removal of countries with significant missing data, the next step is to remove two development indicators that have many missing data: the ‘CPIA transparency’ (73%) and the ‘People using safely managed drinking water services’ (59%). And the replicated indicator ‘Unemployment (National estimate)’ (41%) to keep ‘Unemployment (modelled ILO estimate)’ (9%).

The merged dataset has 22 indicators for 198 countries, of data collected between 1996 and 2019. This dataset represents the raw data that will be used in this research to draw the graph that will be used to assess the crisis.

6.1.9.2 ACLED data Preparation

The ACLED data contains a hundred thousand rows. Getting a sample data set with high-quality data needs preparation. The preparation will consist of selecting among the whole database, the period capturing the maximum data and the source of the most reliable information.

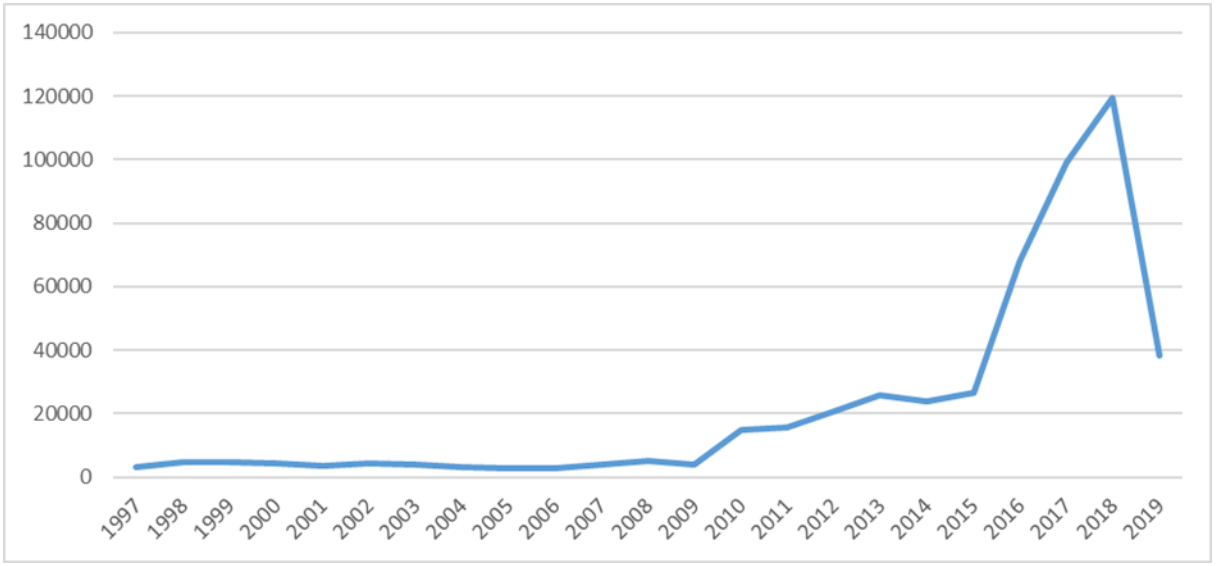


Figure 6.1.9-1 ACLED data entered per year

The most active period from Figure 6.1.9-1 is between 2015 and 2018, and it is the denser period (Figure 6.1.9-2) where data were collected from the highest number of countries, with the few countries presenting many records.

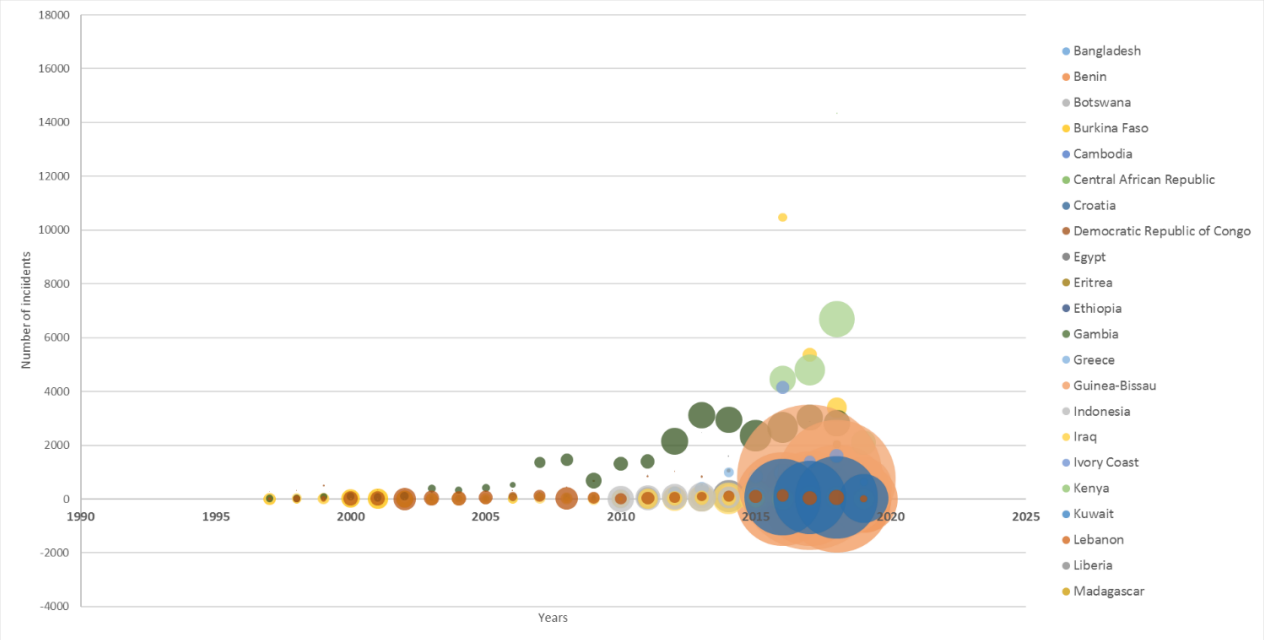


Figure 6.1.9-2 Bubble plot of data entered per year per country

312,685 incidents were recorded between 2015 and 2018 in 93 countries. It is more than half of the whole dataset recorded in four years. Yearly data are missing for some countries. By finding the year's series where most of the countries are represented and removing all countries where data are missing between 2015 and 2018. Between 2016 and 2018, data is available for almost all countries. That makes 200,375 records for 72 countries. This latest selection provides continuing yearly data ideal for time series analysis, with no empty date.

The following selection will filter the data by the source, to find sources coming from international and regional media. From their broader perspective, they are less exposed to bias compared to national or subnational sources of information. It represents 10% of the data, 20,576 records from 72 countries that have certain reliability and can be verified.

To have a homogenous dataset where a machine learning model could apply, the preparation of the data generated a sample piece of the dataset from a period where information is available for most countries from verifiable information sources.

6.1.10 Conclusion

The ACLED database covered almost every form of violence from several countries. The number of incidents is recorded per day, later grouped by years to allow a merging with the WDI and WGI data. The data quality assessment of the WDI, WGI and ACLED data gives various data available for each dataset. The variety of data availability over the years imposes a limitation on the data range to consider for the analysis. The WDI has its denser data availability between 2000 and 2019 (Figure 6.1.10-1); when the WGI covers a more extensive timeline (between 1996 and 2019) with less than 5% of data missing (Figure 6.1.10-2). The final dataset's range of data for analysis is dictated by the ACLED data covering the shortest range of dense data. The timeline where the ACLED data are entered for the maximum country is between 2015 and 2018 (Figure 6.1.10-3).

The WDI and the WGI present a significant number of World indicators. The most representative indicators are those with sufficient data per country and year. In addition to the data generated by the World Bank, a significant amount also comes from other institutions. The historical data collected are universally comparable with data from other countries. The global coverage of the WDI and WGI, offers a macro scale perspective for risk assessment that adds to the model's novelty.

6.2 The data ecology of social network data

In the fields of social science, network analysis, and graph theory, Social Network Analysis (SNA) originated. Social Network Analysis is also called 'structural analysis' (Wellman and Berkowitz, 1988). Otte and Rousseau (2002) documented the growth of Social Network Analysis and drew a co-author network. There should be a network structure that concerns the network analysis with the solution and formulation of the problem: the structure

is captured via a graph. For the graphical analysis, a set of concepts, methods, and abstracts are provided by the graph theory. A combination of analytical tool methods is used for analysing and visualising the social networks. SNA is not a methodology. It is society's function of a unique perspective. The relationship between the groups, social institutions and individuals is centred instead of focusing on macroscopic social structures, attributes and individuals.

To study individuals connected in network relations, the network perspective of society is studied, and the social behaviour is explained within the structure of networks rather than the alone individual. Social Network Analysis has a brief history in social science, and the advanced methods come from computer scientists, mathematicians, biologists, and physicists. In social science, the networks of relationships are essential. Having the data availability, methodology, and computing advances make Social Network Analysis much easier to solve problems.

Beyond social science, Social Network Analysis has many applications; hence, there is a significant advance for humans to generate the study of structures. Some applications are:

- To study the internet traffic, webpages, and information dissemination, network analysis methods are used by computer scientists.
- By the use of network analysis in life sciences, food chains are studied in different ecosystems.
- To analyse the networks, physicists and mathematicians focussed on producing complex and new methods which are relevant to the network.
- To improve and analyse the flow of communication in the organisation SNA used businesses, customers, or the network of partners.
- To identify terrorist and criminal networks, SNA used Law enforcement agencies and the army by tracing the communication which they collect, and key players are identified in the networks.

- To recommend friends based on friends-of-friends which is potential, and to identify SNA used the primary social network site of Facebook.
- To uncover the hidden connections and conflicts of interest between businesses, lobbies, and government bodies, SNA used Civil society organisations.
- To optimise the capacity and structure of the networks, SNA used Network operators like mobile, cable, and telephony.

6.2.1 Social network data Social Media and Political Marketing

With the recent development of analytical techniques, especially the analysis of textual information, social media is a source of data impossible to exclude from research on data science. Considering the particular context created in a country by a crisis, what people expressed on social media could provide additional information to assess the level of crisis. The data ecology of social network data, in relation to the political crisis, will help to assess if social media can be a reliable source of information to consider using in this research.

Social media is also associated with disinformation and political marketing. A review of the literature on political marketing and the application of bots in social media reveals the impact to be generally detrimental to democratic norms and social trust. That is unethical, in some countries, illegal and has dangerous societal consequences on the society, such as the decline in public faith in institutions and journalism associated with the rise of "fake news."

In *The Marriage of Politics and Marketing* (2001), Lees-Marshment describes political marketing as the union between traditional tenets of political science and marketing concepts. Recent definitions of political marketing have come to include the utilisation of marketing tools like social media and artificial intelligence as instruments that support the realisation of political objectives.

The development of a marketing strategy for an election campaign is very similar to a marketing strategy to promote a new product. This includes defining goals and objectives, developing a good understanding of the target group as well as opponents, and estimating the resources and channels required. The key factors of a political marketing strategy are its goals and objectives, which makes it different from product promotion. The usual target of a political marketing strategy is a group of constituents whose political sentiments are to be swayed in some way, typically to look at specific candidates, parties, or agendas (un) favourably. Ultimately, those deploying a political marketing strategy hope that this sentiment will result in the target group acting in a certain way, such as voting for or donating to a political party. The recent use of political marketing has demonstrated that these objectives are increasingly pursued to promote an idea and influence people to adhere to it by engaging them with online tools.

With the significant improvement in the capacity of algorithms to collect and analyse an enormous quantity of data from social media users, several authors have analysed the role of digital technology in recent political campaigns. In the United States presidential 2016 election, Chester and Montgomery (2017) studied the role played by technology and stressed that technology is making an increasingly significant contribution to the political sphere. They raise serious concerns related to privacy, manipulation, discrimination, and lack of transparency. Which also comes with legal challenges raised by the power of social media. For instance, some practices are becoming standard operating procedures, like "fake news." The practice of "fake news" is becoming part of the business, and it is reinforced by the lack of policy to protect user rights. In the 2016 United States presidential election campaign, the Trump campaign used Facebook's digital marketing system to target specific voters who did not initially support him. In addition, the campaign used psychographic messaging designed to discourage its opponents from voting (Green & Issenberg, 2016, cited in Chester &

Montgomery, 2017, p.8). Algorithms and social networks have also been deployed outside North America by political operatives looking to play on voters' emotions, as demonstrated in a number of political campaigns, from the Brexit referendum in the United Kingdom to electoral campaigns in Kenya.

6.2.1.1 The (negative) influence of political marketing on human behaviour

On social media, it is possible to filter a targeted population of users by gender, age, nationality, or location. This option is available when advertising with Facebook. Less refining criteria are available for Twitter, but selecting a group based on their location or interest is still possible. To analyse how social networks influence human behaviours, Althoff, Jindal, and Leskovec (2017), studied 791 million online and offline actions of six million users over the course of five years. They established "a causal effect of how social networks influence user behaviour". The usage of Facebook and Twitter for sharing political information results in a higher level of participation (Halpern et al., 2017). If some authors identified social media as a means of targeting people, Halpern et al. (2017) found that the platforms people choose influence how they participate in politics. For instance, choosing between Facebook or Twitter depends on criteria like the social network used by friends or relatives. However, the influence of both is noticeable when exposed to political content.

Social networks can make a telling contribution to political engagement. Facebook helped to facilitate the organisation of protests like the Euromaidan protests in Istanbul. On Facebook, for instance, there is a network of "friends," and on Twitter, it is a network of "followers" in which each user can decide whom to "follow." In both cases, the informational source is essentially ignored to the benefit of the sender of the information (Jost et al., 2018). The role of social networks could also be based on the type of user information managed by

the platform. Facebook collects private information from users to facilitate the "Friends" connection, whereas Twitter connects users based on their interest "Hashtag."

Regardless of the social media platform, the user operates and interacts inside a limited view of the entire platform, resulting in a perspective that is necessarily dictated by the information received from the user's idiosyncratic network. There is no or little contact with news from other people that are not affiliated as a friend or followers group. That raises the issue of independent verification of shared news items and the problem of confirmation bias in a so-called "echo chamber."

The existence of an echo chamber itself is arguably a necessary but insufficient condition to influence individual behaviour. Instead, the nature of the content shared within this network can generate an extreme emotional reaction and ultimately trigger action. With the rise of intentionally partial or false information, there are many examples where an online campaign has generated physical action. For instance, an investigation using Facebook data has shown that social media can act as a propagation mechanism between online hate speech and real-life violent crime (Müller & Schwarz, 2018). The propagation mechanism will have to consider the users' emotion to lead to real-life violence. This negative effect cannot only be obtained with fake information, a piece of accurate information can also lead to the same results. The use of bots as a means of automating and multiplying this process ensures that the impact of such practices can be used to influence national election results.

6.2.2 Political bots and disinformation

Algorithms or bots used in social media run a series of codes, which aim to simulate the human act of performing informational tasks. In a 2017 study, between 9 and 15 percent of Twitter accounts were identified as bots, equivalent to approximately 30 million accounts (Varol et al., 2017). Similarly, in 2018 Facebook announced that their platform hosts more than

300,000 bots (Khari Johnson, 2018). Bots appear to be a major source for diffusing information, including political information, which anyone with enough computing skills can build into Facebook or Twitter a bot.

Political bots are among the most recent tools for specialists in political communication. They play a key role in digital democracy and electoral propaganda (Howard et al., 2018). Political campaigners need personal records from citizens to narrow the political debate. Individuals' information available on social networks helps target a segment of voters and tailor the messaging accordingly (Kreiss & Howard, 2010 cited in Howard et al., 2018, p.85). With the development of new technologies, data are tracked more extensive layover time for better monitoring and marketing.

6.2.2.1 Models of political bots

Bots or chatbots used for conversational purposes have the same structure. A sample conversation is initiated when the user enters the text, and the bot responds with an appropriate message. The bot could also initiate the conversation to engage the user in a conversation and respond to the user's request (Ravindra, 2018). In the case of a political bot, the objective of the conversation could be to influence a voter. Many bots employ machine learning to adapt their actions based on the training data.

Ravindra (2018) identified two types of information as part of a feedback loop to become more sophisticated. To engage in a conversation, a chatbot must first know the "intents" or the user's intent. For a political bot, this relates to the information that can influence the political consumer. The second type of information is the "entities", in other words, the specific data used by the political bots to influence a voter to attend an event or donate to a campaign. The two types of information constitute the dictionary object. The current intent, current entities, and persisted information that users would have provided during previous interactions

with the bot. Another type of information not mentioned by Ravindra (2018) is context. The context of a conversation among a group of relatives is not the same within a group of colleagues. A bot should be able to make this distinction; the predicted value should fit into the context. Different data-driven models of political marketing exist. Chester and Montgomery (2017) listed seven models.

- Cross-device targeting: the use of various matching processes to deliver over time a specific message on the device where the user might be more receptive at a specific time. The matching could be with the IP (Internet Protocol) address, the email, the MAC (Media Access Control) address, or the device type.
- Programmatic advertising: the automated advertisement of political content on various social media platforms.
- Lookalike modelling: this model clones data from those identified as valuable customers to identify and target prospective customers.
- Geo-location targeting: the information provided by the GPS, Wi-Fi, or Bluetooth help to target consumers based on their interests and buying behaviour.
- Online video advertising is a highly effective way to deliver content that triggers desired emotional reactions in viewers.
- Targeted television advertising: broadcast and television turned into powerful micro-targeting abilities to deliver precise and personalised advertising messages to individual voters.
- Psychographic, Neuromarketing, and emotion-based targeting: psychographics, mood measurement, and emotional testing, with recent development in neuroscience, cognitive computing, data analytics, and behavioural tracking are used to target and influence voters. That could help to predict the emotional impact of a campaign.

Models could be merged to better respond to the objectives of the bots. A cross-device targeting, coupled with a lookalike modelling, could be developed in a conversational structure.

6.2.2.2 The architecture of Conversational bots

Different social media platforms have different architectures. The Facebook News Feed algorithm can predict what someone wants to see based on past interactions with friends or brands. Like Facebook, the Instagram feed algorithm first shows posts a user is willing to see on time-based. Twitter recommends tweets based on interaction history. LinkedIn shares content based on user preferences. As the platform source code is not typically open, these characteristics make each platform unique (Christine, 2018). All platforms have "relevance" as a common factor but deploy idiosyncratic means to engage users. The result is that political marketers select different platforms to support different objectives. The bots manipulate content, and regardless of the characteristics of each social media platform, the process of engaging political customers will follow a conversational model.

A bot, like any application, is a program manipulating data and interacting with external services via an Application Programming Interface. Bots have different architecture according to the objectives of the political marketers. The design of a classic conversational chatbot with lookalike data-driven modelling could create political bots to engage people via a conversation on Facebook or Twitter. Based on lookalike modelling, data from an influencer on a social network will be gathered and used as input to create the bot's personality. This clone bot is a virtual representation of an existing influencer, fed and trained with data to understand how to have a conversation with a political customer. With various Machine Learning algorithms, the bot will learn to give the best suitable answer that could contribute to obtaining the objectives of the political marketer.

Sébastien (2017) presents an overview of technologies powering chatbots in Figure 6.2.2-1. The core technologies developed in the messaging platform get the message processed by Natural Language Processing. The bot will respond to real-time questions with data from the information sources computed in the "bot logic" with Machine Learning algorithms.

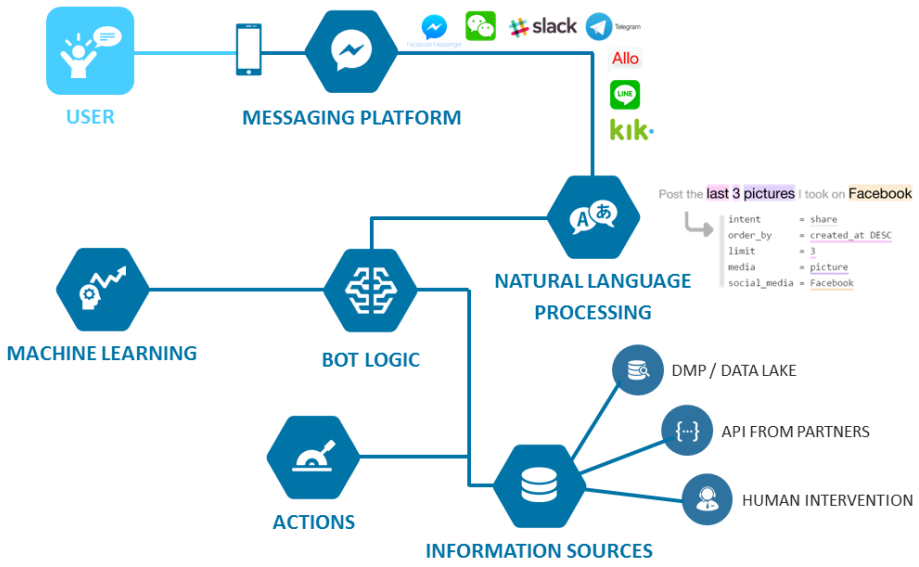


Figure 6.2.2-1 How To Explain The Technologies Powering a Chatbot? (Sébastien, 2017)

Despite their significant impact on human behaviour, bots have some limitations. Those limitations might be technical or ethical. Bots represent the intention of their creator; it is not the bot independently influencing human behaviour, but humans using bots to generate a specific emotion and action. With bots, politicians can use fake news for misinformation or for political gains.

6.2.2.3 The limitations of bots

While bots could influence humans, they fail to reproduce their behaviour. They still need a human contribution to be efficient. Some findings demonstrated that social bots are effective in manipulating users with low credibility content (Shao et al., 2018). Shao et al. provide evidence of the key role played by social bots in spreading fake news. Bots play the

role of fake news amplifiers and user influencers by targeting and interacting with people. This replacement of humans with bots does not mean the bot is the author of the fake news, but it contributes to spreading the news, which raises an ethical issue.

6.2.2.4 Ethical Limitations: The spread of Fake News

Twitter claimed to have removed more than 220,000 applications responsible for more than 2.2 billion low-quality tweets (Twitter Public Policy, 2018). This was a response to how the platform was used to influence the 2016 United States Presidential election. Social media bots are part of the political environment, and social media is the new arena for politicians.

The unique characteristic of fake news is the intention of the author to deceive the audience using false information. The source of the false information could be opinion-based or fact-based (Kumar & Shah, 2018). Social bots target a specific group of people to amplify the fake information and create a misleading perception of the reality within a given "information bubble."

From a data mining perspective, Shu *et al.* (2017) developed a mathematical formula to detect fake news on social media. They identified two components that can be used to identify fake news, the "News Article" and the "Social News Engagements." The news article contains the publisher $\vec{P}a$, which includes information about the author, like the name, domain, or age; the Content $\vec{C}a$, which is about the attributes of the news and includes the headline, text, or images. The second component, social media engagements, is defined by how news spreads over time $\varepsilon = \{e_{it}\}$ among n users $u = \{u_1, u_2, \dots, u_n\}$ and their corresponding posts $p = \{p_1, p_2, \dots, p_n\}$ on social media regarding news article a. With the time t the engagement $e_{it} = \{u_1, p_1, t\}$ of a user u_1 to spread the article a over the post p_1. These attributes and parameters help to determine the veracity of a news article shared on social media. In addition

to the technologies developed to counteract fake news, bots require a large amount of data to be performant.

6.2.2.5 Technical Limitations: The data requirements of bots

To engage in a meaningful conversation, bots require much information. The Users' Stories Framework is a technique to document the data requirements of bots (Katherine, 2018). To develop a social media user story framework, information is required concerning:

- User's information: Information on the social media users relates to their preferences, behaviour, and habits.
- User's roles: Identify roles and groups of users and their preferred affiliations.
- User's expectations: Identify why a user is on the social media platform to build the stories around the user's needs and expectations.
- User's map: Create a system with information collected to get a case for different groups.

To add to the user's data requirement, the classification of information for the training data of bots, previously described by Ravindra (2018), requires persistent information. Persistent information about the intent and the entities, as well as persistent information on the interaction between the users and the bots. That represents a large amount of information to collect and process, which requires computers and data centres to handle big data. The designer of the bot has to consider all the data required for a fruitful conversation, highlighting the key role played by humans to get the optimum result from a bot.

6.2.2.6 Human intervention

A bot cannot be fully automated; it requires human intervention. On certain occasions, a bot cannot perform an action with real-time information from the database; it needs to be

handover to a human (Sébastien, 2017). There are three main reasons why a bot requires human intervention (Amanda, 2017). The first reason is the requirement of an engineer to train the bot with the available data to respond with natural language. Even if all data are available, processing the data to respond like a human is a gap a human must fill. The second reason concerns decision-making. Bots can process the data but not make decisions. Not all decisions are binary; some decisions involve building trust or solving problems. The last reason is emotion. It is hard to program empathy or to give human emotions to a bot, which is a key factor of trust among social users.

Bots apply machine-learning algorithms to data in order to interact with humans on social media. They are traine d to answer questions. Human intervention will be required if a request is not in the training data. As said previously, bots are just automated programs that follow instructions.

6.2.3 Conclusion

Algorithms are becoming more sophisticated in understanding and interpreting human behaviour. The desire to harness this processing power for political ends could have an unpredictable impact on human behaviour. This section has analysed how artificial intelligence and social media negatively influence human behaviour in political participation from a data perspective. For that reason, social media data is not a reliable source of information for this research.

CHAPTER SEVEN: RESULTS

7 Results of the research

7.1 Data Preparation

Before analysis and processing, raw data is processed and changed. Preparing data for processing is a crucial step that typically involves correcting errors and merging different data sets to analyse the results (Talend, 2021).

In this work, we have considered three steps for preparing the data.

Step 1 – Eliminate fields with missing data

- As part of the WDI (217 countries, 19 indicators) and WGI (215 countries, six indicators), data before 1996 were excluded, including 1997, 1999, 2001, and 2020 as in WDI.
- The WGI and WDI transformed data is comprised of a country-by-country network of indicators. All the matrices are categorised by year.

Step 2 – Remove sample that does not fit the WDI and WGI

- Countries that do not have data in both the WGI and WDI have been eliminated. Table 6.1.9-1 shows the countries with several indicators from which countries are excluded.
- Table 6.1.9-2 was also updated to eliminate any countries recognised in the data preparation phase as small data countries; which are countries with more than 50% empty data.
- There are 25 indicators across 198 nations as in WDGI's final.

Step 3 – Eliminate sample with missing ACLED data

- ACLED data was used to create a yearly and country-by-year matrix of ACLED fatalities. The ACLED has a list of 101 countries.

- Any countries from WDGI list whose ACLED data did not match were eliminated. Due to the lack of data as in WDGI, South Sudan and Somalia were removed from the ACLED list. The Romanian Code ROM is being changed to ROU to match WDGI data.
- WDGI data for 100 countries was removed since ACLED data for such countries was not accessible.
- WDGI data for 99 countries were combined, and data for 101 countries were removed.
- The years 1996 and 1999 are removed from WDGI, and the years 1999 and 2001 are removed from ACLED.

The 3 steps above were conducted to get the final dataset that was used for this research. The merging and cleaning of the WDI, WGI and ACLED datasets resulted in a WDGI dataset of fatalities has 1980 observations and 32 variables (Appendix I-2).

7.2 Data Selection

Deletion of empty values

The process of selecting the suitable data source and type and appropriate methods to gather data is known as data selection. A selection of data follows the activity of data gathering. The data exploration with R is used to select indicators for analysis.

The indicators chosen for this study are assigned with a value for fatalities. With 0 as an appropriate value because a country has no fatalities as a result of political unrest in a given year. Indicators removed were the ones with missing value in ‘fatalities’ and with several missing data (Appendix I-4).

- There are 840 empty values as in the fatalities variable and indicators discovered as in data analysis. The indication 'IQ.CPA.TRAN.XQ' contains 1,197 empty data in 1,980 samples; 'SH.H2O.SMDW.ZS' contains 1,187 empty data; and 'SL.UEM.TOTL.NE.ZS' does have 989 empty values.

- When the fatalities' empty values were deleted, the indicator 'EG.USE.ELEC.KH.PC,' (which has 50% of empty entries), when removed from the list, gives less than 30% of the empty data as in the remaining indicators.

Deletion of outliers

After the deletion of empty values followed the deletion of outliers in ‘fatalities’. Outliers correspond to records where the number of fatalities is significantly higher than other records (Figure 7.2-1). An overview of the outliers shows there were 191 records summing 570,295 fatalities from 33 countries, between the years 1998 and 2019 (Table 7.2-1).

Region	Number of records	Sum of fatalities
Africa	141	265,052
Asia	50	305,243
Total	**191**	**570,295**

Table 7.2-1 summary of outliers

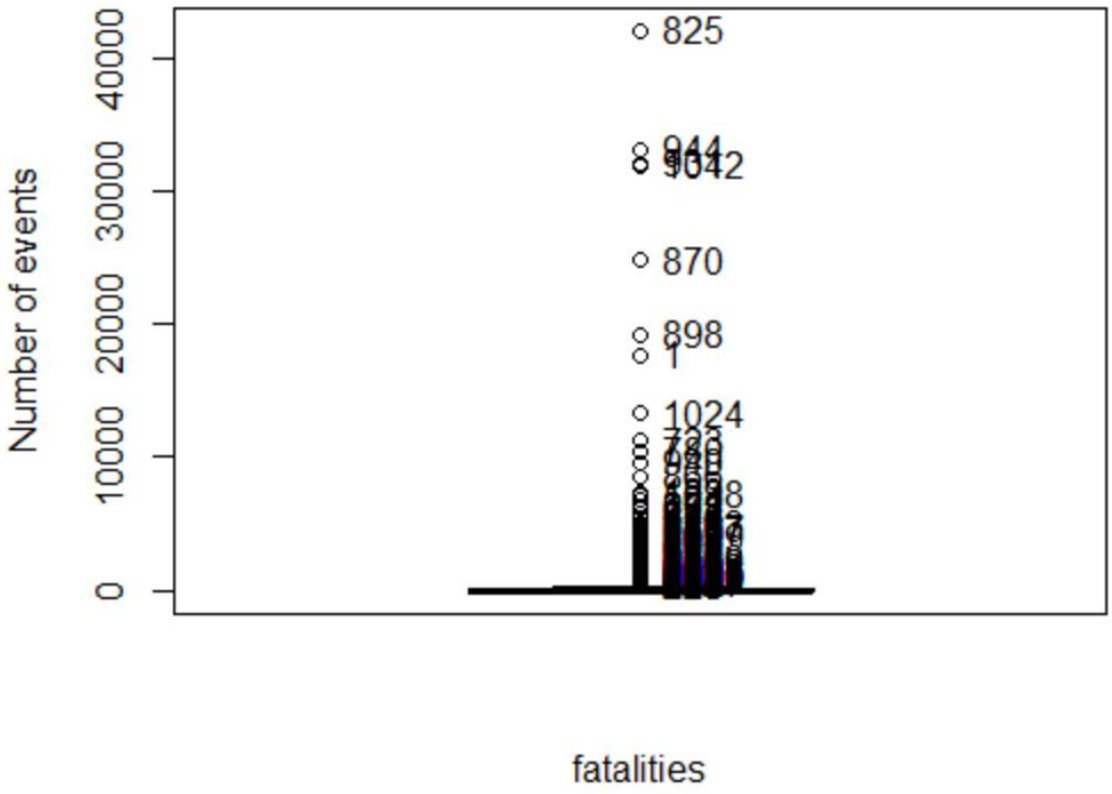


Figure 7.2-1 Boxplot of fatalities outliers

Based on previous experience with the regression analysis with crisis data and WDI, the variables used in Chapter four must be standardised, including fatalities. Once the standardisation equalised the ranges and data variability of fatality, the scale function using R subtracted the average of the variable and did not divide by standard deviation. The data stays as is to better react towards grouping data with fatalities. Standardising or normalising the data would not alter the correlation between the variables, so data stays as is to respond effectively to grouping data with fatalities.

7.3 Clustering

Image clustering seems to be an important but challenging problem in computer vision and machine learning. It is common for existing clustering and feature learning approaches to be overlooked. To achieve this, clustering attempts to group data points based on similarity. It is essential for several data processing and visualisation applications to have well-separated clusters of data points (Chang *et al.*, 2017).

Clustering has been broadly applied to map the input data into the feature space wherein separation becomes more accessible due to the dimensionality reduction with representation learning. Clustering was a process of breaking fatalities data into a predetermined number of groups so that records corresponding to every cluster possess characteristics that seem to be comparable to each other. In other terms, clusters represent areas with a high density of the number of fatalities that are comparable to each other. The sort of algorithm we employ would determine how well the clusters were formed, which will be determined by the algorithm we select. In addition, because there would be no established criterion regarding good clustering, all inferences drawn from sets of data depending on the user.

In this work, because the data is organised by countries, the clustering of fatalities is not determined by the number of fatalities among countries with a relatively large degree of violence. As in univariate analysis of deaths, the outliers are determined by finding nations with fewer than three records and deleting them. As a result, countries that have experimented with war had outliers eliminated from fatality totals. The clustering somehow does not assume the cluster's order. The cluster 1 value is assigned to the first fatality in the data set. As a result of this discovery, we now give the highest rating to 3 and the lowest number to 1. To address this, we first ordered rows with fatalities having ascending order sorted, and afterwards, the data gets separated and clustered (Appendix I-6). Table 7.3-1 shows Burkina Faso's clustering results with a cluster 3,4,5 and 6.

Number of fatalities	0	0	1	1	1	4	5	6	6	8	12	13	14	16	28	89	106	107	334
Difference	0	1	0	0	3	1	1	0	2	4	1	1	2	12	61	17	1	227	334
Cluster 3	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	2	2	2	3
Cluster 4	1	1	1	1	1	1	1	1	1	2	2	2	2	2	2	3	3	3	4
Cluster 5	1	1	1	1	1	2	2	2	2	2	3	3	3	3	3	4	4	4	5
Cluster 6	1	1	1	1	1	2	2	2	2	2	3	3	3	3	3	4	5	5	6

Table 7.3-1 Number of fatalities with clustering

Clusters of Burkina Faso fatalities in ascending order. All substantial variance is represented by Clusters 5 and 6, according to comparing results of all clusters. Nonetheless, Cluster 6 was a waste of effort. Cluster 5 seems to be the best choice for this circumstance. The fatalities clustering by country required each country to get a minimum of 5 entries for Cluster 5. Therefore, all country data with less than five records were deleted before the clustering.

7.4 Data calibration

Data calibration was not a step identified in the literature review, and it is not in the methodology but became necessary once the experimentation started because the data images of non-calibrated country data gave unique images of countries not differentiated by years, because years by country are grouped to create the correlation matrix. When calibrated with data each year could represent a unique image.

The 5 top countries with the highest fatalities before clustering have been used for the calibration. Data calibration used in this research has been extensively discussed in the next chapter (Table 7.4-1).

Country	Years	Fatalities
Cameroon	2015	387
Central Africa Republic	2016	382
Cote d'Ivoire	2004	374
Algeria	2008	366
Egypt, Arab Rep.	2019	359

Table 7.4-1 5 countries used for the data calibration

7.5 Database of images

The algorithm that creates the database of images is a double loop 'for' iteration, that created 5 folders representing each cluster group (Appendix I-7). In each cluster document there is:

- Data_raw giving the raw data from each country used to generate the image
- Data_corr that has the correlation of the raw data with crisis data as an output variable
- Data_img is the results, giving images that represent a visual representation of the correlation

The image generated is a visual representation of the centrality of the correlation of calibrated data (Figure 4.4-2). The database contains 833 images that will be used to train and test the CNN.

7.6 Convolution Neural Network

There are many different convolutional neural networks trained and used for various purposes. Fast R-CNN can detect regions of interest in an image, extracting characteristics before identifying areas of interest enhances the performance. Large-scale CNN architecture known as GoogleNet, also known as Inception v1, won the ImageNet Challenge in 2014. Error rates of fewer than 7% were attained, which is very near towards the human level. Small convolutions, known as "inceptions," batch normalisation, as well as other methods, were used to reduce the number of variables in previous architectures by tens of millions, around four million.

VGGNet (2014) To attain its performance, it employs 3x3 convolutions and then trained on four GPUs over two weeks. VGGNet contains 138 million parameters, making it tough to execute just at the inference stage, compared to GoogleNet. ResNet, a CNN having up to 152 layers, is an example of a ResNet. Employing "gated units," ResNet omits several convolutional layers. It utilises a similar batch normalisation approach as GoogleNet. Considering ResNet's novel design, additional convolutional layers could be operated simultaneously without adding complexity. Like a contestant in the 2015 ImageNet Challenge, it obtained a very acceptable error rate of around 3.57 percent, outperforming humans upon that trained dataset.

7.6.1 DCNN-11

Before the convolution, the image information (Table 7.6-1) shows that each image is a colour image with 3 colours (Red, Green, Blue), and 800x600 pixels size. The images needed a resize to reduce the weights and the computation time. The new size is one-third of the initial size (267x200 pixels).

```
Image
  colorMode    : Color
  storage.mode : double
  dim          : 800 600 3
  frames.total : 3
  frames.render: 1

imageData(object)[1:5,1:6,1]
     [,1] [,2] [,3] [,4] [,5] [,6]
[1,]    1    1    1    1    1    1
[2,]    1    1    1    1    1    1
[3,]    1    1    1    1    1    1
[4,]    1    1    1    1    1    1
[5,]    1    1    1    1    1    1
```

Table 7.6-1 image information

The resized images split into training data (583 images) ad test data (250 images) served to find the accuracy of the DCCN-1 model chosen. DCNN-11 (which includes four convolutional layers, four max-pooling layers, with three fully connected layers) appears to be the best model based on previous work. In terms of Convolutional modelling, at first, DCNN-11 uses four convolutional layers, lour max-pooling layers, including three fully connected layers (Appendix I-8).

7.6.1.1 Layers of convolution

Convolution, pooling, activation, and fully connected are the four most common types of filters used together in a convolutional network. A convolution filter is applied to the image to identify the image's features. For example, Weights are multiplied with inputs from the neural network through a convolution.

As part of the multiplication process, an image is passed by a kernel (for 2D arrays of weights) and perhaps a filter (for 3D structures). It is done diagonally right to left and top to bottom to cover the full image. Dot and scalar product—the mathematical operation is done during convolution and results in the convolution. Weights are multiplied by distinct input values for each filter. Every filter location has a unique value consisting of the sum of all the inputs.

ReLU Activation Layer

A non-linear activation layer, like the Rectified Linear Unit, has replaced negative integers as in filtered images using zeros.

Pooling Layer

With each layer, the image gets smaller until all the most relevant information remains. It is termed max pooling, and it can be used to keep either the pixel with the highest value or the pixel with the lowest value (i.e., average pooling). Overfitting is reduced by limiting network parameters and calculation numbers. Traditional multi-layer perceptrons and "fully connected" neural networks are found just at the end of deep convolutional neural network architectures after numerous repetitions of convolution or pooling layers.

Fully Connected Layer

Activation, as well as pooling layers, separate many of the fully-connected layers used in CNN systems. Convolution and pooling layers filter, adjust and decrease the image's flattened pixels before feeding them into fully connected layers. The softmax method is designed to such fully connected layers' results, indicating the probability of the image belonging to that class.

The model chosen includes fully connected, pooling and ReLU activation layers as follows:

- Layer one is the input layer having a convolutional layer of 32 3x3 filters, with activation function of ReLU, and 150 x 150 x 3 input dimensions.
- Layer two contains a pooling layer of 2x2 filter size.
- Layer three contains a convolutional layer filter of 64 3x3 with an activation function of ReLU.
- Layer four contains a Pooling layer of 2x2 filter size.

- Layer five contains a convolutional layer filter of 128 3x3 with an activation function of ReLU.
- Layer six contains a Pooling layer of 2x2 filter size.
- Layer seven contains a convolutional layer filter of 256 3x3 with an activation function of ReLU.
- Layer eight contains a Pooling layer of 2x2 filter size.
- Layer nine involves a flattening layer that converts the matrix to the vector with a fully connected layer.
- Layer ten involves the fully connected layer having 512 neurons with the activation function of ReLU.
- Layer eleven contains the output layer, a Fully connected layer having five neurons (corresponding to the five groups of the cluster) having the activation function of softmax for the output of probability.

The CNN model described above to classify images operates numerous elements of the images as "weights" or determinants of importance in categorising them as one from another. Its architecture consists of a series of components and standards, all of which revolve around the concept of "convolution". When the number of epochs grows, so does the accuracy gets increased.

7.6.2 CNN operation

There are three steps to the operation of a CNN. The first step is to ensure that all images are within the same pixel range and area. During the second stage, the data is reduced in dimensionality by deconstructing the images into many components, reflecting a distinct visual feature. All the required information is extracted from the images to classify them into

accessible output labels in the final step. This section explains the reasoning behind each of these steps, the hyperparameters they entail, and the available research choices. Torres and Cantú (2022) described a comprehensive method for processing images.

7.6.2.1 Pre-Processing of Images

To begin, we converted the images to a computer-readable format. Every pixel value in a picture is represented as an item in a numerical array in this image representation method. To ensure that the images utilised in a CNN have the same size, shape, and contrast range, all files must be optimised. Before anything else, it is important to make sure all the pictures are the same size squares. It is critical as squared images reduce the number of linear algebra calculations needed for the convolution (Rosebrock, 2017). Squeezing the image's most significant side, adding black bordering to the shortest side, and cutting it in centre are three common frequent methods of squaring the image. As a result, the latter approach is the most frequent, as it decreases how many pixels the model has to evaluate, allowing it to learn more quickly (Torres and Cantú, 2022).

The same pixel range must be maintained throughout all images. A bias towards images with high-value pixels is eliminated in this step. To keep the pixels in the [0,1] range, a popular scaling method divides the raw data by 255, the most significant pixel value. In the centred scaling procedure, the standard deviation of pixel values in information is used to split the difference between every pixel and the image's mean pixel value. Centring, instead of normalising, yields pixel values with a mean of 0 and a variance of 1.

The central unit of analysis is the input matrix. For CNN, the most essential information is extracted from the matrix when gradually shrinking the dimensions. Because of the pattern that convolution operated, CNNs do not give as much attention to the image's edges. To avoid this issue, it is recommended to add zero padding to the input matrix.

Finally, we split our database into two subsets: one for training, and one for testing. A model's training data is comprised of the samples it uses to learn the patterns associated with each of its outputs. For example, we can use test datasets to identify the parameters that best maximise categorisation and verify the model's predictions.

7.6.2.2 Representation of Features

There are many smaller matrices, called filters, in each convolutional layer representing a specific visual feature. In the first layer, the filters represent simple shapes, like straight or diagonal lines. Progressing through the levels, it is possible to see how the work evolves from simple lines to more complicated contours, shapes, and eventually, actual objects. Image complexity is increased by increasing the number of layers of a CNN (Buduma and Locascio, 2017). Each filter aims to get a distinct piece of information out of the matrix input. Each time a filter moves across an input matrix's width and height, it performs a cross-correlation, calculating the dot product among itself and that image subregion. Filter dot products generate a new matrix named the receptive field, that shows how well the filter "matches" a particular image area (Torres and Cantú, 2022).

7.6.2.3 Learning

The final stage of the CNN employs the extracted elements for predicting the image's label. Nodes in the output layer suggest that an input belongs to every label via a specific activation function when information reaches it. The output layer uses an inverse logit to estimate probabilities if the classification problem only includes two labels. Any other method uses a softmax layer, a multinomial logistic function comparable. The final layer of network comprises ten neurons and outputs, and one for each digit value from the number 0 to 9. These are used to recognise the image as a digit value. Each neuron will assess how likely the image

corresponds to a specific digit, and the model would use this information to assign the label to the number with the highest likelihood estimate.

The model returns from the output layer towards the input layer to identify the high-level features of images in this procedure. Weights between neurons are calibrated to reduce prediction errors with every step away from the output layer. As a result, the learning procedure is going over a collection of labelled instances numerous times to discover the best mixture of the feature weights and maps to minimise the gap between the predicted and actual labels.

Such discrepancies are quantified using the loss function, which serves as our aim for optimisation. The classifier's predictions are more accurate when the loss is reduced. For an OLS regression, the loss function seems to be the mean square error, and the difference between the predicted values $y\hat{}^{i}$ and observed values y_i is represented by the regression line of y_i. A binary cross-entropy of loss and categorical cross-entropy of loss is the default loss functions for binary classification in CNN (and practically all deep learning applications).

The CNN model was trained on 749 images and tested on 84 images; it obtained the loss value of 7s 2s/step value is 4.9942 with an accuracy of 0.7719.

The obtained actual and predicted values are presented in the matrix below

	0	1	2	3	4
0	244	0	5	26	15
1	6	127	7	26	13
2	5	2	79	13	15

Table 7.6-2 Actual and predicted values

The number of epochs, and the number of times each training sample is passed through the network, must be specified because of the gradient descent's iterative nature. The model gets more likely to be fit to its weight because the accuracy can be improve with training examples. The prediction errors in validation and training data are supposed to decrease with the number of epochs. As a result, we should end training when a loss in validation dataset does not drop any further. It prevents the model from learning aspects of images that were not applicable outside the training data.

Another consideration is how many training images must be sent across a network before the weights are updated. Unless a batch size is greater than or equal to the total number of the training images, the model would only update the weights after collecting prediction errors from all the training images. This process can use a significant amount of computing memory. The gradient descent is also prone to catch in local minimum due to static error surface. However, the batch size might be as small as one, meaning each training sample is updated in the model's weights. Every time a new case is analysed, it updates the model and reduces the danger of being trapped in a flat area. Training data is divided into small chunks, allowing the model that updates the parameters multiple times in an iteration. It is a middle-way strategy.

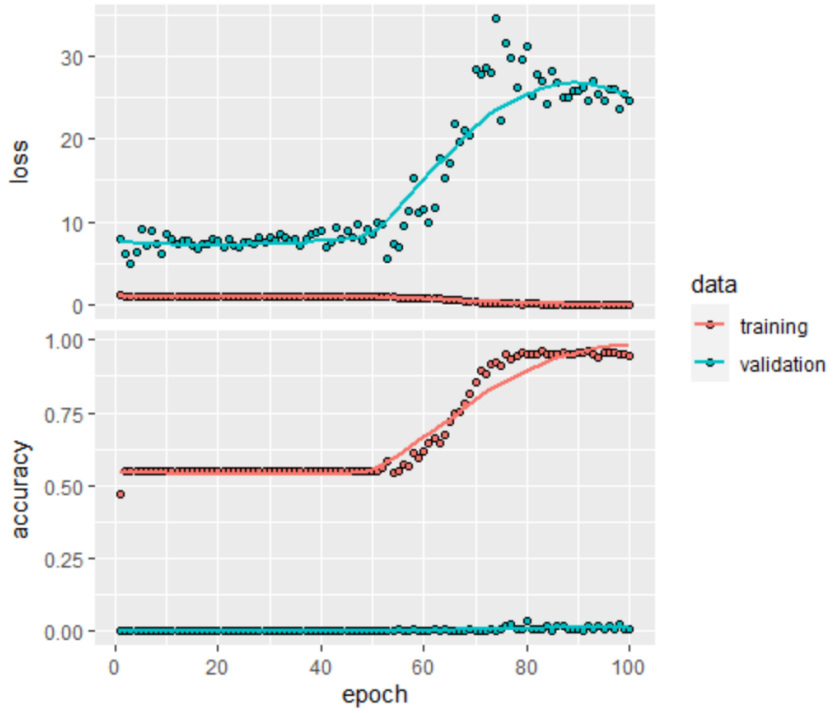


Figure 7.6-1 Epochs plot

7.6.2.4 Overfitting

The accuracy of a CNN model's predictions improves with practice, just as it does with any other human-performed work. For a model to learn relevant patterns in training data, it must be trained across several algorithm iterations. Hence, the results in underfitting, which occurs when the model gets too simplistic and accounts for all variation in the images. When attempting to classify the data, an under-fitted model makes erroneous assumptions, which results in a high degree of bias in its predictions. We need to run it through more training rounds to improve the model's ability to recognise patterns in the training data. However, overfitting occurred with the model trained for an excessive number of iterations. It violates the principle of machine learning, which pertains to the model's performance with instances outside the training set. The model starts classifying random oscillations in the training images as meaningful concepts during overfitting. Overfitting can be checked by monitoring the model's performance in validation and training sets for every epoch and then stopping while there is a

loss (or an increase in the model's accuracy) in the validity data (Figure 7.6-1). To avoid overfitting while still learning significant features of images, there is a practical issue that depends on both the number of training iterations and how well the model architecture is constructed. Changing the number of training iterations and the model architecture could not avoid the overfitting simply because of the nature of the images to classify. The groups used to train the models reflects cluster group and, in each group, images are not necessarily similar. The principal objective of the model is not to learn to classify images but to identify similar images.

7.6.3 Image similarity

The CNN model learned the features of the graphs and it is now able to identify similar images using the Cosine similarities. Cosine similarity is all about transforming two parameters into vectors in a multidimensional space. The angle between two images is measured to measure how the Cosine of that angle measures "similar" they are. The Cosine angle between vectors is the dot product of the two vectors if they are both equal in length.

The similarities between the images in the database and the image representing the country where risks should be assessed were identified following a process. The steps to find similar images are described below:

- Prepare the image database
- Download the trained model,
- Select the image of the country to assess
- Compute similarities between images' feature vectors using the Cosine similarity
- For the selected image, select the images with the top-k similarity scores to build the recommendation

The 833 images created from the correlation analysis have similar shapes, which makes it difficult for the human eye to distinguish what are similar images. The trained CNN model learnt the features of images and can identify similarities between one image and all images in the database. It is the trained weight of the CNN model that stores the extracted features from images.

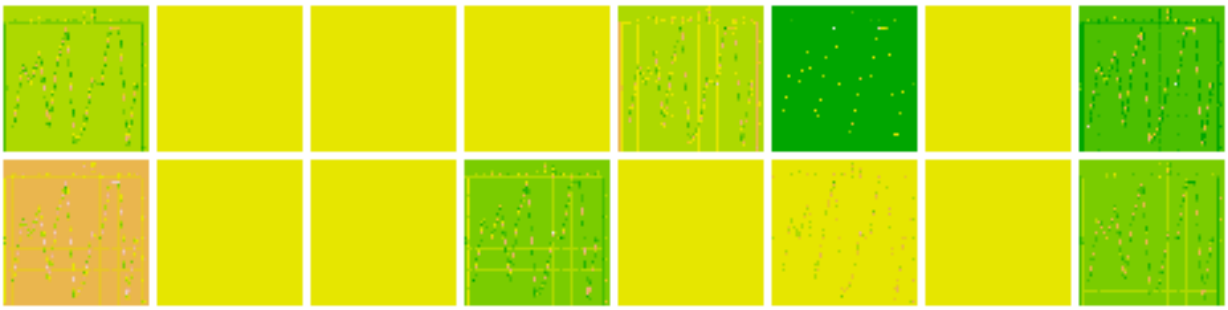


Figure 7.6-2 Sample features extracted from a graph

Figure 7.6-2 is a visual representation of the sample features learnt by the model. The features extracted from the image will help find similar images for risk identification. Feature extraction is particularly relevant for images with the same shape and visual representation.

7.7 Risk Identification

Compiling, aggregating, and providing proof to support an assertion about an activity and event's risk can be considered risk assessment. Many established methods, approaches, tools, and models are available to assist in the evaluation. It is possible to carry either qualitative or quantitative risk assessments or a hybrid approach.

Non-numerical estimations of risk are produced through the characterisation of risk through qualitative assessments. The level of risk is expressed numerically using quantitative instruments. Quantitative risk assessments, in contrast to qualitative method, is more easier to verify and models used in quantitative risk assessment can be basic. Risk assessment quantitative techniques are the emphasis of this work.

The process of expressing uncertainty (and unwanted assurance, dependability) like a future occurrence is known as risk identification. It is frequently accomplished through observation and collection, including both quantitative data (like indications of status, estimations of probability, and rates of past incidence) or qualitative data (like observations and collection of qualitative data) (specialist views, priorities, and opinions). Historical information serves as a feedback loop because companies could learn about prior risk incidents, which could then be used as input for risk identification. Organisational learning, interactions, and evaluation are used in leadership attempt to describe this process. Socio-political beliefs and socio-cultural norms could impact whether or not uncertainty is viewed as a possible danger, as well as how severe the consequences of its existence would be.

The identification steps in this work involve, if a country has fewer than five rows, the risk identification will build a correlation graph with values ranging between -0.5 to 1.

The performance of the models was improved by decreasing the quantity of countries wherein the convolutional network was applied. It is impossible to cluster five as every correlation graph within cluster 5 has the values as 1 or -1.

7.7.1 MCDM TOPSIS method

Following the predefined method of The Technique for the Order Preference by Similarity to Ideal Solution (TOPSIS), first developed by Hwang and Yoon in 1981, is a Multi-Criteria Decision Making method for selecting the best option from a limited number of options. It starts with creating a decision matrix that depicts the satisfaction level of each criterion with every possible alternative. The values are then multiplied by the criterion weights once the matrix has been normalised according to the selected normalising scheme. Then, the negative and positive ideal solutions were measured using a distance measure, and their

distances were determined. Finally, the options are rated according to their proximity to the ideal solution.

In the TOPSIS method, decision-makers can organise problems to be solved, compare and rank the alternatives, and analyse and compare the results. The standard TOPSIS method can solve problems with a clear set of numerical decision facts. However, the structure of most real-world situations is more convoluted. More than a dozen variations on the original TOPSIS technique are based on it. These variations include the inclusion of an interval and "fuzzy" criteria and interval, as well as "fuzzy" weights that account for various types of ambiguity or vagueness in the underlying model. The fuzzy TOPSIS applied to various areas like the environment, energy or healthcare (Palczewski and Sałabun, 2019).

Deciding with several, often competing criteria is referred to as "Multi-Criteria Decision Making" (MCDM). MCDM refers to selecting the best option from a limited number of options. When it comes to making a multi-criteria decision, the following are the most important steps:

- We create particular criteria for evaluating system capabilities that are linked to specific aims
- New approaches for achieving the desired outcomes (generating alternatives),
- Comparing the merits of several options based on predetermined criteria,
- Normative multi-criteria decision-making techniques,
- Rejecting all other options and accepting only one as "ideal" (preferred),
- Gathering new data and trying again if the final solution is not acceptable.

Applying multi-criteria decision-making approaches can aid in selecting options in discrete situations. In particular, these methods, made more accessible for users by computers, have gained wide adoption in many areas of economic and managerial decision-making

processes. ELECTRE, SMART, TOPSIS, AHP, SAW, MAX MAX, and MAX MIN are the most often utilised multi-criteria methodologies. Choosing, ranking, or sorting are all strategies that have different suggestions depending on the problem they are trying to solve. Evaluation criteria for selecting models and methodologies include, but are not limited to:

- Consistency and logical soundness, both internally.
- Transparency,
- Accessibility, and convenience
- The amount of data needed is proportional to the significance of the subject being studied
- Analytical process time and labour resources must be realistically estimated an audit trail, and the ability to offer
- The availability of software, if necessary.

While the MAX MIN method relied on the weakest attribute to assess an alternative's overall performance, the MAX MAX method prioritises the best attribute value. In the SAW (Simple Additive Weighting) approach, the normalised value of the criterion for every alternative gets multiplied by the importance of requirements, and the alternative having the most significant score was chosen as the preferred option. TOPSIS finds the option closest to a perfect solution as well as the farthest away from the ideal negative alternative. When using the classical TOPSIS technique, the only subjective inputs are weights, based on the decision maker's knowledge of attribute values. Paired comparisons and a hierarchical structure are used in the Analytical Hierarchy Process (AHP). Multiple criteria that describe different solutions are at the middle of an AHP hierarchy while competing alternatives are at the bottom.

7.7.2 P-Value method

The P-value test the hypothesis for a ‘country at risk’, what is the country that can best inform about the indicators to target to reduce the number of fatalities. A country that reduced the number of fatalities over the years has proceeded with some changes in the economy, governance or human development. To advise a country on the indicators that can affect the number of fatalities, the CNN identifies for the ‘country at risk’, the country with a similar representation of indicators’ centrality, with a cluster above 1, representing a high number of fatalities. The ‘similar country’ with a high number of fatalities, with a similar social, political and economic context, and that over the years reduced the number of deaths due to crisis will give a list of ‘improved situation’ to inform the ‘country at risk’ with the development factors that have been significantly addressed. Over the years, the improved situation of the ‘similar country’ represents many alternatives for the ‘country at risk’, and the P-value will help to identify the best alternative. This research has tested this method as an alternative approach to identifying indicators for risk assessment. It is simple and easy to use compared to the TOPSIS and emphasises the variation of indicators when the TOPSIS helped more on the countries' ranking. If TOPSIS and P-value gave similar results, the P-value required much less computing resources and gave more flexibility for a qualitative review of the results.

The P-value used in Chapter four for the regression analysis was used to analyse the distribution of the variables. In the risk assessment, the P-value is used to identify the country with the significant changes in the variables compared to the country at risk to assess. The Null-hypothesis usually presented as the absence of a relationship between two sets of variables, in this case the absence of a relationship between the alternatives of the ‘improved situation’ and the country at risk’ implies that the P-value is close to 1. The P-value varies between 0 and 1, and the more the P-value is close to 0, the more a statistical relationship could exist between variables. The best alternative for a ‘country at risk’ is the one that statistically has variables

not close to ones of the ‘country at risk’. The best alternative will then be the alternative that has the highest P-value.

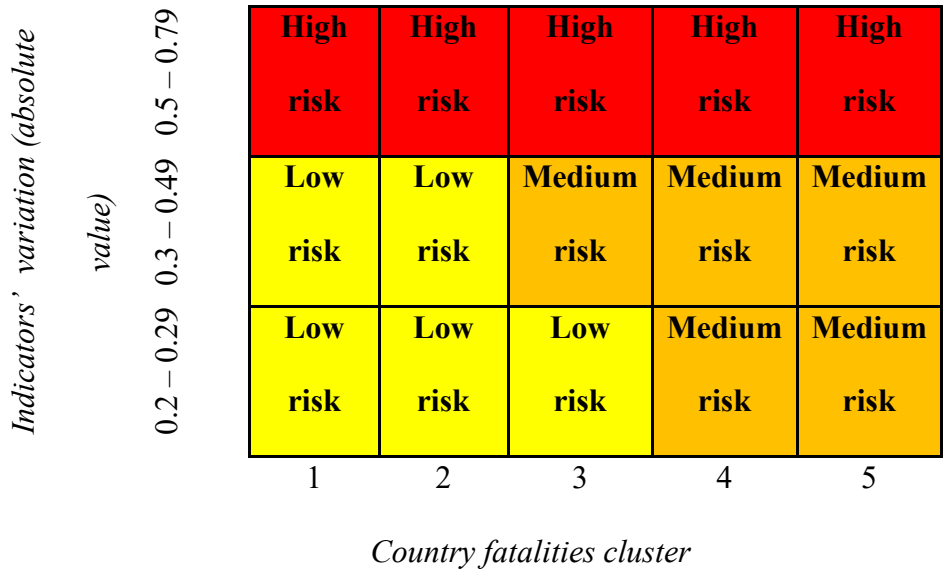


Figure 7.7-1 Risk identification matrix of a ‘country at risk’

The risk identification matrix in Figure 7.7-1 assesses the level of risk of an indicator based on the variation of the indicators and the cluster of the ‘country at risk’. The P-value determine the best alternative for a ‘country at risk’, and the variation of the indicator is calculated with the difference between the value of the indicator of the best alternative and the same indicator for the ‘country at risk’. Indicators above 0.5 difference are essential to reduce the number of fatalities due to crises.

7.8 Use case: Risk assessment of Niger, Mali and Burkina Faso

In this work, I have considered the three research cases namely, Niger, Mali, and Burkina Faso. By analysing the case of three countries that have faced repeated crises in the last 10 years, I have found that Niger, Mali and Burkina Faso are neighbouring countries that also face similar socio-political and economic and terrorism related challenges. It will be interesting to analyse with the newly developed model what could be the indicators that could

contribute to improving the situation in the country without causing more deaths. The more recent number of fatalities recorded for NMB is also the highest number of fatalities recorded in Burkina Faso and Mali. The below table shows the images taken from the three different countries.

Burkina Faso	2019	5	5/data_img/BFA.1890[5].jpeg
Mali	2019	5	5/data_img/MLI.1938[5].jpeg
Niger	2019	3	3/data_img/NER.1946[3].jpeg

Table 7.8-1 List of different countries having different images

To identify the indicators that can significantly contribute to reducing the risk of violent crisis, the first step will consist of identifying countries that have been in a similar situation that will be made with the CNN; and the second step will consist of identifying the indicators that have helped the countries that have faced a similar situation to improve.

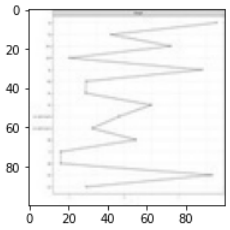
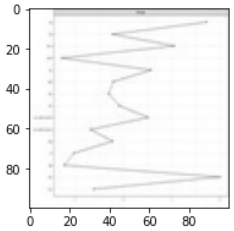
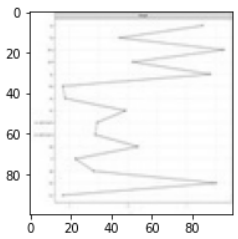
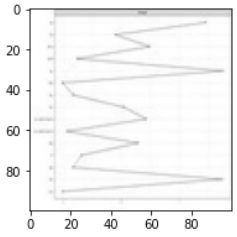
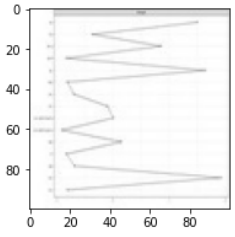
7.8.1 Finding similar countries with CNN

7.8.1.1 Burkina Faso – 2019 – Cluster 5

The CNN algorithm identifies three countries that had a relatively similar situation at a particular time regarding the political, social and economic indicators in relation to fatalities. These countries are: Niger in 2006 for a Cluster 1; Senegal in 2012 for a Cluster 2; again Senegal in 2002 for a Cluster 3; and Mali 2006 for a Cluster 1. The countries identified are all from the same region, Western Africa.

The country that faced a similar simulation and managed to improve (by reducing the number of fatalities throughout the year is Senegal. Niger and Mali are Cluster 1, therefore cannot inform about a way to improve the situation in Burkina, but it is possible to learn from Senegal, finding the Cluster 1 image from Senegal that came after the year 2012, which gives

4 images for Senegal. The below table shows the results for Burkina Faso, Senegal, Mali, and Niger.

BFA.1890[5].jpeg	2019	1.00000000000000 02	
NER.659[1].jpeg	2006	0.99999984335238 52	
SEN.1266[2].jpeg	2012	0.99999976602809 85	
SEN.276[3].jpeg	2002	0.99999974034751 55	
MLI.651[1].jpeg	2006	0.99999949188609 97	

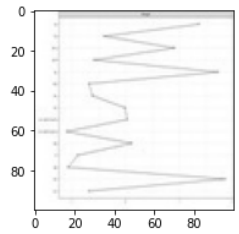
BFA.1296[2].jpeg	2013	0.99999930949883 66	

Table 7.8-2 Centrality plot for Burkina Faso, Senegal, Mali, and Niger

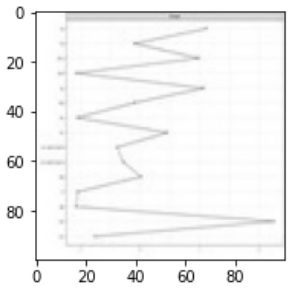
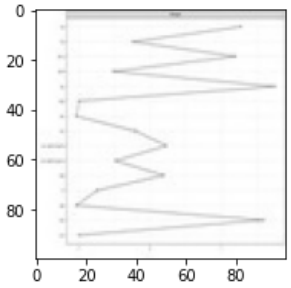
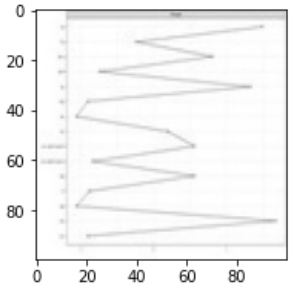
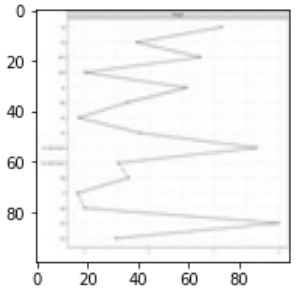
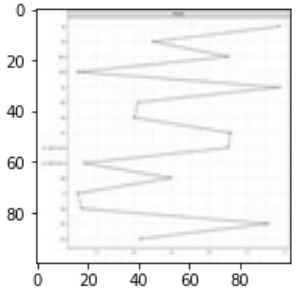
Senegal	2019	1	1/data_img/SEN.1959[1].jpeg
Senegal	2017	1	1/data_img/SEN.1761[1].jpeg
Senegal	2016	1	1/data_img/SEN.1662[1].jpeg
Senegal	2015	1	1/data_img/SEN.1563[1].jpeg

Table 7.8-3 List of Senegal image data for different years

7.8.1.2 Mali – 2019 – Cluster 5

For Mali in 2019, Cluster 5, the algorithm identifies two countries with similar images from a different region, Eastern Africa. Uganda Cluster 1 (2017) and Madagascar Cluster 1 (2011). A new country that was not identified with Burkina Faso, is Benin (2000, Cluster 1). Benin is also from Western Africa like Niger (2017, Cluster 3) also has a similar representation of the latest context in Mali.

Like for Burkina, Senegal is the country that improves its situation regarding fatalities over the year. For Burkina, that will allow a study of all images on Senegal with Cluster 1. Because it is similar to Cluster 5 in Senegal, the assessment could include Cluster 4 images or above, for the year after 2006. That could also include the same four images to use for Burkina Faso. The below table shows the results of Burkina Faso, Senegal, Mali, Nigar, Uganda, and Madagascar.

MLI.1938[5].jpeg	2019	0.9999999999999999	
SEN.672[5].jpeg	2006	0.9999998438671809	
BEN.107[1].jpeg	2000	0.9999998196589783	
UGA.1774[1].jpeg	2017	0.9999997857893917	
NER.1748[3].jpeg	2017	0.9999997331906536	

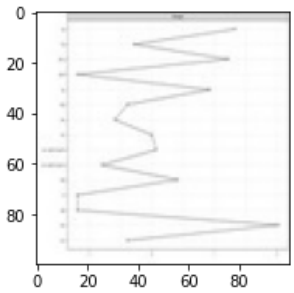
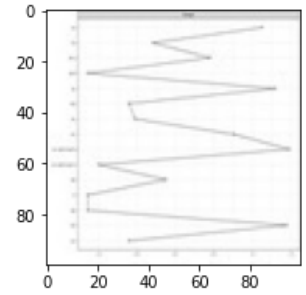
MDG.1144[1].jpeg	2011	0.99999972904398 54	

Table 7.8-4 Centrality Plot for Burkina Faso, Senegal, Mali, Nigar, Uganda, and Madagascar

7.8.1.3 Niger- 2019 – Cluster 3

The more recent image of Niger is similar to the image of Madagascar (2011, Cluster 1), Benin (2000, Cluster 1), but also to the image of Sri Lanka (2019, Cluster 4) and Sri Lanka (2018, Cluster 3). This context similarity between Western African countries and Southern Asia countries is interesting to explore.

Sri Lanka is the country where the context will be studied, analysing the recent images above Cluster 3. There is no data after 2018, so this will allow analysing from past experiences with Cluster 1. The below table shows the results for Niger, Madagascar, Benin, and Sri Lanka.

NER.1946[3].jpe g	2019	1.0	

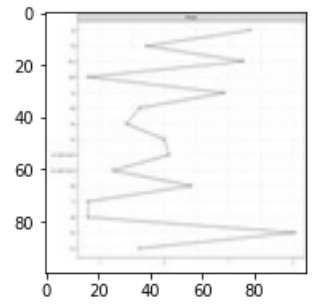
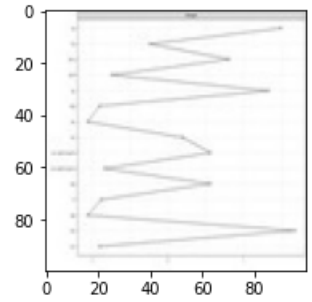
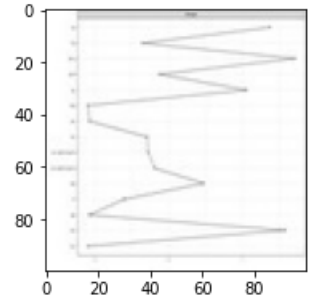
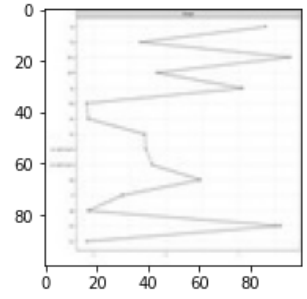
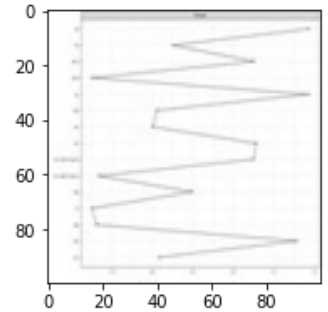
MDG.1144[1].jpeg	2006	0.9999998596271662	
BEN.107[1].jpeg	2000	0.9999998498163879	
LKA.1932[4].jpeg	2017	0.9999997937355272	
LKA.1833[3].jpeg	2017	0.9999997937355272	
NER.1748[3].jpeg	2011	0.9999997454510112	

Table 7.8-5 Centrality Plot for Nigar, Madagascar, Benin, Sri-Lanka

Sri Lanka	2017	1	1/data_img/LKA.1734[1].jpeg
Sri Lanka	2016	1	1/data_img/LKA.1635[1].jpeg

Table 7.8-6 List of Sri Lanka image data for different years

7.8.2 Identification of indicators at risk

The below table shows the analysis of the centrality of various countries shows that for Burkina Faso, the centrality corresponding to the highest P-Value is Senegal (2014, Cluster 1), while for Niger is Senegal (2016, Cluster 1). The four records of Senegal in the table correspond to the year where Senegal has the lowest level of fatalities. The P-value is calculated with the centrality of each respective year with the centrality data from Burkina Faso and Mali. The highest P-Value reflects the year where the centrality is the most different to the ones from the country we are assessing.

	Years	**Cluster**	**Burkina**	**Mali**
Senegal	2019	1	0.2 99403	0.165959
Senegal	2017	1	0.34413	0.44251
Senegal	2016	1	0.272923	0.489087
Senegal	2015	1	0.272923	0.489087
Senegal	2014	1	0.391046	0.402007

Table 7.8-7 Analysis of Centrality of various countries with P-values

That allows seeing the socio-politic and economic context of Burkina is significantly different than the context in Senegal in 2014, with an easy identification of indicators varying.

For Niger, the learning is coming from an Asian country, Sri Lanka. (It will be interesting to analyse what Niger and Sri Lanka have in common in relation to a crisis). The table below shows that the situation of Sri Lanka in 2016 will give the more prominent indicators to target in Niger to reduce the number of fatalities in Niger.

	Years	Cluster	Niger
Sri Lanka	2017	1	0.206601
Sri Lanka	2016	1	0.279259

Table 7.8-8 Analysis of Centrality for Sri Lanka and Niger with P-values

By calculating the difference between the centrality of countries identified as relevant with the P-Value, and the centrality of the country in crisis, we can identify prominent indicators to target. From the 15 indicators used for this analysis, 5 indicators in total seem to be critical. For Burkina Faso, the first and more critical indicator is the "GDP per capita growth" (0.64) and the second most critical indicator is the "Voice and accountability" (-0.54). The negative and the positive value of the difference have to be taken in absolute value because they reflect the centrality, of one indicator in relation to the other indicators. This analysis does not inform about any causality but about the correlation of changes in socio-politic and economic indicators in relation to the crisis.

For Mali, the key indicators are the "GDP per capita growth" (0.55) and "Domestic credit to private sector by banks" (-0.59). And for Niger, the "GDP per capita growth" (-0.78), the "Political Stability and Absence of Violence/Terrorism" (-0.65) and "Population"(0.56) are the principal indicators to analyse, as shown in Table 7.8-9.

In an attempt to interpret the results, the crisis in Burkina, Mali and Niger has affected the GDP growth. But the growth of the GDP is also related to poverty reduction, which could also mean that poverty has been a trigger factor for the crisis. Particularly in Burkina, poverty is linked to accountability and local governance when the ACLED data shows that most of the fatalities happened against the killing of the journalist, Norbert Zongo. In Mali, the GDP is linked with financial support to the private sector, which is also reflected in the ACLED data

of the country in 2019 showing that many people were killed due to ethnic clashes, in relation to natural resources (land and water). In Niger, political instability seems to be the determining factor together with the population (24 million), with Niger reported to have the highest birth rate in the world.

	FD.AST.PRVT.GD.ZS	NY.GDP.PCAP.KD.ZG	SP.POP.TOTL	TG.VAL.TOTL.GD.ZS	PV.EST	RL.EST	VA.EST	
2014	-0.10554	0.642245	0.287023	0.242172	-0.30766	-0.19983	-0.54995	Burkina
2016	-0.59963	0.556631	0.472685	0.127066	-0.31796	0.025786	-0.33604	Mali
2016	-0.14161	-0.7872	0.564797	0.472728	-0.65807	-0.4111	0.030906	Niger

Table 7.8-9 Indicators at risk in Burkina, Mali, and Niger

The data that NMB shared in common is the GDP. Poverty has to be targeted as the most important regional problem for Burkina, Mali and Niger, addressed together with internal and external factors of political instability; the support of banks to the private sector; a policy to handle the growing population; and the need for accountability from the government. From the data perspective, that might contribute to reducing the number of fatalities and simultaneously bring development.

CHAPTER EIGHT: DISCUSSION AND CONCLUSION

This research aims to develop a new method of data for assessing the risk of violent crisis by using graph convolution and network analysis. This study used network analysis to discover and explore the systems of development index which anticipated a damaging crisis and then used graph convolution to find the development index changes that would lead to a violent crisis. A considerable number of variables, including non-linearities and non-dependencies, enhance decision-making. The objectives of this research are attained by using weighted correlation patterns and networks for the result of socio-political crises. The risk assessment model is developed depending on the graph convolution network. It is capable of examining the patterns of variations in development indicator networks.

There are only two components to convolutional neural networks: convolutional and pooling layers. Despite their simplicity, these layers can be arranged in nearly infinite ways to solve a given computer vision problem (Brownlee, 2019). As a result, it is possible to build every CNN model using common patterns for layer configuration and architectural innovations.

Designing model architectures that best use these essential elements is challenging for putting convolutional neural networks to practice. It is effortless to do because of the extensive research and application of CNNs for the ILSVRC (ImageNet Large Scale Visual Recognition Challenge) from 2012 to 2016. As a result of this challenge, the state-of-the-art in computer vision has advanced rapidly, as has the architectural style of CNN model models in general. As a follow-up to our previous discussion of LeNet-5, other networks such as AlexNet, VGG, Inception, and ResNet also play winning architectural innovations for ILSVRC's convolutional neural network (Szegedy *et al.*, 2014).

Inception and GoogLeNet, two groundbreaking convolutional layer innovations, were introduced in 2015, and both performed admirably in the 2014 iteration of the ILSVRC challenge. ResNet, or Residual Network, is a deep convolutional neural network model that won the 2015 ILSVRC challenge with impressive results.

This research extends the limit of the innovation with CNN by experimenting with a correlation centrality plot to find similarities.

8.1 Preliminary research

8.1.1 Comparison to other risk assessment methods

This research proposed a data-oriented risk assessment method, including a hybrid approach with a qualitative and quantitative risk assessment. A combination of quantitative and qualitative risk assessment methods is widespread. Qualitative rating or ranking approaches are employed to identify key variables in conjunction with quantitatively oriented analytical tools that enable the weighing and prioritising of the criteria. The risk assessment method designed in this research is used to identify and evaluate preventative and mitigating strategies.

Similar to the **Root Cause Analysis**, also known as Root Cause Failure Analysis. This method is used to discover why, what, and how something went wrong. After failed occurrence corresponding to a deadly crisis, it is used to examine substantial social economic and political changes. Even if this method investigates the causes, it does not look at all probable causes of risk as required by a Cause-and-Effect Analysis.

The regression analysis in Chapter seven shows the limit of multiple regression analysis to predict country risk, underlining the importance of assessment instead of prediction for a country. Why it is better to refer to risk assessment and not risk prediction? Because predicting a risk it is less relevant. The probability between 0 and 1 about whether a crisis could happen is important, because crises are a vital part of a country's evolution. However, this research

intended to identify factors that could help to avoid significant death as a result of a crisis. Putting human life as the pivot indicator to find the changes that could help minimising the deadly effects of the crisis. That is why this research is an assessment and not a prediction, as could be expected from data science research. For prediction, in future research, the discriminant analysis approach can be used with the discriminant coefficient as a method that gives early alerts in case of future risks.

8.1.2 Statistical Method: Regression analysis

The regression analysis investigated the relationship between DI and socio-political crisis events in low-income countries from a data perspective. The regression analysis has two main objectives. The first objective consisted of the analysis of the DI to identify indicators that will best fit a regression model, and the second objective listed indicators that are key in a regression model to predict the socio-political crisis. The regression analysis applied to 15 WDI as independent variables and the crisis event data generated by the ACLED project, as a dependent variable.

Two regression models were experimented. The multilinear and the Poisson regression came to the same conclusions, the growth of the population increases the risk of social instability and population data can contribute to predicting it, as well as associated with other indicators (the GDP, the military forces personnel and the external debt). The merchandise trade indicator alone or associated with foreign investment indicators as factors decrease the risk of a crisis, and can contribute to predicting it. This has demonstrated the importance of development data in relation to the crisis, in which further analysis such as time series or country analysis can contribute to deepening.

The regression analysis demonstrated that a link exists between the development indicators and crisis data, but did not suffice to identify the indicators triggering death in a

crisis. This research proposed a method not only to identify triggering indicators but also for specific countries and a precise year.

8.2 Data limitations

8.2.1 Social Media Data

This research has investigated how artificial intelligence and social media are being used to influence human behaviour in the realm of political participation. A review of the literature on political marketing and the application of bots in social media reveals the impact to be generally detrimental to democratic norms and social trust. That is unethical, in some countries illegal and has dangerous societal consequences on the society, such as the decline in public faith in institutions and journalism associated with the rise of "fake news" or false information presented as true. Internet-based information and social media, in particular, have the potential to improve emergency preparedness and response. Streaming feeds from users provide constant information updates to social media platforms like Facebook, Twitter, and electronic newspapers. Increasingly, organisations are depending on this type of data to detect brand crises — that is, when a brand has an unexpectedly high frequency of unfavourable comments on online channels.

The proliferation of data-intensive issues had accelerated since the introduction of social media data over the world. Modern software tools and technologies have a difficult time seeing, exploring, managing and analysing digital data because of the abundance and rapid growth of this data. The most important reasons for and methods by which SM data increased are the abundance of the data volume, diversity of data variety, as well as incoming/outgoing data velocity (notion called as the 3Vs). According to a National Security Agency assessment, for example, more than 1K Petabytes of data has been processed daily via the Internet. Digitised data increased ninefold between 2006 and 2011 and is expected to reach 35 trillion gigabytes

by 2020 (Chen and Lin, 2014). Various fields, like scientific research, government administration, business, industrial sector, health, education, and then others will benefit greatly from this massive increase in digital data, but disinformation is poisoning the quality of the social media data.

Disinformation highly affects the quality of the data, and with the power of political bots, the impact of disinformation is amplified. The social media data have quality issues, together with ethical issues and bias. They do not always reflect the real context and can hardly be used with development indicators.

The development of algorithms to mitigate the spread of political bots is a solution considered by many researchers. This must be followed by laws and policies to regulate the use of social media and algorithms in politics. Despite the significant effort accomplished by some companies like Facebook or Twitter to reduce the negative effect of bots, they also have to consider their business interests. The intervention of states or civil society organisations is then useful to protect the citizen against manipulation caused by the use of bots and social media. This research will not include data from social media due to its low quality and the ethical and bias issues observed. When the WDI merged with the WGI creates a new dataset which will give a complete representation of a country from various perspectives. The new development data set, the WDGI will be used with the ACLED, providing disaggregated data on reported political violence and protest events.

8.2.2 Data bias in this research

The World Bank is a data supplier of development data. It provides detailed country data, informing about the population, the economy, the environment and the health situation. World Bank data is used as a primary source for several aggregated indexes, like the

Corruption perception Index or the Fragile States Index. The World Bank compiled cross-country comparable data on development called the World Development Indicators (WDI) (The World Bank, no date). The WDI is a compilation of comparable statistics that contains 1,600 time series indicators for 217 economies and above 40 country groups, with many indicators going back more than 50 years, and has been widely used to interpret and analyse development over the world. National agencies or companies collect the data, but most of the human data are collected via surveys. The data quality assessment of 19 political and socio-economic indicators from 217 countries identifies the quantity of the data that can be used for research. The data collected covers the period between 1960 and 2020. The country data analysis to identify countries with several empty data reveals that most countries with more than 50% of empty data are countries under high tension, or islands (Appendix E).

Of the 27 countries excluded from the analysis, 24 countries not encountering major crises, are micro countries with a high and upper-level income. For a population of fewer than 300,000 habitats, they have more than 50% of the total social, political and economic data collected between 2000 and 2019 is missing. The three low incomes countries are in a particular social and political context (South Sudan, Somalia and Korea, Dem. People's Rep.).

This research excluded from the analysis 27 countries, where it will not be possible to experiment with the designed risk assessment model. That need further investigation to find if an alternative source of information exists and if the bias caused by this data gap can be mitigated.

8.3 Alternatives for limited data and computation resources

8.3.1 Data calibration

This research required an alternative to fix the issue of missing or incomplete data for generating correlation centrality, and the proposed solution is a data calibration. Without the

calibrated country data used, it would not have been possible to get images that have some similarities when uncalibrated data gave distinct images representing the uniqueness of the data representation of each country. Calibrated data also provided an opportunity to get a high number of accurate images. The images are generated from the correlation of indicators regarding the crisis. Then, countries with missing values generated almost the same centrality representing a correlation equal to 1 or -1. The data calibration fixed that by using the data records to balance all country information. This approach significantly contributed to this research by giving accurate images considering the limited quantity of images this research initially produced.

8.3.2 Convolutional neural network

Many scientific and industrial applications rely on convolutional neural network (CNN)-based algorithms to discover patterns and recognise objects. These algorithms have garnered a significant deal of interest in image processing. Network selection and hyper-parameter optimisation are two of the most important issues for CNNs. There is a need to severely constrain the search space for the optimum architecture and hyper-parameters for CNNs because the techniques for automatically discovering the best architecture and hyper-parameters are computationally demanding. CNN parameter optimisation statistical method that can be used in various CNN applications and produces findings that are easier to understand (Akbarzadeh, Ahderom and Alameh, 2019).

The use of convolutional neural networks (CNNs) for image identification, including hyperspectral data classification and video classification, has become widespread. Large amounts of memory and processing power are required for most CNNs. As a result, the CNN's size is for applications with restricted resources. The research used primitive blocks from the well-known CNN architecture DCNN-11.

8.4 Future research: Danger at Sahel

It happens that young people from rural areas wind up travelling to cities searching for a job, only to find themselves unemployed. Studies undertaken in several Middle Eastern countries imply that the young age framework combined with the lack of economic opportunity could be more explosive in these countries. For this reason, researchers in the emerging topic of "demographic security" view an increasing population as a cause for concern. A country's likelihood of civil unrest, violence, as well as even extremism can be increased by a baby boom. The risk of young revolt is greatest when elected leaders fail to respond to dissatisfied people. Tunisia's 2010 revolution, which began the first Arab Spring uprising, had these ingredients (May, 2019). Sahelian countries are primarily agrarian. However, there are not enough agricultural jobs to go around with so many individuals in their 20s and 30s.

The rise of al-Qaida into Africa has coincided with the group's decline in the Middle East. In North Africa and the Sahel, terrorist organisations had an estimated several thousand fighters, many of whom have teamed up with Boko Haram. Niger, formerly tranquil, has lately been infiltrated by some of these organisations from Nigeria and Mali. Refugee camps within Niger's borders have also taken in people fleeing Boko Haram from those countries.

Many countries in the Sahel region of sub-Saharan Africa had seen their fertility rates fall at a very slow rate, compared to the dramatic reduction witnessed in the rest of the world. There are still 7.2 children born to the average Nigerien woman, as per the World Population Data Sheet 2018 from Population Reference Bureau. In underdeveloped countries, the average number of children a woman has is 2.6, terrorism is inevitable unless more is done to promote family planning and improve economic prospects.

8.5 Conclusion

The data used in the research are global data collected from the primary sources of information of global data that gives timely data on several countries from economic, socio-political and governance perspectives.

This research in addition to providing additional knowledge about the correlation between development indicators and social crisis data, proposed a method to calculate the risk of changes in indicators that can harm society.

This research enhanced knowledge in developing quantitative risk assessment for crisis prevention with development indicators. A risk assessment that integrates multi-country risk assessment could help assess the risk for a specific region, without assessing each country separately.

This research proposed a novel analysis method that will consider indicators connected with their correlation instead of taking development indicators as a separate indicator. The indicators, connected with their correlations are transformed into images. These images are then compared with a non-conflict image fingerprint for comparison; to find with network convolutional indicators varying and assess the risk linked to the proximity with a "crisis fingerprint".

This research experimented with a new data approach to classify networks based on their patterns or differences. This data approach will compare networks with graph convolution, networks generated from the centrality measurement of the weighted correlation information network. The results of the experimentation of this data approach on the sample case of Niger, Mali and Burkina Faso helped to identify key development factors that can tremendously reduce the risk of getting deadly clashes in the country. Poverty seems to be the root of all deadly crises, followed by political (in)stability, financial support to the private sector and a growing population. This new approach will be further tested with more recent

data on the Caribbean countries because it is the region where most of the countries with data missing were identified. This future research will retain the maximum quantitative and qualitative data from the data assessment that could allow investigating similarities and differences among Caribbean countries and compared to other countries from other regions.

A guide developed for the World Bank's Identification for Development (ID4D) initiative gives the pillars and principles that the World Bank recommends for a country developing an identification system (The World Bank, 2019). In this guide, the World Bank recommends the pillars: inclusion, design and governance (Table 8.5-1).

Pillar 1: Inclusion	*Principle 1. Ensuring universal coverage for individuals from birth to death, free from discrimination.*
	Principle 2. Removing barriers to access and usage and disparities in the availability of information and technology
Pillar 2: Design	*Principle 3. Establishing a robust—unique, secure, and accurate—identity.*
	Principle 4. Creating a platform that is interoperable and responsive to the needs of various users.
	Principle 5. Using open standards and ensuring vendor and technology neutrality.
	Principle 6. Protecting user privacy and control through system design.
	Principle 7. Planning for financial and operational sustainability without compromising accessibility
Pillar 3: Governance	*Principle 8. Safeguarding data privacy, security, and user rights through a comprehensive legal and regulatory framework.*
	Principle 9. Establishing clear institutional mandates and accountability.
	Principle 10. Enforcing legal and trust frameworks through independent oversight and adjudication of grievances

Table 8.5-1: World Bank's pillars on Identification for Development

The aspect of popular disturbances that can result from such initiative has not been taken on board in the design of this guidance. Turbulence can arise even for a necessary initiative like civil registration. An example is a revolt against the Government of Brazil after they announced a decree in 1852 making mandatory civil registration of birth and deaths, which caused a massacre (Loveman, 2007). For the Brazilian government, civil registration was innovative and would bring modernity to the country, when for a certain amount of the

population, it was perceived as an arbitrary and suspect measure. Taking human lives as a pillar would have imposed a different approach to bringing modernity without causing a revolt.

"Modernization projects that appeared perfectly reasonable on paper regularly hit up against the realities of local conditions and popular understandings of how things ought to be." (Loveman, 2007)

With the Covid pandemic, The World Bank in supporting countries listed a new set of pillars associating life with development (Worldbank, 2022) :

- saving lives
- protecting the poor and most vulnerable
- ensuring sustainable business growth and job creation
- and strengthening policies, institutions, and investments for rebuilding better

This research proposes the adoption of "Saving lives and protecting human capitals" as the first pillar of every development initiative and its integration into quantitative risk assessment as a variable that can influence a policy, similar to financial variables.

References

ACLED Project (2019) *Terms of Use and Attribution Policy*. Available at: https://www.acleddata.com. (Accessed: 6 May 2019).

Akbarzadeh, S., Ahderom, S. and Alameh, K. (2019) 'A statistical approach to provide explainable convolutional neural network parameter optimization', *International Journal of Computational Intelligence Systems*, 12(2), pp. 1635–1648. doi: 10.2991/IJCIS.D.191219.001.

Al-Khouri, R. and Khalik, M. U. A. (2013) 'Does Political Risk Affect the Flow of Foreign Direct Investment Into the Middle East North African Region?', *Journal of Global Business & Technology*, 9(2), pp. 47–59. Available at: http://search.ebscohost.com/login.aspx?direct=true&db=egs&AN=92620570&lang=es&site=ehost-live.

Albert, R. and Barabási, A. L. (2002) 'Statistical mechanics of complex networks', *Reviews of Modern Physics*, 74(1), pp. 47–97. doi: 10.1103/RevModPhys.74.47.

Aljarallah, R. A. and Angus, A. (2020) 'Dilemma of Natural Resource Abundance: A Case Study of Kuwait', *SAGE Open*, 10(1), p. 2158244019899701. doi: 10.1177/2158244019899701.

Alpar, P. and Winkelsträter, S. (2014) 'Assessment of data quality in accounting data with association rules', *Expert Systems with Applications*, 41(5), pp. 2259–2268. doi: 10.1016/j.eswa.2013.09.024.

Althoff, T., Jindal, P. and Leskovec, J. (2017) 'Online actions with offline impact: How online social networks influence online and offline user behavior', in *WSDM 2017 - Proceedings of the 10th ACM International Conference on Web Search and Data Mining*. New York, NY, USA: Association for Computing Machinery, Inc, pp. 537–546. doi: 10.1145/3018661.3018672.

Amanda, R. (2017) *Understanding bot abilities—and limitations | Zendesk Blog*. Available at: https://www.zendesk.com/blog/understanding-bot-abilities-limitations/ (Accessed: 28 December 2018).

Amidi, A. and Amidi, S. (2020) *Convolutional Neural Networks cheatsheet*, *CS 230 - Deep Learning*. Available at: https://stanford.edu/~shervine/teaching/cs-230/cheatsheet-convolutional-neural-networks (Accessed: 10 December 2020).

Avramovic, D. (1964) *Economic growth and external debt*. Baltimore, Md. : Hopkins. Available at: https://www.econbiz.de/Record/economic-growth-and-external-debt-avramovic-dragoslav/10000591827 (Accessed: 24 November 2022).

Balcilar, M. *et al.* (2020) 'Bridging the Gap Between Spectral and Spatial Domains in Graph Neural Networks'. Available at: http://arxiv.org/abs/2003.11702.

Bao, C., Li, J. and Wu, D. (2018) 'A fuzzy mapping framework for risk aggregation based on risk matrices', *Journal of Risk Research*, 21(5), pp. 539–561. doi: 10.1080/13669877.2016.1223161.

Batini, C. *et al.* (2009) 'Methodologies for data quality assessment and improvement', *ACM Computing Surveys*, 41(3), pp. 1–52. doi: 10.1145/1541880.1541883.

Batushansky, A., Toubiana, D. and Fait, A. (2016) 'Correlation-Based Network Generation, Visualization, and Analysis as a Powerful Tool in Biological Studies: A Case Study in Cancer Cell Metabolism', *BioMed Research International*, 2016. doi: 10.1155/2016/8313272.

Benuwa, B. B. *et al.* (2016) 'A review of deep machine learning', *International Journal of Engineering Research in Africa*. Trans Tech Publications Ltd, pp. 124–136. doi: 10.4028/www.scientific.net/JERA.24.124.

Berger-Schmitt, R. (2002) 'Considering Social Cohesion in Quality of Life Assessments: Concept and Measurement', *Social Indicators Research*, 58(1/3), pp. 403–428.

doi: 10.1111/j.1541-0072.1977.tb01338.x.

Brownlee, J. (2019) *Convolutional Neural Network Model Innovations for Image Classification*. Available at: https://machinelearningmastery.com/review-of-architectural-innovations-for-convolutional-neural-networks-for-image-classification/ (Accessed: 8 February 2022).

Buduma, N. and Locascio, N. (2017) *Fundamentals of Deep Learning_ Designing Next-Generation Machine Intelligence Algorithms*. First edit, *Nature*. First edit. Available at: https://en.wikipedia.org/wiki/Word_embedding%5Cnhttp://files.sig2d.org/sig2d14.pdf#page=5 (Accessed: 26 January 2022).

Cambridge Academic Content Dictionary (2020) *Governance | meaning in the Cambridge English Dictionary*, *Cambridge University Press*. Available at: https://dictionary.cambridge.org/dictionary/english/governance (Accessed: 10 January 2021).

Cariolle, J., Goujon, M. and Guillaumont, P. (2016) 'Has Structural Economic Vulnerability Decreased in Least Developed Countries? Lessons Drawn from Retrospective Indices', *The Journal of Development Studies*, 52(5), pp. 591–606. doi: 10.1080/00220388.2015.1098631.

Castle, J. L., Clements, M. P. and Hendry, D. F. (2016) 'An Overview of Forecasting Facing Breaks', *Journal of Business Cycle Research*, 12(1), pp. 3–23. doi: 10.1007/s41549-016-0005-2.

Casula, M., Rangarajan, N. and Shields, P. (2021) 'The potential of working hypotheses for deductive exploratory research', *Quality and Quantity*, 55(5), pp. 1703–1725. doi: 10.1007/s11135-020-01072-9.

Chang, J. *et al.* (2017) 'Deep Adaptive Image Clustering', pp. 5879–5887.

Chapman, P. *et al.* (2000) *CRISP-DM 1.0 Step-by-step*, *ASHA presentation*. doi: 10.1109/ICETET.2008.239.

Chen, X. W. and Lin, X. (2014) ‘Big data deep learning: Challenges and perspectives’, *IEEE Access*, 2, pp. 514–525. doi: 10.1109/ACCESS.2014.2325029.

Chester, J. and Montgomery, K. C. (2017) ‘The role of digital marketing in political campaigns’, *Internet Policy Review*, 6(4), pp. 1–20. doi: 10.14763/2017.4.773.

Chojnacki, S. *et al.* (2012) ‘Event Data on Armed Conflict and Security: New Perspectives, Old Challenges, and Some Solutions’, *International Interactions*, 38(4), pp. 382–401. doi: 10.1080/03050629.2012.696981.

Christine, W. (2018) *This Is Exactly How Social Media Algorithms Work Today*. Available at: https://www.skyword.com/contentstandard/marketing/this-is-exactly-how-social-media-algorithms-work-today/ (Accessed: 26 December 2018).

Clarke, D. and Dercon, S. (2019) ‘Beyond banking: Crisis risk finance and development insurance in IDA19’, *International Development Association*. Available at: www.disasterprotection.org. (Accessed: 25 July 2022).

Cupal, M., Deev, O. and Linnertova, D. (2015) ‘The Poisson Regression Analysis for Occurrence of Floods’, *Procedia Economics and Finance*, 23, pp. 1499–1502. doi: 10.1016/s2212-5671(15)00465-7.

Demeke, Y. H. (2022) ‘Youth unemployment and political instability: evidence from IGAD member countries’. doi: 10.1080/23322039.2022.2079211.

Desta, A. (1985) ‘Assessing Political Risk in Less Developed Countries’, *Journal of Business Strategy*, 5(4), p. 40. doi: 10.1108/eb039086.

Donaubauer, J., Herzer, D. and Nunnenkamp, P. (2019) ‘The Effectiveness of Aid under Post-Conflict Conditions: A Sector-Specific Analysis’, *Journal of Development Studies*, 55(4), pp. 720–736. doi: 10.1080/00220388.2017.1400013.

Duan, Q. *et al.* (2015) ‘Accurate analysis and prediction of enterprise service-level performance’, *ACM Transactions on Design Automation of Electronic Systems*, 20(4), pp.

52:1--52:23. doi: 10.1145/2757279.

Eck, K. (2005) *A beginner's guide to conflict data. Finding and using the right dataset*, *UCDP Research Paper Series*. Available at: www.ucdp.uu.se (Accessed: 10 May 2019).

Eck, K. (2012) 'In data we trust? A comparison of UCDP GED and ACLED conflict events datasets', *Cooperation and Conflict*, 47(1), pp. 124–141. doi: 10.1177/0010836711434463.

Fagbemi, F. and Fajingbesi, A. (2022) 'Political violence: why conflicts can result from sub-Saharan African socioeconomic conditions', *Journal of Business and Socio-economic Development*. doi: 10.1108/jbsed-12-2021-0178.

Fearon, J. D. (2007) 'Economic development, insurgency, and civil war', *Institutions and economic performance*, pp. 292–328. Available at: http://books.google.com/books?hl=en&lr=&id=6lPiGQXSzZkC&oi=fnd&pg=PA292&dq=Economic+development,+insurgency,+and+civil+war+?&ots=XeHPKobKwL&sig=PYONG6NGc_egGORzm8UnrRQP6Xc.

Fragile States Index | The Fund for Peace (no date). Available at: https://fragilestatesindex.org/ (Accessed: 24 April 2022).

Gema Bueno de la Fuente (no date) *What is Open Science? Introduction | FOSTER*. Available at: https://www.fosteropenscience.eu/content/what-open-science-introduction (Accessed: 6 May 2019).

Goldsmith, A. A. (2001) 'Foreign aid and statehood in Africa', *International Organization*, 55(1), pp. 123–148. doi: 10.1162/002081801551432.

Gopaldas, R. and Menzi, N. (2021) *Rising food prices could ignite unrest and instability in Africa - ISS Africa*. Available at: https://issafrica.org/iss-today/rising-food-prices-could-ignite-unrest-and-instability-in-africa (Accessed: 28 June 2022).

Grace-Martin, K. (no date) *Regression Models for Count Data*, *The analysis factor*. Available at: https://www.theanalysisfactor.com/regression-models-for-count-data/ (Accessed: 24 April 2020).

Halpern, D., Valenzuela, S. and Katz, J. E. (2017) 'We Face, I Tweet: How Different Social Media Influence Political Participation through Collective and Internal Efficacy', *Journal of Computer-Mediated Communication*, 22(6), pp. 320–336. doi: 10.1111/jcc4.12198.

Haug, A., Zachariassen, F. and van Liempd, D. (2011) 'The costs of poor data quality', *Journal of Industrial Engineering and Management*, pp. 168–193. doi: 10.3926/jiem.2011.v4n2.p168-193.

Hegre, H. *et al.* (2019) 'ViEWS: A political violence early-warning system', *Journal of Peace Research*, 56(2), pp. 155–174. doi: 10.1177/0022343319823860.

Howard, P. N. *et al.* (2018) 'Algorithms , bots , and political communication in the US 2016 election : The challenge of automated political communication for election law and administration administration', *Journal of Information Technology & Politics*, 15(2), pp. 81–93. doi: 10.1080/19331681.2018.1448735.

Hu, J. *et al.* (2019) 'A Security Risk Plan Search Assistant Decision Algorithm Using Deep Neural Network Combined with Two-Stage Similarity Calculation', *Personal Ubiquitous Comput.*, 23(3–4), pp. 541–552. doi: 10.1007/s00779-019-01236-x.

Hunter, L. Y. and Biglaiser, G. (2020) 'The Effects of the International Monetary Fund on Domestic Terrorism', *Terrorism and Political Violence*, pp. 1–25. doi: 10.1080/09546553.2019.1709448.

Hwang, C.-L. and Yoon, K. (1981) 'Methods for Multiple Attribute Decision Making', in. doi: 10.1007/978-3-642-48318-9_3.

Jena, R. *et al.* (2020) 'Earthquake hazard and risk assessment using machine learning

approaches at Palu, Indonesia', *Science of The Total Environment*, 749, p. 141582. doi: https://doi.org/10.1016/j.scitotenv.2020.141582.

Jost, J. T. *et al.* (2018) 'How Social Media Facilitates Political Protest: Information, Motivation, and Social Networks', *Political Psychology*, 39(3), pp. 85–118. doi: 10.1111/pops.12478.

Katherine, L. (2018) *How to Document Chatbot Requirements – Chatbots Magazine*, *Chatbots Magazine*. Available at: https://chatbotsmagazine.com/how-to-document-chatbot-requirements-7df81275cc66 (Accessed: 28 December 2018).

Kaufmann, D., Kraay, A. and Mastruzzi, M. (2011) 'The worldwide governance indicators: Methodology and analytical issues', *Hague Journal on the Rule of Law*, 3(2), pp. 220–246. doi: 10.1017/S1876404511200046.

Khari Johnson (2018) *Facebook Messenger passes 300,000 bots | VentureBeat*. Available at: https://venturebeat.com/2018/05/01/facebook-messenger-passes-300000-bots/ (Accessed: 24 December 2018).

Kim, N. and Conceicao, P. (2010) 'The Economic Crisis, Violent Conflict, and Human Development', *International Journal of Peace Studies*, 15(1), pp. 29–43.

Kitsak, M. *et al.* (2010) 'Identification of influential spreaders in complex networks', *Nature Physics 2010 6:11*, 6(11), pp. 888–893. doi: 10.1038/nphys1746.

Kotsiantis, S. B., Zaharakis, I. D. and Pintelas, P. E. (2006) 'Machine learning: A review of classification and combining techniques', *Artificial Intelligence Review*, 26(3), pp. 159–190. doi: 10.1007/s10462-007-9052-3.

Kumaraswamy, M. and Ramaswamy, R. (2016) 'Performance Evaluation of Software Projects using Criteria Importance through Inter-criteria Correlation Technique', 6(3), pp. 28–36.

Langfelder, P. and Horvath, S. (2008) 'WGCNA : an R package for weighted

correlation network analysis'. doi: 10.1186/1471-2105-9-559.

Lees-Marshment, J. (2001) 'The marriage of politics and marketing', *Political Studies*. doi: 10.1111/1467-9248.00337.

Liu, Y. *et al.* (2018) 'Fragile states metric system: An assessment model considering climate change', *Sustainability (Switzerland)*, 10(6). doi: 10.3390/su10061767.

LIU, Z. (2018) *Interlinked nature of the Sustainable Development Goals — SDG Indicators*, *The Sustainable Development Goals Report*. Available at: https://unstats.un.org/sdgs/report/2018/interlinkages/ (Accessed: 1 September 2020).

Loshin, D. (2010) 'Evaluating the Business Impacts of Poor Data Quality', *Software Engineering Institute*, (301), pp. 1–10. Available at: www.knowledge-integrity.com (Accessed: 7 February 2021).

Loveman, M. (2007) 'Blinded like a state: The revolt against civil registration in nineteenth-century Brazil', *Comparative Studies in Society and History*, 49(1), pp. 5–39. doi: 10.1017/S0010417507000394.

Marsland, S. (2014) *Machine learning: An algorithmic perspective*, *Machine Learning: An Algorithmic Perspective, Second Edition*. CRC Press. doi: 10.1201/b17476.

May, J. F. (2019) *Niger has the world's highest birth rate – and that may be a recipe for unrest*. Available at: https://theconversation.com/niger-has-the-worlds-highest-birth-rate-and-that-may-be-a-recipe-for-unrest-108654 (Accessed: 8 February 2022).

MBFC News (no date) *Daily Star UK - Media Bias/Fact Check*. Available at: https://mediabiasfactcheck.com/daily-star-uk/ (Accessed: 6 May 2019).

Mohamad, D., Liyana, N. and Halim, A. (2020) 'Possibility Based TOPSIS with Inter Criteria Correlation and Similarity Measure', 42(1), pp. 22–33.

Moksony, F. and Hegedűs, R. (2014) 'The use of Poisson regression in the sociological study of suicide', *Corvinus Journal of Sociology and Social Policy*, 5(2), pp. 97–

114. doi: 10.14267/cjssp.2014.02.04.

Mooij, J. M. *et al.* (2016) 'Distinguishing cause from effect using observational data: Methods and benchmarks', *Journal of Machine Learning Research*, 17.

Muhamedyev, R. I. (2015) 'Machine Learning Methods : an overview', *Computer Modelling & New Technologies*, 19(6), pp. 14–29.

Müller, K. and Schwarz, C. (2018) 'Social Media and Hate Crime Centre for Competitive Advantage in the Global Economy Fanning the Flames of Hate : Social Media and Hate Crime *', *Warwick Working Paper Series*, (373).

Nagle, G. and Guinness, P. (2014) 'Cambridge IGCSE Geography 2nd Edition.pdf', p. 19. Available at: https://books.google.de/books?hl=en&lr=&id=FVyHBAAAQBAJ&oi=fnd&pg=PT6&dq=Cambridge+IGCSE+Geography+2nd+Edition&ots=p2HM8Yppie&sig=Yvutta8dO_ngB9XN1yoBmsQjZ40&redir_esc=y#v=onepage&q=Cambridge IGCSE Geography 2nd Edition&f=false (Accessed: 12 July 2022).

Nardulli, P. F., Althaus, S. L. and Hayes, M. (2015) 'A Progressive Supervised-learning Approach to Generating Rich Civil Strife Data', *Sociological Methodology*, 45(1), pp. 148–183. doi: 10.1177/0081175015581378.

Negre, C. F. A. *et al.* (2018) 'Eigenvector centrality for characterization of protein allosteric pathways', 115(52). doi: 10.1073/pnas.1810452115.

Nelder, J. A., Chatterjee, S. and Price, B. (1979) 'Regression Analysis by Example.', *Biometrics*. doi: 10.2307/2529957.

Newman, M. E. J. (2003) 'The structure and function of complex networks', *SIAM Review*, pp. 167–256. doi: 10.1137/S003614450342480.

Newman, M. E. J. (2010) *Networks: An Introduction*, *Networks: An Introduction*. doi: 10.1093/acprof:oso/9780199206650.001.0001.

Otte, E. and Rousseau, R. (2002) ‘Social network analysis: a powerful strategy, also for the information sciences’, *Journal of Information Science*, 28(6), pp. 441–453. doi: 10.1177/016555150202800601.

Palczewski, K. and Sałabun, W. (2019) ‘The fuzzy TOPSIS applications in the last decade’, *Procedia Computer Science*, 159, pp. 2294–2303. doi: 10.1016/J.PROCS.2019.09.404.

Pew Research Center (no date) *Pew Global Research Center Global Attitudes and Trends*. Available at: https://www.pewresearch.org/global/datasets/ (Accessed: 20 December 2019).

Plonsky, L. and Ghanbar, H. (2018) ‘Multiple Regression in L2 Research: A Methodological Synthesis and Guide to Interpreting R2 Values’, *Modern Language Journal*, 102(4), pp. 713–731. doi: 10.1111/modl.12509.

Potter, R. *et al.* (2012) *Key concepts in development geography*, *Key Concepts in Development Geography*. doi: 10.4135/9781473914834.

Puurunen, A., Majava, J. and Kess, P. (2014) ‘Exploring incomplete information in maintenance materials inventory optimization’, *Industrial Management and Data Systems*, 114(1), pp. 144–158. doi: 10.1108/IMDS-01-2013-0025.

R: What is R? (no date). Available at: https://www.r-project.org/about.html (Accessed: 29 April 2019).

Raleigh, C. *et al.* (2010) ‘Introducing ACLED: An Armed Conflict Location and Event Dataset: Special Data Feature’, *Journal of Peace Research*, 47(5), pp. 651–660. doi: 10.1177/0022343310378914.

Ravindra, K. (2018) *Conversational AI chat-bot — Architecture overview – Towards Data Science*. Available at: https://towardsdatascience.com/architecture-overview-of-a-conversational-ai-chat-bot-4ef3dfefd52e (Accessed: 25 December 2018).

Rodney, R. (2006) ‘Regressions: Why Are Economists Obessessed with Them?’, *FINANCE and DEVELOPMENT*, March, p. Volume 43, Number 1. Available at: https://www.imf.org/external/pubs/ft/fandd/2006/03/basics.htm.

Rosebrock, A. (2017) ‘Deep Learning for Computer Vision with Python - Starter Bundle’, *Pyimagesearch*, p. 330.

Ryan, A. (2006) ‘Post-Positivist Approaches to Research’, *Researching and Writing your thesis: a guide for postgraduate students*, pp. 12–26.

Savun, B. and Tirone, D. C. (2012) ‘Exogenous shocks, foreign aid, and civil war’, *International Organization*, 66(3), pp. 363–393. doi: 10.1017/S0020818312000136.

Sébastien, F. (2017) *The Ultimate Guide To Designing A Chatbot Tech Stack*. Available at: https://chatbotsmagazine.com/the-ultimate-guide-to-designing-a-chatbot-tech-stack-333eceb431da (Accessed: 26 December 2018).

Shao, C. *et al.* (2018) ‘The spread of low-credibility content by social bots’, *Nature communications*, 9(1), p. 4787. doi: 10.1038/s41467-018-06930-7.

Shu, K. *et al.* (2017) ‘Fake News Detection on Social Media: A Data Mining Perspective’, (i). doi: 10.1016/j.tvjl.2007.06.012.

Shyti, B. and Valera, D. (2018) ‘The Regression Model for the Statistical Analysis of Albanian Economy’, *International Journal of Mathematics Trends and Technology*, 62(2), pp. 90–96. doi: 10.14445/22315373/ijmtt-v62p513.

Silvola, R. *et al.* (2016) ‘Data quality assessment and improvement’, *International Journal of Business Information Systems*, 22(1), pp. 62–81. doi: 10.1504/IJBIS.2016.075718.

Szegedy, C. *et al.* (2014) ‘Going Deeper with Convolutions’, *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 07-12-June-2015, pp. 1–9. doi: 10.1109/CVPR.2015.7298594.

Talend (2021) ‘DPA (v2021.07.06) 2 Data Protection Terms and Definitions’.

Tebaldi, E. and Alda, E. (2017) 'Quality of Institutions and Violence Incidence: a Cross-Country Analysis', *Atlantic Economic Journal*, 45(3), pp. 365–384. doi: http://dx.doi.org/10.1007/s11293-017-9547-5.

The Fund for Peace (2014) 'CAST Conflict Assessment Framework Manual', p. 28.

The World Bank (2018) *Where do the world's poorest people live today?* Available at: http://datatopics.worldbank.org/world-development-indicators/stories/where-do-the-worlds-poorest-people-live-today.html.

The World Bank (2019) 'ID4D Practitioner's Guide (English)'.

The World Bank (2022) *World Bank Country and Lending Groups – World Bank Data Help Desk*. Available at: https://datahelpdesk.worldbank.org/knowledgebase/articles/906519-world-bank-country-and-lending-groups (Accessed: 28 November 2022).

The World Bank (no date) *World Development Indicators*. Available at: http://datatopics.worldbank.org/world-development-indicators/ (Accessed: 29 November 2019).

Torres, M. and Cantú, F. (2022) 'Learning to See: Convolutional Neural Networks for the Analysis of Social Science Data', *Political Analysis*, 30(1), pp. 113–131. doi: 10.1017/pan.2021.9.

Twitter Public Policy (2018) *Update on Twitter's review of the 2016 US election*. Available at: https://blog.twitter.com/official/en_us/topics/company/2018/2016-election-update.html (Accessed: 27 December 2018).

Tzanis, G., Katakis, I. and Vlahavas, I. (2006) 'Modern Applications of Machine Learning', (May 2014).

UN-OHRLLS (2020) *About the Least Developed Countries*.

UNDP (2015) *What is Human Development? | Human Development Reports*. Available at: https://hdr.undp.org/content/what-human-development (Accessed: 12 July

2022).

UNDP (no date) *Human Development Index | Human Development Reports*. Available at: https://hdr.undp.org/data-center/human-development-index#/indicies/HDI (Accessed: 12 July 2022).

United States Army Corps of Engineers (2018) 'Corps Risk Analysis Gateway Training Module: Risk Assessment - Quantitative Methods', p. 51.

Uusitalo, O. (2014) 'Research methodology', *SpringerBriefs in Applied Sciences and Technology*, (9783319068282), pp. 25–39. doi: 10.1007/978-3-319-06829-9_3.

Uyanık, G. K. and Güler, N. (2013) 'A Study on Multiple Linear Regression Analysis', *Procedia - Social and Behavioral Sciences*, 106, pp. 234–240. doi: 10.1016/j.sbspro.2013.12.027.

Varol, O. *et al.* (2017) 'Online Human-Bot Interactions: Detection, Estimation, and Characterization'. doi: 10.1016/S0009-9260(05)80025-X.

Vijaya, R. M. *et al.* (2018) 'Economic underpinnings of violent extremism: A cross country exploration of repeated survey data', *World Development*, 109, pp. 401–412. doi: 10.1016/J.WORLDDEV.2018.05.009.

Wang, D., Hu, G. and Lyu, C. (2020) 'Spatial-Related Correlation Network for 3D Point Clouds', *IEEE Access*, 8, pp. 116004–116012. doi: 10.1109/ACCESS.2020.3004472.

Wang, H. *et al.* (2020) 'Landslide identification using machine learning', *Geoscience Frontiers*. doi: https://doi.org/10.1016/j.gsf.2020.02.012.

van Weezel, S. (2019) 'Local warming and violent armed conflict in Africa', *World Development*, 126, p. 104708. doi: 10.1016/j.worlddev.2019.104708.

Wellman, B. and Berkowitz, S. D. (1988) 'Introduction: Studying social structures', in *Social Structures: A Network Approach*.

What is Open Data? (no date). Available at:

http://opendatahandbook.org/guide/en/what-is-open-data/ (Accessed: 6 May 2019).

Worldbank (2022) ‘Helping Countries Adapt to a Changing World’, p. 116. Available at: www.worldbank.org/annualreport.

Yamashita, R. *et al.* (2018) ‘Convolutional neural networks: an overview and application in radiology’, *Insights into Imaging*, 9(4), pp. 611–629. doi: 10.1007/s13244-018-0639-9.

Yang, P., Liu, X. and Xu, G. (2018) ‘A dynamic weighted TOPSIS method for identifying influential nodes in complex networks’, *https://doi.org/10.1142/S0217984918502160*, 32(19). doi: 10.1142/S0217984918502160.

Zadeh, L. A. (1965) ‘Fuzzy sets’, *Information and Control*, 8(3), pp. 338–353. doi: 10.1016/S0019-9958(65)90241-X.

Zhang, S. *et al.* (2019a) ‘Graph convolutional networks : a comprehensive review’, *Computational Social Networks*. doi: 10.1186/s40649-019-0069-y.

Zhang, S. *et al.* (2019b) ‘Graph convolutional networks: a comprehensive review’, *Computational Social Networks*, 6(1), pp. 1–23. doi: 10.1186/s40649-019-0069-y.

Appendices

Appendix A: Results of the systematic literature review

Motivation	Search terms	Comments	Threads
Question 1**: What is the selection of development indicators that might conduct to socio-political crises?**	**SAGE: All journals** **Query**: [All combin*] AND [All "development indicators"] AND [[All fragil*] OR [All uncertain*] OR [All delicat*]] AND [All social* OR politic*] AND [[All cohesion] OR [All cooperat*] OR [All participat*]]		18
	Journal: ***Taylor & Francis Journals Complete*** **Query**: [All: combin*] AND [All: "development indicators"] AND [[All: fragil*] OR [All: uncertain*] OR [All: delicat*]] AND [All: social*] AND [[All: cohesion] OR [All: cooperat*] OR [All: participat*]] AND [Publication Date: (01/01/2015 TO 12/31/2020)]	This search was refined to get literature published in The Journal of Development Studies, and Terrorism and Political Violence.	24
	World Bank Open Knowledge Repository "development indicators" AND (fragil* OR uncertain* OR delicat*) AND (association* OR Correlation* OR match*) AND (social* OR politic*)	No refining	1
Question 2**: What are the patterns in the correlation network of development indicators?**	**Journal:** ***EBSCOhost*** **Query:** pattern AND correlation AND (network OR graph) AND (social or	Searching into all databases, this request	32

	economic or political) AND indicators	. The search with development indicators gave 2 results, both are not relevant for the question. Replacing the term "development" social OR economic OR political, gave more results. Results selected from peer-reviewed articles published between 2015 and 2020, in the database Academic Search Complete, Business Source Complete, and GreenFILE.	
	Journal: Proquest-*ABI/INFORM Collection* **Query:** pattern AND correlation AND ("network theory" OR "graph theory") AND ("development indicators")	This request was refined to include peer reviewed and published since 2015. The search with "development indicators" and "network theory" OR "graph theory"gave precise results	25
	Journal: *Wiley Online Library*	To reduce the results, the search	26

Question 3**: Are the risks of a brutal social crisis assessable with graph convolution?**	**Request:** "risk+assessment+OR+risk+analysis" anywhere and "violen* OR fatal* OR harmful OR dangerous OR brutal*" anywhere and "crisis OR conflict*" anywhere and "graph convolution" anywhere and "NOT biolog*" anywhere	excluded all articles related to biology. The results came from journals published between 2015 and 2020	
	Journal: *IEEE Xplore* **Request:** (("All Metadata":graph convolution) AND "All Metadata":risk)	The search of risk assessment terms related to graph convolution gave no results. The search of risk assessment in the database gave many results applied to several sectors. The search was the reduced to Risk and Graph convolution.	6
	ACM digital library **Request**: [Abstract: risk+assessment] AND [Abstract: graph+convolution] AND NOT [All: biolog*] AND [Publication Date: (01/01/2015 TO *)]	The ACM journal does not give results for social science and WDI. Instead, it gives results for risk assessment and graph convolution, they were search in the abstract. Articles Journal published since 2015	16
	Science Direct **Request**: Year: 2015-2020	The request search risk and convolution in the abstract.	25

	Title, abstract, keywords: (convolution AND ("risk assessment" OR "risk analysis"))		

Appendix B: Preliminary reviewed articles

Author	Title	Journal	Year	Relevance	Interest
Martín, C. J. & Carnero, M. C.	Evaluation of Sustainable Development in European Union Countries	Applied Sciences	2019	Medium	Patterns in the correlation network
Moaniba, Igam M.; Su, Hsin-Ning; Pei-Chun, Lee	Does reverse causality explains the relationship between economic performance and technological diversity?	Technological and Economic Development of Economy	2018	Medium	Patterns in the correlation network
Feng, Fuli; He, Xiangnan; Wang, Xiang; Luo, Cheng; Liu, Yiqun; Chua, Tat-Seng	Temporal Relational Ranking for Stock Prediction	ACM Trans. Inf. Syst.	2019	High	Graph convolution
Green, Ben; Chen, Yiling	The Principles and Limits of Algorithm-in-the-Loop Decision Making	Proc. ACM Hum.-Comput. Interact.	2019	Medium	Graph convolution
Gritzalis, Dimitris; Iseppi, Giulia; Mylonas, Alexios;	Exiting the Risk Assessment Maze: A Meta-Survey	ACM Comput. Surv.	2018	High	Graph convolution

Stavrou, Vasilis					
Hu, Jun; Fang, Jun; Du, Yanhua; Liu, Zhe; Ji, Pengyang	A Security Risk Plan Search Assistant Decision Algorithm Using Deep Neural Network Combined with Two-Stage Similarity Calculation	Personal Ubiquitous Comput.	2019	High	Graph convolution
Li, Zhaoju; Zhou, Zongwei; Jiang, Nan; Han, Zhenjun; Xing, Junliang; Jiao, Jianbin	Spatial Preserved Graph Convolution Networks for Person Re-Identification	ACM Trans. Multimedia Comput. Commun. Appl.	2020	Medium	Graph convolution
Soler-Dominguez, Amparo; Juan, Angel A.; Kizys, Renatas	A survey on financial applications of metaheuristics	ACM Computing Surveys	2017	Medium	Graph convolution
Zhang, Yuan; Sun, Fei; Yang, Xiaoyong; Xu, Chen; Ou, Wenwu; Zhang, Yan	Graph-Based Regularization on Embedding Layers for Recommendation	ACM Trans. Inf. Syst.	2020	Medium	Graph convolution
????????, ???????? ???????????	ЭКОНОМИКО-СТАТИСТИЧЕСКИЙ АНАЛИЗ ГАЗОРАСПРЕДЕЛИТЕЛЬН	Financial Analytics	2017	Medium	Patterns in the correlation network

	ОГО КОМПЛЕКСА РОССИИ				
Andrews, Naomi C. Z.	Prestigious Youth are Leaders but Central Youth are Powerful: What Social Network Position Tells us About Peer Relationships.	Journal of Youth {\&} Adolescence	2020	Medium	Patterns in the correlation network
Baydilli, Yusuf Yargi; T{\{u}}rker	Ä°lker"	Is the world small enough? â€" A view from currencies.	2019	Medium	Patterns in the correlation network
Ning, Xiaodong; Yac, Lina; Wang, Xianzhi; Benatallah, Boualem; Dong, Manqing; Zhang, Shuai	Rating prediction via generative convolutional neural networks based regression.	Pattern Recognition Letters	2020	Medium	Patterns in the correlation network
Pandove, Divya; Goel, Shivani; Rani, Rinkle	General correlation coefficient based agglomerative clustering.	Cluster Computing	2019	High	Patterns in the correlation network
Xu, Zhixi; Wu, Shufan; Huang, Xinran; Cai, Zhongliang; Su, Shiliang; Weng, Min;	Identifying the Geographic Indicators of Poverty Using Geographically Weighted Regression: A Case Study from Qiandongnan Miao and Dong Autonomous Prefecture, Guizhou, China.	Social Indicators Research	2019	Medium	Patterns in the correlation network

Liu, Ji; Sun, Junying					
Gama, F.; Marques, A. G.; Ribeiro, A.; Leus, G.	MIMO Graph Filters for Convolutional Neural Networks		2018	High	Graph convolution
Lv, L.; Cheng, J.; Peng, N.; Fan, M.; Zhao, D.; Zhang, J.	Auto-encoder based Graph Convolutional Networks for Online Financial Anti-fraud		2019	High	Graph convolution
Parsa, B.; Narayanan, A.; Dariush, B.	Spatio-Temporal Pyramid Graph Convolutions for Human Action Recognition and Postural Assessment		2020	High	Graph convolution
Sun, J.; Zhang, J.; Li, Q.; Yi, X.; Liang, Y.; Zheng, Y.	Predicting Citywide Crowd Flows in Irregular Regions Using Multi-View Graph Convolutional Networks	IEEE Transactions on Knowledge and Data Engineering	2020	Medium	Graph convolution
Zhang, J.; Gong, J.; Barnes, L.	HCNN: Heterogeneous Convolutional Neural Networks for Comorbid Risk Prediction with Electronic Health Records		2017	High	Graph convolution
Zheng, J.; Zhou, X.; Riga, C.; Yang, G.	Real-Time 3-D Shape Instantiation for Partially Deployed Stent Segments From a Single 2-D Fluoroscopic Image in Fenestrated Endovascular Aortic Repair	IEEE Robotics and Automation Letters	2019	High	Graph convolution
Aljarallah, Ruba A.;	Dilemma of Natural Resource Abundance: A Case Study of Kuwait	SAGE Open	2020	High	Development indicators

Angus, Andrew					
Chen, Hao; Hongo, Duncan O.; Ssali, Max William; Nyaranga, Maurice Simiyu; Nderitu, Consolata Wairimu	The Asymmetric Influence of Financial Development on Economic Growth in Kenya: Evidence From NARDL	SAGE Open	2020	Medium	Development indicators
Estrin, Saul; Korosteleva, Julia; Mickiewicz, Tomasz	Schumpeterian Entry: Innovation, Exporting, and Growth Aspirations of Entrepreneurs	Entrepreneurship Theory and Practice	2020	Medium	Development indicators
Hegre, H{\aa}vard; Allansson, Marie; Basedau, Matthias; Colaresi, Michael; Croicu, Mihai; Fjelde, Hanne; Hoyles, Frederick; Hultman,	ViEWS: A political violence early-warning system	Journal of Peace Research	2019	Medium	Development indicators

Lisa; H{\{o}}gbladh					
Khalid, Usman; Okafor, Luke Emeka; Shafiullah, Muhammad	The Effects of Economic and Financial Crises on International Tourist Flows: A Cross-Country Analysis	Journal of Travel Research	2019	Medium	Development indicators
Chen, Jiayao; Yang, Tongjun; Zhang, Dongming; Huang, Hongwei; Tian, Yu	Deep learning based classification of rock structure of tunnel face	Geoscience Frontiers	2020	Medium	Graph convolution
He, Rui; Li, Xinhong; Chen, Guoming; Chen, Guoxing; Liu, Yiwei	Generative adversarial network-based semi-supervised learning for real-time risk warning of process industries	Expert Systems with Applications	2020	Medium	Graph convolution
Jena, Ratiranjan; Pradhan, Biswajeet; Beydoun, Ghassan; Alamri, Abdullah M.; Ardiansyah;	Earthquake hazard and risk assessment using machine learning approaches at Palu, Indonesia	Science of The Total Environment	2020	High	Graph convolution

Nizamuddin; Sofyan, Hizir					
Kwag, Shinyoung; Gupta, Abhinav	Probabilistic risk assessment framework for structural systems under multiple hazards using Bayesian statistics	Nuclear Engineering and Design	2017	High	Graph convolution
Liu, Mingliang; Grana, Dario	Petrophysical characterization of deep saline aquifers for CO2 storage using ensemble smoother and deep convolutional autoencoder	Advances in Water Resources	2020	Medium	Graph convolution
Wang, Fan; Yang, Jing-Fang; Wang, Meng-Yao; Jia, Chen-Yang; Shi, Xing-Xing; Hao, Ge-Fei; Yang, Guang-Fu	Graph attention convolutional neural network model for chemical poisoning of honey bees' prediction	Science Bulletin	2020	Medium	Graph convolution
Wang, Haojie; Zhang, Limin; Yin, Kesheng; Luo, Hongyu; Li, Jinhui	Landslide identification using machine learning	Geoscience Frontiers	2020	High	Graph convolution
Yu, Rongjie; Wang, Yiyun; Zou, Zihang;	Convolutional neural networks with refined loss functions for the real-time crash risk analysis	Transportation Research Part C: Emerging Technologies	2020	Medium	Graph convolution

Wang, Liqiang					
Joël Cariolle, Michaël Goujon & Patrick Guillaumont	Has Structural Economic Vulnerability Decreased in Least Developed Countries? Lessons Drawn from Retrospective Indices		2016	Medium	Development indicators
Donaubauer, Julian; Herzer, Dierk; Nunnenkamp, Peter	The Effectiveness of Aid under Post-Conflict Conditions: A Sector-Specific Analysis	Journal of Development Studies	2019	Medium	Development indicators
Hunter, Lance Y.; Biglaiser, Glen	The Effects of the International Monetary Fund on Domestic Terrorism	Terrorism and Political Violence	2020	Medium	Development indicators
Gates, Scott; Hegre, H{\aa}vard; Nyg{\aa}rd, H{\aa}vard Mokleiv; Strand, H{\aa}vard	Consequences of Civil Conflict		2010	High	Development indicators
Lu, Yafeng; Garcia, Rolando; Hansen, Brett; Gleicher, Michael;	The State-of-the-Art in Predictive Visual Analytics	Computer Graphics Forum	2017	Medium	Graph convolution

Maciejewski, Ross					
Luo, Xiaochun; Li, Heng; Yang, Xincong; Yu, Yantao; Cao, Dongping	Capturing and Understanding Workers' Activities in Far-Field Surveillance Videos with Deep Action Recognition and Bayesian Nonparametric Learning	Computer-Aided Civil and Infrastructure Engineering	2019	Medium	Graph convolution
Modarres, Ceena; Astorga, Nicolas; Droguett, Enrique Lopez; Meruane, Viviana	Convolutional neural networks for automated damage recognition and damage type identification	Structural Control and Health Monitoring	2018	High	Graph convolution
Nayyeri, Fereshteh; Hou, Lei; Zhou, Jun; Guan, Hong	Foregroundâ€"background separation technique for crack detection	Computer-Aided Civil and Infrastructure Engineering	2019	Medium	Graph convolution

Appendix C: List of World Development Indicators used in this research

Indicator Name	Code	Long definition
Domestic credit to private sector by banks (% of GDP)	FD.AST.PRVT.GD.ZS	Domestic credit to private sector by banks refers to financial resources provided to the private sector by other depository corporations (deposit taking corporations except central banks), such as through loans, purchases of nonequity securities, and trade credits and other accounts receivable, that establish a claim for repayment. For some countries these claims include credit to public enterprises.
Electric power consumption (kWh per capita)	EG.USE.ELEC.KH.PC	Electric power consumption measures the production of power plants and combined heat and power plants less transmission, distribution, and transformation losses and own use by heat and power plants.
Oil rents (% of GDP)	NY.GDP.PETR.RT.ZS	Oil rents are the difference between the value of crude oil production at world prices and total costs of production.
GDP per capita growth (annual %)	NY.GDP.PCAP.KD.ZG	Annual percentage growth rate of GDP per capita based on constant local currency. Aggregates are based on constant 2010 U.S. dollars. GDP per capita is gross domestic product divided by midyear population. GDP at purchaser's prices is the sum of gross value added by all resident producers in the economy plus any product taxes and minus any subsidies not included in the value of the products. It is calculated without making deductions for depreciation of fabricated assets or for depletion and degradation of natural resources.

GDP per capita (current US$)	NY.GDP.PCAP.CD	GDP per capita is gross domestic product divided by midyear population. GDP is the sum of gross value added by all resident producers in the economy plus any product taxes and minus any subsidies not included in the value of the products. It is calculated without making deductions for depreciation of fabricated assets or for depletion and degradation of natural resources. Data are in current U.S. dollars.
Merchandise trade (% of GDP)	TG.VAL.TOTL.GD.ZS	Merchandise trade as a share of GDP is the sum of merchandise exports and imports divided by the value of GDP, all in current U.S. dollars.
Trade (% of GDP)	NE.TRD.GNFS.ZS	Trade is the sum of exports and imports of goods and services measured as a share of gross domestic product.
CPIA transparency, accountability, and corruption in the public sector rating (1=low to 6=high)	IQ.CPA.TRAN.XQ	Transparency, accountability, and corruption in the public sector assess the extent to which the executive can be held accountable for its use of funds and for the results of its actions by the electorate and by the legislature and judiciary, and the extent to which public employees within the executive are required to account for administrative decisions, use of resources, and results obtained. The three main dimensions assessed here are the accountability of the executive to oversight institutions and of public employees for their performance, access of civil society to information on

		public affairs, and state capture by narrow vested interests.
Mobile cellular subscriptions	IT.CEL.SETS	Mobile cellular telephone subscriptions are subscriptions to a public mobile telephone service that provide access to the PSTN using cellular technology. The indicator includes (and is split into) the number of postpaid subscriptions, and the number of active prepaid accounts (i.e. that have been used during the last three months). The indicator applies to all mobile cellular subscriptions that offer voice communications. It excludes subscriptions via data cards or USB modems, subscriptions to public mobile data services, private trunked mobile radio, telepoint, radio paging and telemetry services.
School enrollment, primary (% gross)	SE.PRM.ENRR	Gross enrollment ratio is the ratio of total enrollment, regardless of age, to the population of the age group that officially corresponds to the level of education shown. Primary education provides children with basic reading, writing, and mathematics skills along with an elementary understanding of such subjects as history, geography, natural science, social science, art, and music.
Maternal mortality ratio (modeled estimate, per 100,000 live births)	SH.STA.MMRT	Maternal mortality ratio is the number of women who die from pregnancy-related causes while pregnant or within 42 days of pregnancy termination per 100,000 live births. The data are estimated with a regression model using

		information on the proportion of maternal deaths among non-AIDS deaths in women ages 15-49, fertility, birth attendants, and GDP measured using purchasing power parities (PPPs).
Urban population (% of total population)	SP.URB.TOTL.IN.ZS	Urban population refers to people living in urban areas as defined by national statistical offices. The data are collected and smoothed by United Nations Population Division.
Air transport, registered carrier departures worldwide	IS.AIR.DPRT	Registered carrier departures worldwide are domestic takeoffs and takeoffs abroad of air carriers registered in the country.
Unemployment, total (% of total labor force) (modeled ILO estimate)	SL.UEM.TOTL.ZS	Unemployment refers to the share of the labor force that is without work but available for and seeking employment.
Unemployment, total (% of total labor force) (national estimate)	SL.UEM.TOTL.NE.ZS	Unemployment refers to the share of the labor force that is without work but available for and seeking employment. Definitions of labor force and unemployment differ by country.
People using safely managed drinking water services (% of population)	SH.H2O.SMDW.ZS	The percentage of people using drinking water from an improved source that is accessible on premises, available when needed and free from faecal and priority chemical contamination. Improved water sources include piped

		water, boreholes or tubewells, protected dug wells, protected springs, and packaged or delivered water.
Military expenditure (% of GDP)	MS.MIL.XPND.GD.ZS	Military expenditures data from SIPRI are derived from the NATO definition, which includes all current and capital expenditures on the armed forces, including peacekeeping forces; defense ministries and other government agencies engaged in defense projects; paramilitary forces, if these are judged to be trained and equipped for military operations; and military space activities. Such expenditures include military and civil personnel, including retirement pensions of military personnel and social services for personnel; operation and maintenance; procurement; military research and development; and military aid (in the military expenditures of the donor country). Excluded are civil defense and current expenditures for previous military activities, such as for veterans' benefits, demobilization, conversion, and destruction of weapons. This definition cannot be applied for all countries, however, since that would require much more detailed information than is available about what is included in military budgets and off-budget military expenditure items. (For example, military budgets might or might not cover civil defense, reserves and auxiliary forces, police and paramilitary forces, dual-purpose forces such as military and civilian

		police, military grants in kind, pensions for military personnel, and social security contributions paid by one part of government to another.)
Total natural resources rents (% of GDP)	NY.GDP.TOTL.RT.ZS	Total natural resources rents are the sum of oil rents, natural gas rents, coal rents (hard and soft), mineral rents, and forest rents.
Population, total	SP.POP.TOTL	Total population is based on the de facto definition of population, which counts all residents regardless of legal status or citizenship. The values shown are midyear estimates.
Control of Corruption: Percentile Rank	CC.PER.RNK	Control of Corruption captures perceptions of the extent to which public power is exercised for private gain, including both petty and grand forms of corruption, as well as "capture" of the state by elites and private interests. Percentile rank indicates the country's rank among all countries covered by the aggregate indicator, with 0 corresponding to lowest rank, and 100 to highest rank. Percentile ranks have been adjusted to correct for changes over time in the composition of the countries covered by the WGI.
Government Effectiveness: Percentile Rank	GE.PER.RNK	Government Effectiveness captures perceptions of the quality of public services, the quality of the civil service and the degree of its independence from political pressures, the quality of policy formulation and implementation, and the credibility of the government's

		commitment to such policies. Percentile rank indicates the country's rank among all countries covered by the aggregate indicator, with 0 corresponding to lowest rank, and 100 to highest rank. Percentile ranks have been adjusted to correct for changes over time in the composition of the countries covered by the WGI.
Political Stability and Absence of Violence/Terrorism: Percentile Rank	PV.PER.RNK	Political Stability and Absence of Violence/Terrorism measures perceptions of the likelihood of political instability and/or politically-motivated violence, including terrorism. Percentile rank indicates the country's rank among all countries covered by the aggregate indicator, with 0 corresponding to lowest rank, and 100 to highest rank. Percentile ranks have been adjusted to correct for changes over time in the composition of the countries covered by the WGI.
Rule of Law: Percentile Rank	RL.PER.RNK	Rule of Law captures perceptions of the extent to which agents have confidence in and abide by the rules of society, and in particular the quality of contract enforcement, property rights, the police, and the courts, as well as the likelihood of crime and violence. Percentile rank indicates the country's rank among all countries covered by the aggregate indicator, with 0 corresponding to lowest rank, and 100 to highest rank. Percentile ranks have been adjusted to correct for changes over time in the composition of the countries covered by the WGI.

Regulatory Quality: Percentile Rank	RQ.PER.RNK	Regulatory Quality captures perceptions of the ability of the government to formulate and implement sound policies and regulations that permit and promote private sector development. Percentile rank indicates the country's rank among all countries covered by the aggregate indicator, with 0 corresponding to lowest rank, and 100 to highest rank. Percentile ranks have been adjusted to correct for changes over time in the composition of the countries covered by the WGI.
Voice and Accountability: Percentile Rank	VA.PER.RNK	Voice and Accountability captures perceptions of the extent to which a country's citizens are able to participate in selecting their government, as well as freedom of expression, freedom of association, and a free media. Percentile rank indicates the country's rank among all countries covered by the aggregate indicator, with 0 corresponding to lowest rank, and 100 to highest rank. Percentile ranks have been adjusted to correct for changes over time in the composition of the countries covered by the WGI.

Appendix D: WDI primary source of data

Primary sources
International Monetary Fund
The International Energy Agency Statistics
Organisation for Economic Co-operation and Development GDP estimates
International Financial Statistics and data files
World Bank national accounts data, and OECD National Accounts data files.
World Trade Organization, and World Bank GDP estimates.
World Bank national accounts data
Organisation for Economic Co-operation and Development National Accounts data files
International Telecommunication Union (ITU) World Telecommunication/ICT Indicators Database
UNESCO Institute for Statistics
UNFPA, World Bank Group, and the United Nations Population Division. Trends in Maternal Mortality: 2000 to 2017. Geneva, , 2019
United Nations Population Division. World Urbanization Prospects: 2018 Revision.
International Civil Aviation Organization
International Labour Organization
The United Nations Children's Fund
Stockholm International Peace Research Institute
World Bank
World Health Organization
United Nations

Appendix E: WDI: Percentage of empty records by country

Country Name	Empty fields 1960-2020	Empty fields 2000-2019
Afghanistan	53%	31%
Albania	42%	8%
Algeria	28%	13%
American Samoa	74%	53%
Andorra	70%	62%
Angola	46%	18%
Antigua and Barbuda	54%	35%
Argentina	26%	14%
Armenia	51%	9%
Aruba	65%	46%
Australia	25%	13%
Austria	27%	9%
Azerbaijan	50%	7%
Bahamas, The	40%	24%
Bahrain	40%	15%
Bangladesh	31%	10%
Barbados	46%	29%
Belarus	49%	9%
Belgium	31%	9%
Belize	42%	22%
Benin	30%	18%

Bermuda	61%	54%
Bhutan	47%	16%
Bolivia	25%	11%
Bosnia and Herzegovina	58%	15%
Botswana	34%	18%
Brazil	29%	16%
British Virgin Islands	82%	80%
Brunei Darussalam	39%	17%
Bulgaria	39%	8%
Burkina Faso	32%	17%
Burundi	36%	24%
Cabo Verde	45%	17%
Cambodia	47%	6%
Cameroon	29%	13%
Canada	24%	12%
Cayman Islands	76%	60%
Central African Republic	39%	26%
Chad	37%	22%
Channel Islands	84%	76%
Chile	25%	9%
China	32%	15%
Colombia	25%	9%
Comoros	52%	31%
Congo, Dem. Rep.	36%	19%

Congo, Rep.	30%	14%
Costa Rica	26%	8%
Cote d'Ivoire	28%	12%
Croatia	54%	9%
Cuba	41%	22%
Curacao	79%	64%
Cyprus	36%	9%
Czech Republic	45%	8%
Denmark	27%	13%
Djibouti	57%	35%
Dominica	55%	39%
Dominican Republic	29%	16%
Ecuador	25%	9%
Egypt, Arab Rep.	27%	14%
El Salvador	29%	13%
Equatorial Guinea	51%	34%
Eritrea	64%	42%
Estonia	52%	9%
Eswatini	39%	27%
Ethiopia	41%	16%
Faroe Islands	76%	59%
Fiji	37%	21%
Finland	27%	9%
France	26%	9%

French Polynesia	68%	77%
Gabon	34%	24%
Gambia, The	41%	22%
Georgia	48%	6%
Germany	33%	9%
Ghana	27%	9%
Gibraltar	78%	74%
Greece	28%	9%
Greenland	61%	47%
Grenada	51%	29%
Guam	71%	48%
Guatemala	28%	13%
Guinea	52%	27%
Guinea-Bissau	46%	29%
Guyana	37%	27%
Haiti	39%	28%
Honduras	28%	13%
Hong Kong SAR, China	37%	22%
Hungary	43%	8%
Iceland	26%	8%
India	29%	16%
Indonesia	30%	13%
Iran, Islamic Rep.	28%	11%
Iraq	42%	26%

Ireland	29%	9%
Isle of Man	79%	69%
Israel	28%	8%
Italy	27%	9%
Jamaica	30%	16%
Japan	27%	14%
Jordan	31%	10%
Kazakhstan	50%	8%
Kenya	29%	15%
Kiribati	52%	39%
Korea, Dem. People's Rep.	72%	60%
Korea, Rep.	24%	9%
Kosovo	79%	47%
Kuwait	33%	10%
Kyrgyz Republic	50%	4%
Lao PDR	48%	17%
Latvia	54%	11%
Lebanon	47%	18%
Lesotho	39%	25%
Liberia	61%	28%
Libya	51%	24%
Liechtenstein	71%	61%
Lithuania	54%	11%
Luxembourg	32%	9%

Macao SAR, China	52%	22%
Madagascar	33%	17%
Malawi	34%	18%
Malaysia	26%	9%
Maldives	53%	28%
Mali	37%	19%
Malta	35%	10%
Marshall Islands	67%	51%
Mauritania	37%	20%
Mauritius	36%	13%
Mexico	26%	8%
Micronesia, Fed. Sts.	66%	44%
Moldova	52%	4%
Monaco	68%	59%
Mongolia	43%	6%
Montenegro	66%	21%
Morocco	26%	9%
Mozambique	47%	14%
Myanmar	46%	21%
Namibia	49%	17%
Nauru	78%	62%
Nepal	29%	9%
Netherlands	28%	9%
New Caledonia	66%	72%

New Zealand	26%	9%
Nicaragua	29%	13%
Niger	34%	16%
Nigeria	27%	9%
North Macedonia	52%	12%
Northern Mariana Islands	77%	58%
Norway	25%	12%
Oman	33%	14%
Pakistan	24%	5%
Palau	79%	56%
Panama	28%	14%
Papua New Guinea	38%	24%
Paraguay	26%	10%
Peru	24%	8%
Philippines	24%	9%
Poland	44%	10%
Portugal	28%	9%
Puerto Rico	52%	36%
Qatar	40%	13%
Romania	43%	9%
Russian Federation	49%	9%
Rwanda	36%	19%
Samoa	52%	19%
San Marino	74%	58%

Sao Tome and Principe	64%	32%
Saudi Arabia	36%	15%
Senegal	30%	14%
Serbia	64%	11%
Seychelles	41%	25%
Sierra Leone	35%	17%
Singapore	29%	13%
Sint Maarten (Dutch part)	89%	79%
Slovak Republic	51%	11%
Slovenia	51%	9%
Solomon Islands	47%	26%
Somalia	61%	69%
South Africa	29%	13%
South Sudan	79%	58%
Spain	27%	9%
Sri Lanka	26%	10%
St. Kitts and Nevis	57%	49%
St. Lucia	54%	32%
St. Martin (French part)	95%	95%
St. Vincent and the Grenadines	47%	31%
Sudan	31%	16%
Suriname	43%	26%
Sweden	25%	12%
Switzerland	26%	9%

Syrian Arab Republic	36%	42%
Tajikistan	52%	10%
Tanzania	44%	13%
Thailand	26%	13%
Timor-Leste	70%	34%
Togo	30%	16%
Tonga	48%	28%
Trinidad and Tobago	34%	22%
Tunisia	28%	9%
Turkey	26%	13%
Turkmenistan	59%	29%
Turks and Caicos Islands	78%	64%
Tuvalu	72%	56%
Uganda	33%	12%
Ukraine	49%	9%
United Arab Emirates	45%	23%
United Kingdom	24%	8%
United States	24%	10%
Uruguay	28%	16%
Uzbekistan	53%	12%
Vanuatu	48%	21%
Venezuela, RB	30%	25%
Vietnam	45%	14%
Virgin Islands (U.S.)	77%	61%

West Bank and Gaza	72%	33%
Yemen, Rep.	53%	23%
Zambia	34%	14%
Zimbabwe	33%	19%

Appendix F: WDGI Data: Percentage of empty fields per indicator

Indicators	**Empty fields**
CPIA transparency, accountability, and corruption in the public sector rating (1=low to 6=high)	73%
People using safely managed drinking water services (% of population)	59%
Electric power consumption (kWh per capita)	47%
Unemployment, total (% of total labor force) (national estimate)	41%
Maternal mortality ratio (modeled estimate, per 100,000 live births)	25%
Military expenditure (% of GDP)	24%
Air transport, registered carrier departures worldwide	24%
School enrollment, primary (% gross)	22%
Domestic credit to private sector by banks (% of GDP)	13%
Trade (% of GDP)	11%
Oil rents (% of GDP)	9%
Unemployment, total (% of total labor force) (modeled ILO estimate)	9%
Total natural resources rents (% of GDP)	7%
Merchandise trade (% of GDP)	5%

GDP per capita growth (annual %)	3%
GDP per capita (current US$)	2%
Political Stability and Absence of Violence/Terrorism: Percentile Rank	2%
Regulatory Quality: Percentile Rank	2%
Voice and Accountability: Percentile Rank	2%
Government Effectiveness: Percentile Rank	1%
Control of Corruption: Percentile Rank	1%
Mobile cellular subscriptions	1%
Urban population (% of total population)	1%
Rule of Law: Percentile Rank	1%
Population, total	0%

Appendix G: ACLED Data: Number of records entered every year by timestamp

Timestamp	Number of Timestamp missing	1997	1998	1999	2000	2001	…	2015	2016	2017	2018	2019
1552577661	0					945						
1552577662	0				654	297						
1552577663	0				938							
1552577664	0				848							
1552577665	0				883							
1552577666	0			138	769							
1552577667	0			902								
1552577668	0			835								

1552577669	0			812									
1552577670	0			849									
1552577671	0			885									
1552577672	0		518	346									
1552577673	0		831										
1552577674	0		860										
1552577675	0		822										
1552577676	0		898										
1552577677	0	382	423										
1552577678	0	794											
1552577679	0	816											
1552577680	0	827											
1552577681	127061	397											
1552704743	227974						4			6	2		
1552932718	0	2	3	1	4	5	11	11	35	83	105		
1552932719	12349						1			82	345		
1552945069	0								4	16	561		
1552945070	4044							10	5	3	109		
1552949115	0										114		
1552949116	2										1		
1552949119	0										1		
1552949120	1										1		
1552949122	45995								4	20	327		
1552995118	0										543		

1552995119	1											1
1552995121	0									75	55	187
1552995122	0									2		
1552995123	0									133		
1552995124	1						129	115	14			
1552995126	0						66					
1552995127	0											
1552995128	0											
1552995129	1											
1552995131	549701											
1553544833	0	4	11	3	3	4	64	47	54	87	39	

Appendix H: Regression analysis: Source code out outputs

Appendix H-1: Overview of datasets

```
'data.frame':    137 obs. of  16 variables:
 $ poverty_ratio     : num  54.5 NA NA 30.3 29.4 NA NA NA NA NA ...
 $ gini              : num  NA NA NA 42 32.5 NA NA NA NA NA ...
 $ growth_rate       : num  NA NA NA NA NA NA NA NA NA NA ...
 $ population        : int  35383128 28842484 2876101 43590368 2936146 9452
812 10487998 10872067 18646378 ...
 $ education         : num  4.23 NA 3.95 5.57 2.76 NA 2.9 4.69 3.99 NA ...
 $ unemployment      : num  1.63 7.28 15.22 8.02 17.62 ...
 $ electricity       : num  97.7 40.6 100 100 100 ...
 $ natural_resources : num  0.63 10.75 1.05 1.13 4.41 ...
 $ gdp               : num  1.94e+10 1.01e+11 1.19e+10 5.58e+11 1.05e+10 ..
 $ percapita         : int  1910 6410 12060 19690 9010 21880 16280 740 2160
0 ...
 $ foreign_investment: num  0.48 -0.18 8.8 0.58 3.16 ...
 $ military_exp      : num  0.96 2.73 1.1 0.81 4.08 NA 3.72 2.12 0.93 1.23
 $ forces_prsnl      : int  323000 117000 8500 105450 49100 180 81950 51050
11450 ...
 $ debt_extrnl       : num  2.60e+09 5.72e+10 8.74e+09 1.89e+11 9.96e+09 ..
 $ merchandise_trade : num  36.8 40.2 55.9 20.4 48 ...
 $ crisis_event      : int  NA 45 NA NA NA NA NA 1175 17 104 ...
```

Appendix H-2: Summary of variables

```
 poverty_ratio        gini         growth_rate       population
 Min.   : 2.50   Min.   :25.00   Min.   :-1.01   Min.   :1.122e+04
 1st Qu.:17.40   1st Qu.:33.40   1st Qu.:-0.55   1st Qu.:2.079e+06
 Median :24.90   Median :40.95   Median : 1.31   Median :9.655e+06
 Mean   :26.48   Mean   :39.70   Mean   : 1.52   Mean   :4.454e+07
 3rd Qu.:30.40   3rd Qu.:45.17   3rd Qu.: 3.47   3rd Qu.:2.857e+07
 Max.   :82.30   Max.   :53.70   Max.   : 4.78   Max.   :1.379e+09
 NA's   :98      NA's   :101     NA's   :124     NA's   :1
   education       unemployment     electricity     natural_resources
 Min.   : 1.500   Min.   : 0.310   Min.   :  9.30   Min.   : 0.000
 1st Qu.: 3.388   1st Qu.: 3.440   1st Qu.: 55.60   1st Qu.: 1.180
 Median : 4.390   Median : 6.260   Median : 91.80   Median : 3.700
 Mean   : 4.450   Mean   : 7.994   Mean   : 76.25   Mean   : 7.495
 3rd Qu.: 5.312   3rd Qu.:11.020   3rd Qu.: 99.97   3rd Qu.:12.283
 Max.   :10.240   Max.   :26.890   Max.   :100.00   Max.   :42.380
 NA's   :71       NA's   :10                        NA's   :11
      gdp              percapita     foreign_investment  military_exp
 Min.   :3.655e+07   Min.   :  740   Min.   :-37.150    Min.   :0.000
 1st Qu.:4.330e+09   1st Qu.: 3490   1st Qu.:  1.042    1st Qu.:0.950
 Median :1.539e+10   Median : 7390   Median :  2.600    Median :1.430
 Mean   :1.911e+11   Mean   : 8786   Mean   :  3.521    Mean   :1.817
 3rd Qu.:6.879e+10   3rd Qu.:12840   3rd Qu.:  5.633    3rd Qu.:2.440
 Max.   :1.110e+13   Max.   :27460   Max.   : 26.210    Max.   :6.380
 NA's   :5           NA's   :8       NA's   :9          NA's   :36
  forces_prsnl      debt_extrnl       merchandise_trade  crisis_event
 Min.   :    180   Min.   :8.091e+07   Min.   : 11.94    Min.   :    3.00
 1st Qu.:  11725   1st Qu.:2.146e+09   1st Qu.: 36.00    1st Qu.:   43.75
 Median :  35000   Median :7.950e+09   Median : 49.20    Median :  146.50
 Mean   : 177374   Mean   :5.094e+10   Mean   : 54.51    Mean   :  989.68
 3rd Qu.: 158300   3rd Qu.:2.649e+10   3rd Qu.: 70.59    3rd Qu.:  858.25
 Max.   :2981050   Max.   :1.420e+12   Max.   :171.18    Max.   :13305.00
 NA's   :22        NA's   :18          NA's   :9         NA's   :69
```

Appendix H-3: Shapiro-Wilk test results

"population"	W = 0.65033, p-value = 1.094e-10
"unemployment"	W = 0.86353, p-value = 6.74e-06
"electricity"	W = 0.92228, p-value = 0.0008525
"natural_resources"	W = 0.89536, p-value = 0.0001019
"gdp"	W = 0.52882, p-value = 1.769e-12
"percapita"	W = 0.83201, p-value = 1.303e-06
"foreign_investment"	W = 0.66306, p-value = 2.879e-10
"military_exp"	W = 0.88553, p-value = 9.331e-0
"forces_prsnl"	W = 0.6817, p-value = 6.221e-10
"debt_extrnl"	W = 0.49681, p-value = 1.422e-12
"merchandise_trade"	W = 0.87547, p-value = 2.5e-05
"crisis_event"	W = 0.75054, p-value = 7.86e-09

Appendix H-4: Correlation matrix of all variables

	population	unemployment	electricity	natural_resources	gdp	percapita
population	1	-0.32154	0.138978	-0.16636	0.793288	0.007487
unemployment	-0.32154	1	0.319216	-0.09447	0.055124	0.534727
electricity	0.138978	0.319216	1	-0.62222	0.611391	0.824509
natural_resources	-0.16636	-0.09447	-0.62222	1	-0.41594	-0.47298
gdp	0.793288	0.055124	0.611391	-0.41594	1	0.57497
percapita	0.007487	0.534727	0.824509	-0.47298	0.57497	1
foreign_investment	-0.04038	-0.17041	-0.12744	0.116372	-0.09489	-0.16367
military_exp	-0.21277	0.304705	0.27703	-0.16526	0.10007	0.384024
forces_prsnl	0.718865	-0.124	0.463799	-0.39618	0.82114	0.396234
debt_extrnl	0.666247	0.038132	0.605229	-0.49553	0.889232	0.586732
merchandise_trade	-0.38988	0.150638	0.214351	-0.0105	-0.19387	0.288624
crisis_event	0.615881	-0.00106	0.216587	-0.07534	0.641132	0.105894

	foreign_investment	military_exp	forces_prsnl	debt_extrnl	merchandise_trade	crisis_event
population	-0.04038	-0.21277	0.718865	0.666247	-0.38988	0.615881
unemployment	-0.17041	0.304705	-0.124	0.038132	0.150638	-0.00106
electricity	-0.12744	0.27703	0.463799	0.605229	0.214351	0.216587
natural_resources	0.116372	-0.16526	-0.39618	-0.49553	-0.0105	-0.07534
gdp	-0.09489	0.10007	0.82114	0.889232	-0.19387	0.641132
percapita	-0.16367	0.384024	0.396234	0.586732	0.288624	0.105894
foreign_investment	1	-0.08483	-0.06977	0.067124	0.190977	-0.11182
military_exp	-0.08483	1	0.295558	0.067427	0.237326	-0.00599
forces_prsnl	-0.06977	0.295558	1	0.676539	-0.23384	0.62466
debt_extrnl	0.067124	0.067427	0.676539	1	-0.08222	0.576629
merchandise_trade	0.190977	0.237326	-0.23384	-0.08222	1	-0.40406
crisis_event	-0.11182	-0.00599	0.62466	0.576629	-0.40406	1

Appendix H-5: Normalisation of the dependent variable

```
library(bestNormalize)

#First standardise the variable
norm_data <- scale(indicator2$crisis_event)
#Select the best method
best_norm <- bestNormalize(norm_data)
#Result of the transformation
norm_data <- predict(best_norm)

indicator2$crisis_event <- norm_data

> best_norm
Best Normalizing transformation with 60 Observations
       Estimated Normality Statistics (Pearson P / df, lower => more normal):
       - No transform: 3.1833
       - Log_b(x+a): 1.406
       - sqrt(x+a): 1.912
       - exp(x): 4.75
       - arcsinh(x): 3.0833
       - Yeo-Johnson: 2
       - orderNorm: 1.3167
      Estimation method: Out-of-sample via CV with 10 folds and 5 repeats

      Based off these, bestNormalize chose:
      orderNorm Transformation with 60 nonmissing obs and ties
       - 54 unique values
       - Original quantiles:
          0%    25%    50%    75%   100%
      -0.759 -0.683 -0.516  0.164  2.518

> shapiro.test(norm_data)
        Shapiro-Wilk normality test

      data:  norm_data
      W = 0.99858, p-value = 1
```

Appendix H-6: Kaiser-Meyer-Olkin factor adequacy

```
Kaiser-Meyer-Olkin factor adequacy
Call: KMO(r = cor(indicator3))
Overall MSA =  0.67
MSA for each item =
        population       unemployment        electricity  natural_res
ources
              0.63               0.66               0.72
0.63
               gdp          percapita foreign_investment       milita
ry_exp
```

Appendix H-7: Varimax rotation results

```
Principal Components Analysis
Call: principal(r = indicator3, nfactors = 4, rotate = "varimax")
Standardized loadings (pattern matrix) based upon correlation matrix
                     RC1   RC2   RC3   RC4   h2   u2 com
population          0.85 -0.18 -0.14 -0.12 0.78 0.22 1.2
unemployment       -0.23  0.75 -0.17  0.09 0.65 0.35 1.3
electricity         0.44  0.75  0.13 -0.07 0.78 0.22 1.7
natural_resources  -0.25 -0.54 -0.09  0.57 0.69 0.31 2.4
gdp                 0.91  0.22 -0.10  0.00 0.89 0.11 1.1
percapita           0.29  0.77  0.04  0.23 0.73 0.27 1.5
foreign_investment -0.07 -0.31  0.77 -0.07 0.69 0.31 1.4
military_exp        0.08  0.24  0.11  0.83 0.76 0.24 1.2
forces_prsnl        0.84  0.11  0.01  0.19 0.75 0.25 1.1
debt_extrnl         0.70  0.32  0.19 -0.12 0.64 0.36 1.6
merchandise_trade   0.00  0.28  0.79  0.17 0.73 0.27 1.3

                       RC1  RC2  RC3  RC4
SS loadings           3.16 2.43 1.34 1.17
Proportion Var        0.29 0.22 0.12 0.11
Cumulative Var        0.29 0.51 0.63 0.74
Proportion Explained  0.39 0.30 0.17 0.14
Cumulative Proportion 0.39 0.69 0.86 1.00

Mean item complexity =  1.5
Test of the hypothesis that 4 components are sufficient.

The root mean square of the residuals (RMSR) is  0.08
 with the empirical chi square  44.04  with prob <  0.00034

Fit based upon off diagonal values = 0.94
```

Appendix H-8: Factor analysis: Shapiro-Wilk test results

```
data:  fit.data$population_feature
W = 0.81489, p-value = 3.238e-07

data:  fit.data$population_feature2
W = 0.96315, p-value = 0.06718

data:  fit.data$random_factor
W = 0.87082, p-value = 1.328e-05

data:  fit.data$random_factor2
W = 0.91337, p-value = 0.000419
```

Appendix H-9: Spearman rank relation between factors and crisis event data

```
Spearman's rank correlation rho

data:  crisis_event and fit.data$population_feature
S = 15361, p-value = 1.697e-06
alternative hypothesis: true rho is not equal to 0
sample estimates:
      rho
0.5731951


data:  crisis_event and fit.data$population_feature2
S = 37558, p-value = 0.741
alternative hypothesis: true rho is not equal to 0
sample estimates:
       rho
-0.0435725


data:  crisis_event and fit.data$random_factor
S = 49298, p-value = 0.003639
alternative hypothesis: true rho is not equal to 0
sample estimates:
       rho
-0.3697827


data:  crisis_event and fit.data$random_factor2
S = 34352, p-value = 0.7298
alternative hypothesis: true rho is not equal to 0
sample estimates:
      rho
0.0455177
```

Appendix H-10: Multilinear regression: Dataset 1 summary

```
Call:
lm(formula = indicator2$crisis_event ~ indicator2$population +
    indicator2$gdp + indicator2$forces_prsnl + indicator2$debt_extrnl +
    indicator2$merchandise_trade)

Residuals:
     Min       1Q   Median       3Q      Max
-1.96484 -0.37955  0.06871  0.35944  1.84114

Coefficients:
                                Estimate Std. Error t value Pr(>|t|)
(Intercept)                   7.997e-02  2.412e-01   0.332  0.74166
indicator2$population         1.159e-08  4.256e-09   2.724  0.00902 **
indicator2$gdp               -2.181e-12  2.093e-12  -1.042  0.30284
indicator2$forces_prsnl       7.725e-07  8.700e-07   0.888  0.37907
indicator2$debt_extrnl        1.266e-11  4.720e-12   2.682  0.01007 *
indicator2$merchandise_trade -1.206e-02  3.774e-03  -3.197  0.00249 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.731 on 47 degrees of freedom
  (7 observations deleted due to missingness)
Multiple R-squared:  0.483,     Adjusted R-squared:  0.428
F-statistic: 8.781 on 5 and 47 DF,  p-value: 6.137e-06
```

Appendix H-11: Multilinear regression: Dataset 2 summary

```
Call:
lm(formula = indicator2$crisis_event ~ fit.data$population_feature +
    fit.data$random_factor)

Residuals:
     Min       1Q   Median       3Q      Max
-2.16404 -0.64870  0.05262  0.46143  2.05070

Coefficients:
                             Estimate Std. Error t value Pr(>|t|)
(Intercept)                 0.0005296  0.1087146   0.005   0.9961
fit.data$population_feature  0.4975418  0.1096320   4.538 2.98e-05 ***
fit.data$random_factor      -0.2459149  0.1096320  -2.243   0.0288 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.8421 on 57 degrees of freedom
Multiple R-squared:  0.3102,    Adjusted R-squared:  0.286
F-statistic: 12.81 on 2 and 57 DF,  p-value: 2.537e-05
```

Appendix H-12: Multilinear regression: Improved dataset 1 summary

```
Call:
lm(formula = indicator2$crisis_event ~ indicator2$population +
    indicator2$debt_extrnl + indicator2$merchandise_trade)

Residuals:
     Min       1Q   Median       3Q      Max
-1.98481 -0.39121  0.02776  0.41063  1.83152

Coefficients:
                                 Estimate Std. Error t value Pr(>|t|)
(Intercept)                   9.126e-02  2.379e-01   0.384  0.70295
indicator2$population         9.702e-09  2.983e-09   3.253  0.00205 **
indicator2$debt_extrnl        1.065e-11  3.682e-12   2.893  0.00563 **
indicator2$merchandise_trade -1.176e-02  3.637e-03  -3.233  0.00217 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.7233 on 50 degrees of freedom
  (6 observations deleted due to missingness)
Multiple R-squared:  0.4835,    Adjusted R-squared:  0.4525
F-statistic:  15.6 on 3 and 50 DF,  p-value: 2.724e-07
```

Appendix H-13: Multilinear regression of factors: relative importance of variables

```
Response variable: indicator2$crisis_event
Total response variance: 0.9931152
Analysis based on 60 observations

2 Regressors:
fit.data$population_feature fit.data$random_factor
Proportion of variance explained by model: 31.02%
Metrics are normalized to sum to 100% (rela=TRUE).

Relative importance metrics:

                                  lmg
fit.data$population_feature 0.8036693
fit.data$random_factor      0.1963307

Average coefficients for different model sizes:

                                   1X        2Xs
fit.data$population_feature  0.4975418  0.4975418
fit.data$random_factor      -0.2459149 -0.2459149
```

Appendix H-14: Poisson regression: dataset 1 summary

```
Call:
glm(formula = indicator2$crisis_event ~ indicator2$population +
    indicator2$gdp + indicator2$forces_prsnl + indicator2$debt_extrnl +
    indicator2$merchandise_trade, family = "quasipoisson")

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-25.161  -11.552   -6.257    2.402   41.275

Coefficients:
                              Estimate Std. Error t value Pr(>|t|)
(Intercept)                  5.986e+00  4.128e-01  14.502  < 2e-16 ***
indicator2$population        1.153e-08  4.475e-09   2.576  0.01320 *
indicator2$gdp              -3.022e-12  2.380e-12  -1.270  0.21043
indicator2$forces_prsnl      6.823e-07  7.540e-07   0.905  0.37012
indicator2$debt_extrnl       1.710e-11  5.425e-12   3.153  0.00281 **
indicator2$merchandise_trade -2.302e-02  7.842e-03  -2.936  0.00513 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for quasipoisson family taken to be 301.7443)

    Null deviance: 23838  on 52  degrees of freedom
Residual deviance: 11043  on 47  degrees of freedom
  (7 observations deleted due to missingness)
AIC: NA
```

Appendix H-15: Poisson regression: dataset 2 summary

```
Call:
glm(formula = indicator2$crisis_event ~ fit.data$population_feature +
    fit.data$random_factor, family = "quasipoisson")

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-30.507  -15.613   -7.629    3.743   41.355

Coefficients:
                            Estimate Std. Error t value Pr(>|t|)
(Intercept)                   5.5485     0.1699  32.657  < 2e-16 ***
fit.data$population_feature   0.3621     0.1091   3.318  0.00158 **
fit.data$random_factor       -0.6298     0.1889  -3.334  0.00151 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for quasipoisson family taken to be 375.8237)

    Null deviance: 29202  on 59  degrees of freedom
Residual deviance: 18955  on 57  degrees of freedom
AIC: NA
```

Appendix H-16: Poisson regression: dataset 1 summary-2015

```
Call:
glm(formula = indicator2$crisis_event ~ indicator2$population +
    indicator2$gdp + indicator2$forces_prsnl + indicator2$debt_extrnl +
    indicator2$merchandise_trade, family = "quasipoisson")

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-17.300   -9.218   -5.386    1.771   28.536

Coefficients:
                              Estimate Std. Error t value Pr(>|t|)
(Intercept)                  5.226e+00  3.557e-01  14.691  < 2e-16 ***
indicator2$population        1.304e-08  5.843e-09   2.231  0.03151 *
indicator2$gdp              -6.773e-12  3.657e-12  -1.852  0.07162 .
indicator2$forces_prsnl      3.446e-06  1.047e-06   3.291  0.00212 **
indicator2$debt_extrnl       1.379e-11  1.174e-11   1.175  0.24715
indicator2$merchandise_trade -1.466e-02  5.984e-03  -2.449  0.01891 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for quasipoisson family taken to be 158.0927)
```

Appendix I: Image database creation and CNN application: Source code out outputs

Appendix I-1: List of indicators in the WDGI dataset

```
[1] "country"  "country.code"  "region"   "sub.region"
 [5] "intermediate.region" "EG.USE.ELEC.KH.PC"   "FD.AST.PRVT.GD.ZS"   "IQ.CP
A.TRAN.XQ"
 [9] "IS.AIR.DPRT"  "IT.CEL.SETS"  "MS.MIL.XPND.GD.ZS" "NE.TRD.GNFS.ZS"
[13] "NY.GDP.PCAP.CD" "NY.GDP.PCAP.KD.ZG" "NY.GDP.PETR.RT.ZS"  "NY.GDP.TOTL.R
T.ZS"
[17] "SE.PRM.ENRR" "SH.H2O.SMDW.ZS"  "SH.STA.MMRT" "SL.UEM.TOTL.NE.ZS"
[21] "SL.UEM.TOTL.ZS"  "SP.POP.TOTL" "SP.URB.TOTL.IN.ZS" "TG.VAL.TOTL.GD.ZS"
[25] "CC.EST" "GE.EST" "PV.EST" "RL.EST"
[29] "RQ.EST"  "VA.EST" "YEARS" "fatalities"
```

Appendix I-2: summary of initial dataset variables

```
country          country.code          region           sub.region
 Length:1980        Length:1980        Length:1980        Length:1980
 Class :character   Class :character   Class :character   Class :character
 Mode  :character   Mode  :character   Mode  :character   Mode  :character



 intermediate.region EG.USE.ELEC.KH.PC  FD.AST.PRVT.GD.ZS IQ.CPA.TRAN.XQ   IS.AIR.DPRT
 Length:1980         Min.   :   22.76   Min.   :  0.00    Min.   :1.000   Min.   :      0
 Class :character    1st Qu.:  249.18   1st Qu.: 11.15    1st Qu.:2.500   1st Qu.:   6019
 Mode  :character    Median : 1470.57   Median : 21.28    Median :2.500   Median :  16342
                     Mean   : 2492.00   Mean   : 31.94    Mean   :2.667   Mean   :  64633
                     3rd Qu.: 3065.57   3rd Qu.: 44.95    3rd Qu.:3.000   3rd Qu.:  51660
                     Max.   :21508.36   Max.   :255.19    Max.   :4.000   Max.   :1209803
                     NA's   :835        NA's   :137       NA's   :1197    NA's   :396
  IT.CEL.SETS        MS.MIL.XPND.GD.ZS NE.TRD.GNFS.ZS     NY.GDP.PCAP.CD    NY.GDP.PCAP.KD.ZG
 Min.   :0.000e+00   Min.   : 0.0163   Min.   :  0.1674   Min.   :  111.9   Min.   :-62.378
 1st Qu.:8.967e+05   1st Qu.: 1.2284   1st Qu.: 52.4873   1st Qu.:  692.1   1st Qu.:  0.450
 Median :4.346e+06   Median : 1.7651   Median : 73.6482   Median : 1677.1   Median :  2.845
 Mean   :2.309e+07   Mean   : 2.4228   Mean   : 79.0537   Mean   : 5186.3   Mean   :  2.575
 3rd Qu.:1.522e+07   3rd Qu.: 3.0283   3rd Qu.: 97.9540   3rd Qu.: 5043.2   3rd Qu.:  4.855
 Max.   :1.176e+09   Max.   :32.6557   Max.   :347.9965   Max.   :85076.1   Max.   :121.780
 NA's   :16          NA's   :267       NA's   :117        NA's   :33        NA's   :45
 NY.GDP.PETR.RT.ZS  NY.GDP.TOTL.RT.ZS  SE.PRM.ENRR     SH.H2O.SMDW.ZS     SH.STA.MMRT
 Min.   : 0.00000   Min.   : 0.000    Min.   : 20.96   Min.   :  4.529   Min.   :   2.0
 1st Qu.: 0.00000   1st Qu.: 1.740    1st Qu.: 94.24   1st Qu.: 43.013   1st Qu.:  28.0
 Median : 0.08427   Median : 6.467    Median :101.21   Median : 70.852   Median : 190.0
 Mean   : 6.27676   Mean   :11.808    Mean   :100.60   Mean   : 63.947   Mean   : 305.5
 3rd Qu.: 4.12616   3rd Qu.:16.911    3rd Qu.:108.60   3rd Qu.: 91.524   3rd Qu.: 509.0
 Max.   :78.54109   Max.   :84.229    Max.   :149.96   Max.   :100.000   Max.   :2480.0
 NA's   :135        NA's   :131       NA's   :409      NA's   :1187      NA's   :314
SL.UEM.TOTL.NE.ZS SL.UEM.TOTL.ZS    SP.POP.TOTL       SP.URB.TOTL.IN.ZS TG.VAL.TOTL.GD.ZS
 Min.   : 0.11     Min.   : 0.091   Min.   :5.516e+05  Min.   :  7.83    Min.   :0.00001
 1st Qu.: 4.00     1st Qu.: 3.235   1st Qu.:3.930e+06  1st Qu.: 32.75    1st Qu.:0.00004
 Median : 8.05     Median : 6.135   Median :1.106e+07  Median : 47.32    Median :0.00006
 Mean   :10.04     Mean   : 8.501   Mean   :3.897e+07  Mean   : 49.39    Mean   :0.00006
 3rd Qu.:13.57     3rd Qu.:11.557   3rd Qu.:2.947e+07  3rd Qu.: 66.51    3rd Qu.:0.00008
 Max.   :55.00     Max.   :37.250   Max.   :1.366e+09  Max.   :100.00    Max.   :0.00024
 NA's   :989       NA's   :20       NA's   :8          NA's   :28        NA's   :82
     CC.EST            GE.EST            PV.EST            RL.EST            RQ.EST
 Min.   :-1.8264   Min.   :-2.27942   Min.   :-3.1808   Min.   :-2.1300   Min.   :-2.34695
 1st Qu.:-1.0427   1st Qu.:-0.97684   1st Qu.:-1.1740   1st Qu.:-1.0455   1st Qu.:-0.94876
 Median :-0.6380   Median :-0.57657   Median :-0.4928   Median :-0.5878   Median :-0.45550
 Mean   :-0.5699   Mean   :-0.50846   Mean   :-0.5979   Mean   :-0.5594   Mean   :-0.47515
 3rd Qu.:-0.2139   3rd Qu.:-0.07453   3rd Qu.: 0.0489   3rd Qu.:-0.1147   3rd Qu.:-0.03339
 Max.   : 1.5672   Max.   : 1.56367   Max.   : 1.2236   Max.   : 1.2165   Max.   : 1.42331
 NA's   :1         NA's   :11         NA's   :14        NA's   :1         NA's   :12
     VA.EST            YEARS        fatalities
 Min.   :-2.2592   Min.   :1998   Min.   :    0.0
 1st Qu.:-1.2022   1st Qu.:2005   1st Qu.:    1.0
 Median :-0.7113   Median :2010   Median :   13.0
 Mean   :-0.6596   Mean   :2009   Mean   :  536.2
 3rd Qu.:-0.1233   3rd Qu.:2014   3rd Qu.:  158.2
 Max.   : 1.1914   Max.   :2019   Max.   :41981.0
                                  NA's   :840
```

Appendix I-3: R script: Deletion of almost empty indicators

```
> indicator1 <- subset.data.frame(indicator, select = -c(IQ.CPA.TRAN.XQ, S
H.H2O.SMDW.ZS, SL.UEM.TOTL.NE.ZS, EG.USE.ELEC.KH.PC, NY.GDP.PETR.RT.ZS, SE.
PRM.ENRR, SH.STA.MMRT, IS.AIR.DPRT, MS.MIL.XPND.GD.ZS, NY.GDP.TOTL.RT.ZS))
```

Appendix I-4: Summary of the variables after the deletion of almost empty indicators and missing values in fatalities

```
   country           country.code          region           sub.region
 Length:1140        Length:1140        Length:1140        Length:1140
 Class :character   Class :character   Class :character   Class :character
 Mode  :character   Mode  :character   Mode  :character   Mode  :character



 intermediate.region FD.AST.PRVT.GD.ZS  IT.CEL.SETS         NE.TRD.GNFS.ZS
 Length:1140         Min.   :  0.000   Min.   :0.000e+00   Min.   :  0.2004
 Class :character    1st Qu.:  9.528   1st Qu.:6.862e+05   1st Qu.: 47.7739
 Mode  :character    Median : 15.702   Median :4.424e+06   Median : 63.6598
                     Mean   : 24.448   Mean   :2.232e+07   Mean   : 72.9980
                     3rd Qu.: 28.596   3rd Qu.:1.695e+07   3rd Qu.: 90.2855
                     Max.   :137.912   Max.   :1.176e+09   Max.   :347.9965
                     NA's   :76        NA's   :1           NA's   :69
 NY.GDP.PCAP.CD    NY.GDP.PCAP.KD.ZG  SL.UEM.TOTL.ZS    SP.POP.TOTL        SP.URB.TOTL.IN.Z
S
 Min.   :  111.9   Min.   :-62.3781   Min.   : 0.091   Min.   :5.585e+05   Min.   :  7.83
 1st Qu.:  536.4   1st Qu.:  0.1953   1st Qu.: 2.783   1st Qu.:4.561e+06   1st Qu.: 29.53
 Median : 1118.8   Median :  2.2518   Median : 5.093   Median :1.293e+07   Median : 40.34
 Mean   : 2934.8   Mean   :  1.9993   Mean   : 7.851   Mean   :3.167e+07   Mean   : 43.30
 3rd Qu.: 3154.0   3rd Qu.:  4.2123   3rd Qu.:10.640   3rd Qu.:3.153e+07   3rd Qu.: 56.03
 Max.   :62088.1   Max.   :121.7795   Max.   :37.002   Max.   :1.366e+09   Max.   :100.00
 NA's   :15        NA's   :29         NA's   :2        NA's   :8           NA's   :10
 TG.VAL.TOTL.GD.ZS      CC.EST             GE.EST             PV.EST             RL.EST
 Min.   :0.000008   Min.   :-1.8264   Min.   :-2.2794   Min.   :-2.99331   Min.   :-2.1300
 1st Qu.:0.000036   1st Qu.:-1.0642   1st Qu.:-1.1254   1st Qu.:-1.21828   1st Qu.:-1.1186
 Median :0.000048   Median :-0.6987   Median :-0.7099   Median :-0.51874   Median :-0.6783
 Mean   :0.000057   Mean   :-0.6656   Mean   :-0.6882   Mean   :-0.65521   Mean   :-0.6850
 3rd Qu.:0.000071   3rd Qu.:-0.3628   3rd Qu.:-0.3556   3rd Qu.:-0.02898   3rd Qu.:-0.2838
 Max.   :0.000245   Max.   : 1.2167   Max.   : 1.4313   Max.   : 1.20023   Max.   : 1.0722
 NA's   :20
     RQ.EST             VA.EST             YEARS        fatalities
 Min.   :-2.3469   Min.   :-2.2261   Min.   :1998   Min.   :    0.0
 1st Qu.:-1.0249   1st Qu.:-1.2306   1st Qu.:2006   1st Qu.:    1.0
 Median :-0.6131   Median :-0.7989   Median :2011   Median :   13.0
 Mean   :-0.6395   Mean   :-0.7208   Mean   :2011   Mean   :  536.2
 3rd Qu.:-0.2426   3rd Qu.:-0.2072   3rd Qu.:2016   3rd Qu.:  158.2
 Max.   : 1.3126   Max.   : 1.0773   Max.   :2019   Max.   :41981.0
```

Appendix I-5: Outliers in fatalities removed

```
  [1] 17669  1235  2752   948  1234   901   567   490  2430  3248  1081  3856  3881
1416  1247
 [16]  1406  1344   415  1421  2568   537  1104  2136   476  6097   978  3734   404
1053  7399
 [31]   499  1672  1627  1170  3203   565   562   954  3126  2253   536   946   412
6353  1215
 [46]   584   460   813  1624   911  1565  1718   663   563  1404   879  1730   707
2116   639
 [61]  1087   967  1806  5425   458  1850  2184   553  1359  1559  1532  6879  1701
530   782
 [76]   613  1039   506  5532   892  1971  4127  2711   411  1432   715   407   488
533   737
 [91]  2899  5715  2013   568  2058  1547  1531   536   581   844  4378  3982  6361
3217   948
[106]   966   830   608  1717   406 11210  4939  2871   570   446  1414  1480   648
1766   821
[121] 10492  4254  2508   395   557  1175  1067  1138  1325 41981  1981  4813  2005
1564  1741
[136]  2672  3661  8547 24749  1128   635  2152  1063  1143  1206 19128   636  1256
667  1000
[151]  4421  1438  1940  1708 32045  2591  9590 33137   966  1450  1785   703  1213
1537  4419
[166]   519   798  3961   629  1414  1722   831 13262  1596  7118 31902   886  2038
511   859
[181]  2392   692  1075   634  4614   805  1193  1024  5006   644  4436
```

Appendix I-6: Fatalities clustering

```
#Order the dataframe by ascending order
indicator1_c <- indicator1_c[order(indicator1_c$fatalities),]
#Split data by country and store in a class
library(purrr)
dat_list = split(indicator1_c, indicator1_c$country.code)

#Clustering by 5 the fatalities for each country
#1=Low,  5=High
library(cluster)
for (i in 1:49){
  clst <- pam(dat_list[[i]]$fatalities, 5, metric = "manhattan" )
  dat_list[[i]]$fatalities <- clst$clustering
}
```

Appendix I-7: Creation of the database of images

```
#--- Create database of images ------------------

# Initialise variables
cluster_centrality <- NULL

#------ Generate correlation data in CSV and images
# loop for the 5 clusters
for (k in 1:5){

  # Create sub data of countries corresponding to the cluster K
  indicator1_c1_k <- indicator1_c1 %>% filter(fatalities==k)
  # Remove from the K cluster dataset the 5 rows used for the calibration
  indicator1_c1_k <- indicator1_c1_k[!indicator1_c1_k$country.ind %in%
indicator1_c2$country.ind,]

  # Number of countries with cluster k
  nrow(indicator1_c1_k)

  for (i in indicator1_c1_k$country.ind) {
    print(k)
    print(i)
    # A subdataset is created with each single country data with k-cluster
joined with the calibration data
    x= rbind(indicator1_c2, indicator1_c1 %>% filter(country.ind==i &
fatalities==k))
    write.xlsx(x, file=paste0(k,'/data_raw/',i,'[',k,']','.xlsx'), sheetName =
i,
               col.names = TRUE, row.names = TRUE, append = FALSE)

    # Remove the text variables and fatalities and export the correlation matrix
in Excel and the graph
    y = subset.data.frame(x, select = -c(country, country.ind, country.code,
region, sub.region, intermediate.region, YEARS, fatalities))
    # Get the correlation matrix of the subdataset
    cor.matrix = cor(y, use = "pairwise.complete.obs", method = "spearman")
    cor.df = as.data.frame(cor.matrix)
    write.xlsx(cor.df, file=paste0(k,'/data_corr/',i,'[',k,']', '-
corr','.xlsx'), sheetName = i,
               col.names = TRUE, row.names = TRUE, append = FALSE)

    # Generate the correlation weighted graphs
    graph_pcor = qgraph(cor.matrix, graph = "pcor", layout = "spring",
edge.labels = TRUE)

    #Create centrality table
    centRes <- centrality(graph_pcor)
    centRes_val <- centRes$OutDegree
    centRes_val$country <- (x %>% filter(country.ind==i))$country
    centRes_val$years <- (x %>% filter(country.ind==i))$YEARS
    centRes_val$cluster <- k
    cluster_centrality <- rbind(cluster_centrality, centRes_val)
    write.xlsx(cluster_centrality,
file=paste0('centrality/','centrality_cluster5','.xlsx'), sheetName =
"centrality_cluster5", col.names = TRUE, row.names = TRUE, append = FALSE)

    # Save the JPEG device
    jpeg(file = paste0(k,'/data_img/',i,'[',k,']','.jpeg'), width = 800, height
= 600)

    # Plotting centrality indices
    centralityPlot(graph_pcor)

    # Close device
    dev.off()
  }
}
```

Appendix I-8: DCNN-11 model

```
model %>%
  layer_conv_2d(filters = 32, kernel_size = c(3,3), activation = 'relu',
input_shape = c(150, 150, 3)) %>%
  layer_max_pooling_2d(pool_size = c(2,2)) %>%
  layer_conv_2d(filters = 64, kernel_size = c(3,3), activation =
'relu')%>%
  layer_max_pooling_2d(pool_size = c(2,2)) %>%
  layer_conv_2d(filters = 128, kernel_size = c(3,3), activation =
'relu')%>%
  layer_max_pooling_2d(pool_size = c(2,2)) %>%
  layer_conv_2d(filters = 256, kernel_size = c(3,3), activation =
'relu')%>%
  layer_max_pooling_2d(pool_size = c(2,2)) %>%

  layer_flatten() %>%
  layer_dense(units = 512, activation = 'relu') %>%
  layer_dense(units = 5, activation = "softmax") %>%

  compile(loss = "categorical_crossentropy", optimizer = "adam",
          metrics = c('accuracy'))
```

Appendix J: Approval of thesis title change

Change project title - Mr Jean Brice Ghislain Tetka

ResearchUEL
To Jean Brice Ghislain TETKA

ResearchUEL

Change project title - Mr Jean Brice Ghislain Tetka

The ACE Research Degrees Sub-Committee on behalf of the Impact and Innovation Committee has considered your request. The decision is:

Approved

Your new thesis title is confirmed as follows:

Old thesis title: NETWORK ANALYSIS AND GRAPH CONVOLUTION TO ASSESS RISK OF VIOLENT SOCIO-POLITICAL CRISIS

New thesis title: RISK ASSESSMENT OF DEADLY ECONOMIC SOCIO-POLITICAL CRISIS WITH CORRELATIONAL NETWORK AND CONVOLUTIONAL NEURAL NETWORK

Your registration period remains unchanged.

Change project title - Mr Jean Brice Ghislain Tetka