International Journal on Optimization and Applications

VOL 02 - ISSUE 02
2022

Editor in chief
Prof. Dr. Hanaa HACHIMI

ISSN : 2737-8314

International Journal On Optimization and Applications

Vol 02 – Issue 02
2022

Editor in Chief
Prof. Dr. Hanaa HACHIMI

ISSN : 2737 - 8314

FOREWORD

The International Journal on Optimization and Applications (IJOA) is an open access, double blind peer-reviewed online journal aiming at publishing high-quality research in all areas of : Applied mathematics, Engineering science, Artificial intelligence, Numerical Methods, Embedded Systems, Electric, Electronic engineering, Telecommunication Engineering... the IJOA begins its publication from 2021. This journal is enriched by very important special manuscripts that deal with problems using the latest methods of optimization. It aims to develop new ideas and collaborations, to be aware of the latest search trends in the optimization techniques and their applications in the various fields..

Finally, I would like to thank all participants who have contributed to the achievement of this journal and in particular the authors who have greatly enriched it with their performing articles.

Prof. Dr. Hanaa HACHIMI
Editor in chief
Associate Professor in Applied Mathematics & Computer Science
Systems Engineering Laboratory LGS Director, BOSS Team
Sultan Moulay Slimane University

TABLE OF CONTENTS

Understanding Social Big data:

A literature review

Mahmoud HAOUAT

Phd student

Faculty of Legal, Economic and Social Sciences

Laboratory of Economics and Management of Organizations

Ibn Tofail University- Kenitra

Mahmoud.haouat@uit.ac.com

Mohammed QMICHCHOU
Professeur chercheur
Faculty of Legal, Economic and Social Sciences

Laboratory of Economics and Management of Organizations

Ibn Tofail University- Kenitra

qmichchou.mohammed@uit.ac.ma

***Abstract*—The advent and the popularity of both social media and ITC contributed to the birth of a new type of data called social data. The accumulation of this data gave birth to a new big data subgenre formally known as social big data. In the same way as big data, social big data is a new concept draw the attention of many studies. However, there is still a lack of consensus on its definition and also conceptualization. Hence the interest of our work, which aims to bridge the existing gap in the literature by providing a definition of the concept, its characteristics and also types. All of this was done based on a detailed analysis of previous works. In the same way as big data, it should be noted that the exploitation of social big data is very useful for several fields such as psychology, sociology, politics and business to achieve this firms therefore the second point addressed in this work, is how firms and users in general can extract value from all the social big data floating around them, to do so we put forward a series of actions and steps that falls under the name of social big data value chain, and also the main challenges encountered in these steps. With this paper we aim to foster future research activities around this concept.**

Keywords—Big data, Social big data, Social big data value, Digital human, Characteristics, Social media, Acquisition and preparation of data, Organization of data, Data Analysis, Data Exploitation.

I. INTRODUCTION

Since the start of the 21st century Big Data became one of the most discussed topics and terms in the world but at the same time it was the least understood.

Since its inception, the discussion around BD (Big Data) defined the concept using only qualitative terms. At that time BD was mostly defined as a larger than a certain number X of terabytes. This simple definition went through a lot of changes due to the many technological advances such as software tools, size of datasets, storage devices etc.… and also due to different fields and sectors where big data is used like business world, computer science etc.…

The rise of big data has coincided with two major events, the technological boom and the abundance of technological devices such as smartphones, computers etc. And also the birth and the development of a new communication tool called SOCIAL MEDIA. During the last two decades, social media completely reshaped not only the way people communicate but also the way they share and create information, these changes contributed to the birth of what researchers call "an always on society" where people constantly interact with each other thus creating an extensive amount of human generated diverse data called SBD (Social Big Data). According to Gandomi and Haider (2015) "such unstructured/semi unstructured yet semantically rich data has been argued to constitute 95% of all big data". Given its important and quantity that cannot be ignored, many researchers and theorists tried to study and theorize this emerging concept.

Broadly speaking SBD refers to a large amount of data generated through the use of social media, "the sheer volume and semantic richness of such data opens enormous possibilities for utilizing and analyzing it for personal, commercial as well as societal purposes" (Olshannikova et al, 2017).

II. MOTIVATION & METHODOLOGY

A. Motivation

The main goal of this literature review is to clarify and shed some light in the concept of social big data. To do so we will start by identifying the many ways SBD was defined and interpreted in the literature. We will also address the types of SBD and also the many benefits and challenges related to using SBD.

Based on what we said before, the value of this paper can be presented as follows:

- Firstly, we will examine the literature so we can bring clarity on big data and its characteristics.
- Secondly, we will aim to bring some clarity on various SBD concepts; we will also attempt to summarize the relation between SBD and many other fields. All of this will help us provide a synthesized definition of the concept and also identify its characteristics.
- Thirdly, we will focus on the various types of SBD.

— Finally we will address the way firms and users can create value des the collected data.

B. Methodology

Before stating which the mythology was followed in this work, it should be emphasized that the latter takes form of a literature review.

As we know, the literature review is considered the building block of almost all the academic research activities, regardless of the disciplines. It's usually used as way of collecting and synthesizing previous researches, and also a firm formation for advancing and facilitating any theory development. There are three different approaches to conduct a literature review: systemic, semi-systemic and integrative. Each of these approaches differs in their purpose, sample characteristics….

When it comes to our work we chose to follow the semi-systemic approach, this approach was designed is for "topics that have been conceptualized differently and studied by various groups of researchers within diverse disciplines" (Wong et al, 2013). Besides the aim of overviewing a topic, a semi-systemic review "often looks at how a research within a selected field has progressed over time or how a topic has developed across research traditions" (Snyder, 2019). Overall we choose this approach because it will help us understand all relevant research traditions that have implications for studied topic.

III. LITERATURE REVIEW:

A. Big data:

In this section, we will try to first present a list of popular definitions of big data, followed by a list of its essential characteristics. To do so we studied, reviewed and analyzed the related literature found on major databases.

- Definition:

While its ubiquitous big data as a concept has no formal and certain origin, despite its current popularity, there's no single unified definition. Fundamentally speaking, the term big data "applies to datasets that grows so large that they become awkward to work with using traditional datasets management systems" (Elgendy, Ebragal, 2014), this definition coincides with the one cited in the oxford dictionary. According to oxford dictionary big data is an "extremely large datasets that may be analyzed, computationally to reveal patterns, trends and associations especially relating to human behavior and interactions". This definition has since been reiterated by a number of other scholars, some of them even added new characteristics such as veracity, velocity, value etc...

When it comes to the literature, the quick and chaotic evolution of big data has contributed to the birth if several definitions, some of them focused on what big data is or represent, what others try to answer through it, what it does and even what kind of technology is needed to deal with it.

Amongst the most cited definitions, we found the one included in the Gartner report published in 2001. However the Gartner report doesn't mention the term "big data", but it was still considered as an authority in the matter. According to Gartner big data can be defined based on its characteristics known as Vs. The report in question proposed a 3 fold definition encompassing the 3 Vs: Volume, Velocity and Variety. As reported by Gartner "big data is high volume, high velocity and high variety information assets that demand cost effective, innovative forms of information processing that enable enhanced insight, decision making and process automation". Similarly, (Schroeck et al, 2012) considers big data as "a combination of Volume, Variety, Velocity and Veracity that creates an opportunity for organizations to gain competitive advantage in today's digitized marketplace", In the same way Laney (2001) offers the following definition "big data represents voluminous, high-velocity and varied information resources that require innovative forms of processing ...». Recently, the 3Vs method developed by Laney (2001) has been extended to 3 other Vs, including the value cited by (Chen et al, 2014 and Oracle), veracity (Arun and Jabasheel, 2014) and variability / complexity (IAS Inc.). The addition of these new characteristics prompted the birth of new definitions based on the technological aspect of big data, according to Provost and Fawcett (2013) big data represents "data sets that are too large for traditional data processing systems and that therefore require new technologies with names like Hadoop, Hbase, Mapreduce etc…".

The attributes and the technological one were not the only ways to define and describe the concept in hand; some definitions consider big data in terms of crossing a certain thresholds. In his definition Dumbill (2013) conveyed the multidimensionality of big data, by stating that big data is "data too big, moves too fast or doesn't fit the structure of the database". The other aspect mentioned in the literature is big data's impact. Several definitions highlight the huge impact that big data have not only on firms but also on society in general. Boyd and Crowford (2012) proposed the following definition, big data is "a cultural, technological, and scholarly phenomenon", in the same way Mayer-Schonberger and Cuckier (2013) described big data in terms of three main shifts in the way of analyzing the data in hand, the analysis of the data will improve the way we understand the society and the firms in question, according to Mayer Schonberger and Cuckier (2013), the shifts in question include a high volume of messier data and correlation.

The many definitions stated before and the ones stated in table 1 confirm what we stated earlier and go along with the observation made by Bi and Cohen (2014) among other authors.

- Summary of the definitions:

For the rest of this work, we will offer a definition based on the aforementioned definitions as well as observations of the literature. "**Big data is a complex large data set characterized by high velocity and a variety of information; this said data requires new technological tools (for storage, analysis) to substrate value".** This definition highlights the concept of 5Vs; the omission of veracity was done on purpose since it's often attributed to a specific type of data characterized by uncertainties and inconsistencies or lack of precision, example: data collected from social media.

Authors	Definitions
Boyd and Crawford (2012)	A cultural, technological and scientific phenomenon that relies on the interplay of analytical technology and methodology.
Manyika et al (2011)	Refers to datasets whose size is beyond the ability of typical database software tools to capture, store, manage, and analyze.
Zikopoulos et al (2013)	Big data contains four dimensions namely volume, variety, velocity and veracity.
Laney (2001)	Characterized by 3Vs theory: volume, variety, and velocity. Volume: with the generation and collection of data, data scale becomes increasingly big; Velocity: timeliness of big data, specifically data collection and analysis must be rapidly and timely conducted; Variety: the various types of data, which include semi-structured and unstructured, structured data

Table. 1: Some Big data definitions

- Big data characteristics:

The definition cited highlights the following characteristics: Volume, velocity, variety, value and complexity. These terms represent the characteristics of big data commonly known as 5Vs. However these dimensions are constantly developing.

— Volume:

Volume is the first dimension that we will discuss, it refers to the magnitude of data generated and collected (Chen et al, 2014), it is safe to say that the amount of data generated increases each day. That amount can exceed terabytes to reach petabytes, or even extabytes, and this growth is expected to continue over the next few years.

According to Hendler (2013) "the term volume originated to describe the amount of data held in large databases". This enormous amount of data is continuously created from multiple sources such as social media, clouds, business data and the Internet of Things which prompts us to ask the question what's big?, because what may be deemed big today may not meet the threshold in the future, since the storage capacities increase with time.

— Velocity:

Whereas volume refers to the size of data, velocity refers to the speed at which the data in question is generated and also the speed, at which it should be collected, analyzed and used. This was confirmed by Chen and Zhang (2014), they refer to velocity as "the speed characterizing incoming and outgoing data".

— Value:

This is the most important aspect of big data, and the purpose of big data technologies. It was introduced by Oracle as the defining attribute of big data. This view was also shared and expressed by IDC (International Data Corporation). According to them the big data architectures is designed to extract value from large volume of data.

— Variety:

This attribute is an essential characteristic of big data, it refers to different data's structures (structured, semi structured and unstructured) and also the many sources it comes from for example smartphones, social networks, clouds, sensors among many others, and also the formats and types images, videos, texts, audios……

— Complexity:

This dimension was introduced by SAS Inc. It was added in order to extend the dimensional model of big data. It refers to the fact that big data is generated through a multitude of sources, which poses a critical challenge when processing data (the need to connect, cleanse and transform this data into useful information).

According to Mikalef et al. (2018), various researchers focus on different aspects of big data, some authors introduced other characteristics such as veracity, visualization, variability.

Attribute	Definition
Volume	Volume represents the sheer size of the dataset due to the aggregation of a large number of variables and an even larger set of observations for each variable. (George et al. 2016)
Velocity	Velocity reflects the speed at which data are collected and analyzed, whether in real time or near real time from sensors, sales transactions, social media posts, and sentiment data for breaking news and social trends. (George et al. 2016)
Variety	Variety in big data comes from the plurality of structured and unstructured data sources such as text, videos, networks, and graphics among others. (George et al. 2016)
Veracity	Veracity ensures that the data used are trusted, authentic, and protected from unauthorized access and modification. (Demchenko et al. 2013)
Value	Value represents the extent to which big data generates economically worthy insights and/or benefits through extraction and transformation. (Wamba et al. 2015)
Variability	Variability concerns how insight from media constantly changes as the same information is interpreted in a different way, or new feeds from other sources help to shape a different outcome. (Seddon and Currie 2017)
Visualization	Visualization can be described as interpreting the patterns and trends that are present in the data. (Seddon and Currie 2017)
3Vs: volume, velocity, variety (Chen and Zhang 2014)	
4Vs: volume, velocity, variety, veracity (Zikopoulos and Eaton 2011; Schroeck et al. 2012; Abbasi et al. 2016)	
5Vs: volume, velocity, variety, veracity, value (Oracle 2012; Sharda et al. 2013)	
7Vs: volume, velocity, variety, veracity, value variability, visualization (Seddon and Currie 2017)	

Fig.1: The characteristics of big data, adopted from (Mikalef et al., 2018).

B. Social big data:

- Definition:

After defining big data in general, we will now move on to the main concept of our work, called social big data or social big data, as its name suggests, social big data represents the set of simple data and information collected from social media. When it comes to the literature there seems to be a lack of consensus on the general definition of social big data and associated terms. "In general, researchers focus on the analysis and use of social big data, having paid little attention to the concept and understanding of the associated phenomena" (Cambria et al, 2013). Faced with this dilemma it is therefore essential to describe and examine the literature in order to clarify this concept more and more. Therefore the goal of this section is to examine and review the existing literature on the concept by presenting and comparing the various definitions and approaches. While examining the literature, we noticed that social big data was defined and studied in several ways, some of these definitions were "simple" and others were considered "complex". These differences were due to many factors:

- The rapid and constant development of social media.
- The enormous growth of data generated by social media and social networks, approximately 2.5 Exabyte of data is created and shared per second.
- The various fields of research: social media, social networks, social computing etc.…

As its name suggests, social big data represents the set of data and simple or raw information collected via social media. In the literature there seems to be a lack of consensus on the general definition of social big data and associated terms. When it comes to the literature the first definition of social big data is attributed to Lazer and al (2009). According to them social big data represents "data collected from social media platforms". This definition was reiterated in different ways and by many authors such as Mukkamala and al (2014) among others. In a more detailed way Orgaz and al (2016) consider social big data as a "conjunction of two different concepts, social media and big data therefore SBD is a vast amount of data generated from multiple distributed sources but with an emphasis on social media". Another school of thought proposes a definition by focusing on the components of social big data; Tang et al (2014) consider social big data as "the fusion or union of three large parts, user data, and relationships social and generated content". Beyond the cited, definitions other theorists interpreted Social Big Data based on other concepts and approaches. For example, in his book "SBD mining", Ishikawa considers Social Big Data as a science and describes it as "the science of analyzing physical real world data (heterogeneous data with implicit semantics such as science data, event data, and transportation data) and social data (social media data with explicit semantics) by relating them to each other". In the same book Ishikawa also listed the characteristics of the concept. According to the latter since big data is characterized by 3 V's: Volume, Variety and Velocity, Social Big Data shares the same characteristics but he also added a fourth one named Vagueness.

The second approach sees big data as a big part of social computing (social computing is the field of computing that focuses on the interaction between social behavior and computer systems). The most prominent scholars of this school of thought are Guellil and Boukhalfa (2015). Based on the work of other authors such as Mukkamala (2014) and Nguyen (2015), Guellil and Boukhalfa (2015) were able to conclude that SBD is a direct synonym of social media data. Guellil and Boukhalfa provided the following definition, "SBD might be interpreted as a synonym of SMD (Social Media Data) with qualities such as large volume, noisiness and dynamism". In this context noisiness refers to the abundant spam found in the blogosphere as well as the existence of trivial posts on social networks. Even though Guellil and boukhalfa defined SBD, they did not propose any clear conceptualization of said concept. In the other hand, and following the same school of thought Mark coté (2014) attempted not only to define the concept but also to conceptualize it, by first of all distinguishing it from big data and second by defining it and showing its importance. According to Coté (2014) the difference between the two concepts resides in their sources, "big data is any data produced as the result of the quantification of the world that may include data from sensors, multiple industrial and domestic networks as well as financial markets, whereas BSD "comes from the mediated communicative practices of our everyday lives, whenever we go online, use our smartphone, use an app or make a purchase." As for its importance Coté (2014) argues that SBD is not a novel concept, and has a huge impact for many reasons, processing a huge amount of data can provide valuable information.

It should be noted that other authors and researchers believe that social big data plays an essential role in the analysis of enormous of data generated via social media. Among the authors we find Bello-Orgaz (2016). The latter insinuate that SBD represents "a process and methods that are designed to provide sensitive and relevant knowledge to any user or company from social media data sources". The collected data is characterized by their varied formats and content and also their very large size. Based on the work of Bello-Orgaz (2016), SBD incorporates three major different concept, big data (as a processing paradigm), social media (as the main source of data), and data analysis (as a method that will transform the collected data into relevant knowledge).

- Characteristics:

Since SBD is the result of the interaction between social media and big data, we can say that the two concepts share some characteristics (5Vs). However given the nature and the characteristics of social media, SBD is also characterized by Veracity and vagueness.

— Veracity:

This (recently added) feature does not yet have a uniform definition in the academic literature; each author uses / proposes a slightly different definition. The idea of veracity was used long before the birth of social big data and big data; it is deployed in several fields of research (biology, psychology, medicine, etc.). It alludes to the credibility of hypotheses and bibliographic resources used. We have to wait until 2012 that veracity was deployed in the context of social big data. Its first use is attributed to Snow (2012). The reasoning behind its incorporation was the complexity as well as the variety of the data collected and also the reliability of the sources.

When it comes to our research and social big data, there are several definitions of veracity. According to Orgaz et al (2016) it "refers to the accuracy and precision of information". This definition goes hand in hand with the one proposed by Bagiwa (2017), the latter considers veracity as "the disorder or reliability of the data, due to the large amount of data and its varied forms, the quality and precision are less controllable, for example Tweets or Facebook posts which are often characterized by abbreviations, typos and familiar speeches". The definitions do not end there, for example IBM and Microsoft have referred to "veracity as the fourth V, it represents the inherent unreliability of some data sources". For Gandomi and Haider (2015) "veracity refers to disorder and reliability of data", while for Storey and Song (2017) "veracity raises issues related to data quality". Other authors define veracity by citing associated dimensions such as precision, credibility, completeness, availability (Agrawal, 2012), consistency and accessibility (Corbellini et al (2017), integrity and authenticity (Denchenko et al 2013), reliability, authenticity, responsibility, availability (Mohan, 2016).

To conclude we can say that veracity refers to the accuracy, credibility and reliability of a set of collected data, it also refers to the integrity and objectivity as well as the reliability of the sources of these data especially when it's social media.

— Vagueness:

Vagueness is the new characteristic but also perhaps the trickiest one to define and address. In general vagueness refers to the indistinctness of existence in data. When it comes to big data vagueness represents the confusion over its meaning, nature, content availability, tools…. In the literature many authors tried to define vagueness as a characteristic such as Bonne (2014), Svetlana (2014), Panimabat et al (2017). When it comes to SBD, vagueness was introduced by Ishikawa (2015) in his book. Ishikawa (2015) maintains that vagueness "is a result of a combination of various types of data to be analyzed, which lead to inconsistency and deficiency. It also relates to the issues of privacy and data management as social data involves individuals' personal information. The proposed definition should not be mixed with the definition stated earlier that refers to the confusion over the meaning of big data.

When it comes to our work, we will focus on the definition made by Ishikawa.

- SBD types:

The rapid evolution of ICTs has completely transformed the role of users from a simple consumer to an active producer or mediator of information. This big change gave birth to a new type of users, "they are more in control of their profiles they can personalize/model the shared content according to their values, needs and preferences" (Cioffi-Revilla, 2013). These users are called "digital human».

The digital human represents the cornerstone of a society which balances between a physical world and a virtual world, during his interaction with the virtual world, the user gives birth to two types of data: data generated only by the machine and human generated data. Both of them can give users a fair share of social insight.

— Machine generated data:

This type data is considered the lifeblood of the Internet of Things. As its name suggests, machine generated data represents a type of data automatically generated by a computer process, application, or any other mechanism without any human intervention. The data generated by machine is characterized by being produced at very high rate and also by having no single form, type, format or metadata. According to Monash Research's Curt machine generated data can be defined as "data that was produced entirely by machines OR data that is more about observing humans than recording their choices". On the hand Abadi propose a more narrow definition of the concept, according to the author "machine generated data is data that is generated as a result of a decision of an independent computational agent or a measurement of an event that is not caused by a human action".

— Human generated data:

In general, the data generated by the human represents the content created following the interaction between the digital human (the user) and the social media, they exist in several forms created every day (email, audio file , video file, text, etc.). In the literature, human-generated data is divided into three subcategories: digital self-representation, Technology-mediated communication data and digital relationships data.

— Digital self-representation:

In what follows we will deal with the first type of social big data, even before defining this type, it should be noted that self-representation is not a new notion for human beings, because since its appearance humans have resorted to several methods in order to present themselves, starting with a simple painting in a cave, passing by well-detailed sculptures arriving at pictures / portrait photos. Today in the digital context, self-representation has retained its importance, it is considered the first step taken by the user (the digital human) in order to communicate and socialize with other users.

In the literature digital self-representation is defined in several ways, according to Warburton (2010) "digital self-representation constitutes a part of individual identity, and in an increasingly digitalized environment it represents an increasing part in sources of access and possible knowledge of an individual". For their part, Boyd and Heer (2006) consider digital self-representation as "the most frequent strategy in online participation and communication, this strategy is based on data such as virtual profiles, user-published content or a community".

— Technology-mediated communication data:

This type of social big data represents all the data generated from communications via social media, thanks to these communication platforms this type of data is easily created. In the literature there is a very minimal number of definitions of this type. According to, Olshannikova et al (2017) this type of data represents, "the data generated during two-way communication or during the collaborative creation of knowledge during the distribution of knowledge or information via social media platforms".

— Digital relationships data:

Social media platforms such as Facebook and Twitter offer their users the chance to create virtual communities based on already existing connections in the physical world or with other users (virtual connections), this characteristic has given rise to a type of data called digital relationships data, this type of data describes the links between and the implicit and explicit relationships between users of social media.

The analysis of this data offers a general and deep insight into social relationships and structure, as well as a better understanding of social phenomena, this information can be exploited to better know and target customers.

C. Unit Social big data management:

The proliferation of social big data allows the creation of added value in various areas, ranging from optimizing customer relation to significantly increasing margins. According to Stieglitz et al (2018) "social media data can be analyzed to gain insights into issues, trends, influential actors and other kind of information". In order to create and generate the value needed, companies and organizations must use and master several techniques and technologies essential for handling and extracting the coveted information from the data at their disposal. These techniques and technologies fall under what we call data management.

Basically data management "is seen as the gathering, processing, management of data producing new information for end users" (Emani et al, 2015; Krishnan, 2013). However, despite its benefits there's little to no research that focus on the management of SBD even though the latter has always been the order of the day for researchers and practitioners, this is not the case for big data management. The latter was always discussed and covered by many authors and researchers such as Karmasphere (2011), Mork and Miller (2013), Curry (2014), Gandomi and Haider (2015) etc..... Each of these authors proposed its own version of big data management process under the name "big data value chain". The proposed value chains have small differences (that can be overlooked) but also have several points such as the 4 big steps that should be followed: Acquisition, Organization, Analyze and Decision each of these steps contain many actions.

When it comes to SBD, and since it's considered as a subgenre of big data, we can say that the four steps cited before are also applicable here. However and considering its nature and characteristics it's only right to add more steps.

- Social Big Data value chain:

— Big data value chain:

The value chain was created thanks to the work of Porter (1985), "the value chain represents a series of activities that can lead to the creation of value", and based on this principle several authors, researchers and organizations have tried to apply the value chain in their respective fields. Rayport and Sviokla (1995) are considered among the first to adopt the value chain in another domain precisely to information systems. Based on the logic of Porter (1985) and Rayport and Sviokla (1995), Karmasphere (2011) a company specializing in the optimization of contact points proposed its own interpretation of the value chain in the field of data and big data, the merger between the two concepts has given birth to what is called - big data value chain -. The latter is considered by the European Commission as "the cornerstone of the economy of the future, the knowledge economy".

Karmasphere (2011) proposed a framework that can identify the key elements of BD value Chain, this framework is extremely focused on the processing of data already acquired and stored is split into four major stages called the 4As: acquisition, assembly, analysis and finally action. These steps can be segmented into:

- Data management: This category is made up of the following tasks, acquisition and assembly, it is considered as the basic element of the chain value, and during this step the data is retrieved / extracted from several sources and organized according to its types: structured, unstructured, semi-structured data.
- Data analysis: In turn, this category is made up of the following tasks: analysis and action. During this phase companies begin to extract value from the data collected, the extracted value will serve as a basis for decisions.

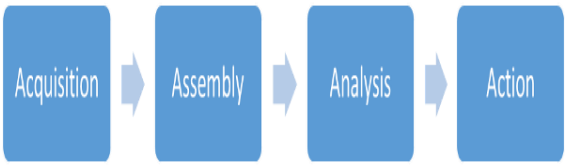


Fig.2: Value chain according to Karmasphere (2011)

Since this framework relegates the analysis, cleaning and filtering of data to the background, it is often judged by theorists to be too limited.

In the same way as Karmasphere (2011), several theorists and organizations have formulated their own versions of the big data value chain, according to Labrindis and Jagadish (2012) have proposed a model composed of five steps; these same steps constitute the two main sub-processes, data management and data analysis.

- Data management: involves the main things that enable the company to acquire, store and prepare data.
- Analysis: refers to the techniques used for the analysis and creation of knowledge and value.

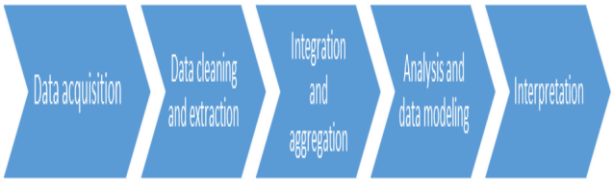


Fig.3: Value chain according to Labrindis and Jagadish (2012).

Unlike the Karmasphere model, the model here emphasizes the phase of integration (or consolidation) and cleaning of data.

- Data cleaning: This is the operation of detecting, correcting or even deleting the various errors that exist on the data acquired and stored. This

operation can be carried out using several technologies which will be mentioned below.

- Data integration: This is the operation that involves combining data from multiple sources (social media, internet of things, traditional data…), in order to provide users with a new idea and a benefit. It should be noted that non-compatible data is not used and is discarded.

Other theorists use the term pipeline to refer to the chain value, according to Gandomi and Haider (2015) the pipeline is made up of six major stages: data acquisition, data extraction and cleaning, data integration and aggregation. (These three steps represent the data management process), while the analysis process has the following two steps: data analysis and interpretation.

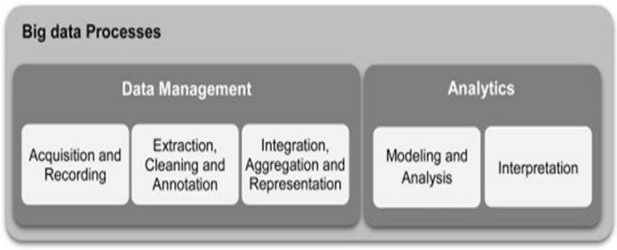


Fig.4: Processes for extracting insights from big data according to Gandomi and Haider (2015)

.

- SBD value chain:

Based on the analysis of the frameworks mentioned above, it is obvious that certain steps are crucial to extracting value; these same steps will serve as a basis for us in the development of our value chain.

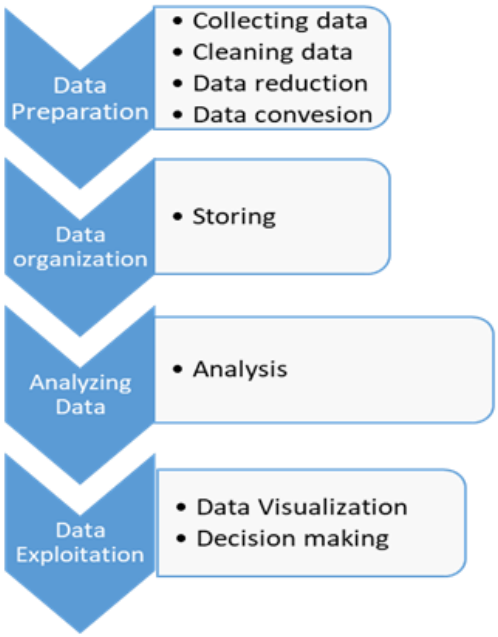


Fig.5. Social Big Data value chain:

— Acquisition and preparation of data:

The preparation of the data represents the first phase of our value chain, it is considered particularly crucial for the analysis of SBD and the creation of value, given the nature and characteristics of this type of data. The objective of this phase is to collect and acquire high speed data from different social media sources by using APIs and WebCrawler. One the data has been collected, firms can now process raw datasets, improve its quality, verify its veracity and limit the disclosure of private and sensitive information as much as possible. Once these tasks are completed the firm can now store the data.

— Organization:

This phase represents the second phase of the value chain; it's mostly composed of a single major task "storing". The purposes if this phase is to store, manage and organize data in an efficient way (even when it's subjected to a heavy load of data) in order to facilitate the next phase. Organization is considered one of the most complex aspects of the social big data world, because unlike other elements of the process, this step does not only rely on software but it also requires a significant infrastructure. The latter should be able to deal with various data formats and store it in the right location and also must be able to parse them and extract the actual information like named entities, relation between them, etc.

It should be noted, that a storage system is said to be efficient when it takes into account in one hand, one of the factors proposed by Brewer (2000): Consistency, Availability and Partition of Tolerance. And on the other hand when it offers users simultaneity the chance to store an unlimited amount of data and manage a high rate of random access to said data.

— Analyze:

During this phase, the collected data is subjected to several techniques in order to extract information deemed necessary to facilitate decision making.

The collected SBD data in itself has no value and requires deep analysis in order to extract and acquire information from the collected data hence the critical status of this step. Some refers to this phase as the element that bridges the gap between data and knowledge.

In the literature, it is defined in several ways. According to Davenport and Harris (2007) it represents "the intensive use of data, quantitative statistical analysis, explanatory and predictive models as well as factual management to guide decisions and actions to be taken". On the hand for (Gudivada et al, 2016), it represents "any actionable information that results from computational analysis of data using mathematical and statistical methods. Data analytics is an interdisciplinary domain encompassing mathematics, statistics, and computer science". From these definitions it is evident that this phase is interdisciplinary and multidimensional, it involves the use of disciplines such as statistics, operational research, information systems to name a few examples.

SBDA is generally categorized into five major categories that differ from each other in complexity, value and techniques used: method that describes what happened, method that describes why it happened, method that describes what will happen, method that facilitates obtaining the desired results and finally the one that describes the best action or decision to be taken. These methods are respectively named as follows: Descriptive, diagnostic, predictive, prescriptive and finally cognitive analysis.

Analytics	Définitions	Technics used
Descriptive	Based on the exploitation of historical data in order to identify patterns and create management reports based on past behavior. (Assunçao et al, 2015)	▪ Standard reports and dashboards ▪ AD / HOC reporting ▪ OLAP ▪ Line graph
Diagnostic	An application of data analytics to investigate the causes and effects of situations. (Fleckenstein et al, 2018).	▪ Correlations ▪ Drill Down
Predictive	Use Data to identify Past patterns to predict the future (Bagiwa, 2017).	▪ Data mining ▪ Text mining ▪ Web / media mining
Perspective	Uses data and mathematics to determine a set of high-value alternative actions given a complex set of objectives, requirements, with the goal of improving business performance (Damirkan and Delen, 2013).	▪ Machine Learning ▪ Neural networks ▪ Prescriptive dashboards
Cognitive	"The natural evolution of both data mining and visual analytics. It removes humans from the loop and is completely Automated, it combines the computing and cognitive science approaches" (Gudivada et al, 2016).	▪ Deep Learning ▪ Computer learning systems ▪ Pattern recognition

Table. 2: 5 Types of Analytics

— Decision:

It represents the last phase of our value chain, during this phase the user interprets the results found during the previous phase, then it relies on these interpretations in order to make the appropriate decisions. It is generally made up of two tasks: data visualization then interpretation / decision making.

▪ Visualization:

Broadly speaking, it can be described as the science of analytical reasoning facilitated by static or interactive visual interfaces, to be precise it represents an iterative process that involves information gathering, data preprocessing, knowledge representation. In the literature visualization is defined by several authors, according to Batrinca and Treleaven (2015) "it is the visual representation of data via diagrams in order to communicate information in a clear, efficient and compact way", according to Nasser and Tariq (2015) "visualization reveals so-called hidden patterns and patterns as well as unknown correlations to improve decision-making".

In general, graphs and dashboards are the most used techniques (for decades) to synthesize data in a coherent, compact and understandable format. However, as the volume of data increases, traditional visualization techniques cannot handle this huge volume, which has given rise to a new field of advanced visualization that uses interactive methods to represent thousands or even millions of points.

▪ Decision making or interpretation:

Decision-making is the last link in the value chain, this step consists in determining the necessary actions to be taken based on the gathered and visualized results, it includes an interpretation and a critical evolution of the reports, diagrams tables, while taking into account the limitations, the validity and the reliability of the methods used. In the literature it is considered to be the phase with the greatest value in the chain.

- Challenges related to the use of SBD value chain:

In order to make the best use of the value chain, companies face several challenges. These challenges can be categorized according to the phases of the chain as well as the tasks to be accomplished.

— Phase 1 : data acquisition and data preparation:

In the quest for data acquisition and preparation, companies face several challenges. Each of these challenges is directly related to the task to accomplish.

▪ Data collection:

In this first task, companies face primarily technological challenges among them we find:

- ✓ Difficulty in collecting, acquiring, deciphering data related to experiences, due to the irregularity as well as to the diversity of the language used, the latter often contains an informal language (sarcasm, acronyms, spelling mistakes), this content is often ambiguous and subject to human interpretation, not algorithms.
- ✓ Difficulty in deciphering opinions/feelings. In order to overcome this obstacle, it is necessary to establish lists of pseudo codes/terms one which designates positive feelings/opinions while the other designates negative feelings/opinions.
- ✓ Despite their necessity, APIs prove to be quite limited in terms of space and the number of units allowed, for example YouTube data API sets a limit of 30,000 units/users, while the total quota per day is set at 50,000,000 units, for its part Twitter allows only 15 requests per minute. These limits prevent a large amount of data from being obtained.

- Data cleaning:

When it comes to data cleaning firms face diverse set of challenges some of them were mentioned in the literature, according to (Freitas and Curry, 2016) "the central challenge of cleaning models is to manage the long data trail and improve the extensibility of data curation». In addition to this challenge, practitioners distinguish many others such as the high costs of curation projects, the time constraint this is directly related to the volume and variety of data collected as well as to security.

- Data reduction and data conversion:

The most obvious challenge is that of technology, companies must at all costs update IT and technological services in order to cope with the complexity of the task and the complexity of the data in their hands this in order to avoid any corruption that can affect data. The other challenge that companies face is directly related to redundancy elimination methods, deleting non-duplicate data, in order to prevent this from happening, it is recommended to create duplicate media.

Another concern is data security due to open source programs, adding to this the loss of the original data format, which in some cases poses problems of compliance with legal constraints.

— Phase 2 : data organization:

Despite the immense technological progress the world has experienced, companies face several challenges:

- Privacy and Security:

Data storage is no longer the main challenge faced by companies following the birth of Cloud computing. The main threat encountered is data privacy and security; if ever the systems are compromised personal data can be disclosed. Therefore, it is of utmost importance to secure and protect data against threats.

- Integrity check:

This challenge is directly linked to the use of the Cloud. When businesses use this option, they lose full or partial control of the data. In this case the data is outsourced and always remains at risk. Verification is of utmost importance, this task can be carried out by companies or by third parties.

- Cost:

The cost of this operation is always high, whether the companies use disks or systems. It should also be noted that the transfer of data to the Cloud as well as their hosting remains very expensive due to the excessive volume.

— Phase 3 data analysis:

In order to apply this operation correctly, companies face certain challenges, according to Ahmed and Ji (2013), companies face the following challenges: Budget and investment costs, also the data availability. Other authors cite data security and confidentiality as the main challenge faced by companies, as well as the talent gap (even if the field of big data is growing; there is still a lack of experts who can perform these tasks quickly, efficiently and correctly).

— Phase 4 data exploitation:

Companies face several challenges when carrying out this step:

- The difficulty in making decisions, since data is collected from social media, the latter are constantly developing. In this case the platforms must be constantly monitored in order to have a major margin of time to take the necessary measures, it is necessary to use real time big data analytics.
- Overestimation of the accuracy of the analysis.
- The use of erroneous data.

IV. CONCLUSION:

The goal of this work was to introduce and study social big data as a separated concept in all its multidisciplinary and multidimensional nature, to accomplish this task we decided to divide our work in two chapters. In the first chapter we defined social big data based on an in depth literature search, while in the second we tried to tackle social big data management.

When it comes to the first chapter, our literature overview showed us the existence of two different ways of treating and studying social big data. The first stream of thought tries to define the concept in its own rights, while the other focuses more on the analysis of social big data and the concepts related to it. These differences led to a lack of consensus and a little bit of vagueness. In order to remedy this problem we tried to define the concept in question based on the other definitions and also our own observations. Based on the proposed definition we outlined social big data's characteristics. Finally, to close the chapter we proposed a brief classification of the concept.

As for the second chapter, it was entirely devoted to social big data management; in this chapter we presented the concept of social big data chain value. This chain value covers both the entire data lifecycle, and also how companies and users can create value from the collected data. In general, the chain value is made up of four main phases each of which play a different and specific role in the value creation process: data preparation (discovery of data), data organization (discovery and preservation of data), analysis (creation of value), and exploitation (realization of value). During the same chapter we defined each phase and identify every challenge the users might face. It should be noted that these challenges vary from one organization to another.

In summary with this paper we aim to draw researchers to this concept in the hopes of developing a better conceptualization and understanding of social big data as a concept.

REFERENCES

[1] Abadi Daniel, 2010. Machine vs. human generated data. (http://dbmsmusings.blogspot.com/2010/12/machine-vs-human-generated-data.html).

[2] Ahmed, Zafor and Ji, Shaobo, Business Analytics: Current State & Challenges. CONF-IRM 2013 Proceedings. 12. 2013.

[3] K. Arum, L. Jabasheela, Big Data: Review, Classification and Analysis Survey, International Journal of Innovative Research in Information Security, Vol. 4, N°3, p.17-23, 2014.

[4] M.D. Assunção, R.N. Calheiros, S. Bianchi, M.A.S. Netto, R. Buyya, Big Data computing and clouds: Trends and future directions, Journal of Parallel and Distributed Computing, Vol. 79, p.3-15, 2015.

[5] L.I. Bagiwa, Big Data: Concepts, Approaches and Challenges, International Journal of Computer Networks and Communications Security, Vol. 5, N°8, p.181–187, 2017.

[6] B. Batrinca, P.C. Treleaven, Social media analytics: a survey of techniques, tools and platforms, AI & SOCIETY, Vol. 30, N°1, p.89-116, 2014.

[7] G. Bello-Orgaz, J.J. Jung, D. Camacho, Social big data: Recent achievements and new challenges, Information Fusion, Vol. 28, p.45-49.

[8] D. Boyd, J. Heer, Profiles as Conversation: Networked Identity Performance on Friendster, Proceedings of the 39th Annual Hawaii International Conference on System Sciences (HICSS'06), Vol. 3, 2006.

[9] D. Boyd, K. Crawford, CRITICAL QUESTIONS FOR BIG DATA, Information, Communication & Society, Vol. 15, N°5, p.662-679, 2012.

[10] L. Burkholdee, "PHILOSOPHY AND THE COMPUTER. ROUTLEDGE". [S.l.], 1992.

[11] E. Cambria, D. Rajagopal, D. Olsher, D. Das, Big Social Data Analysis, Big Data Computing, p. 401-414, 2013.

[12] C.L. Philip Chen, C.Y. Zhang, Data-intensive applications, challenges, techniques and technologies: A survey on Big Data, Information Sciences, Vol. 275, p.314-347, 2014.

[13] M. Chen, S. Mao, Y. Liu, Big Data: A Survey, Mobile Networks and Applications, Volume 19, N°2, p.171-209, 2014.

[14] C. Cioffi-Revilla, Introduction to Computational Social Science. Springer International Publishing, 2017.

[15] M. Coté, Data Motility: The Materiality of Big Social Data, Cultural Studies Review, Vol. 20, N°1, 2014.

[16] T.H. Davenport, J.G. Harris, "Competing on Analytics: The New Science of Winning", Harvard Business Review Press, Boston, 2007.

[17] Dbms2.com. 2021. Examples and definition of machine-generated data | DBMS 2: DataBase Management System Services. [online] Available at: <http://www.dbms2.com/2010/12/30/examples-and-definition-of-machine-generated-data/> [Accessed 17 October 2021].

[18] D. Delen, H. Demirkan, Data, information and analytics as services, Decision Support Systems, Vol. 55, N°1, p.359-363, 2013.

[19] E. Dumbill, Making Sense of Big Data, Big Data, Vol. 1, N°1, p.1-2, 2013.

[20] N. Elgendy, A. Elragal, Big Data Analytics: A Literature Review Paper, Advances in Data Mining. Applications and Theoretical Aspects, p.214-227, 2014.

[21] C.K. Emani, N. Cullot, C. Nicolle, Understandable Big Data: A survey, Computer Science Review, Vol. 17, p.70-81, 2015.

[22] A. Freitas, E. Curry, Big Data Curation, New Horizons for a Data-Driven Economy, p.87-118, 2016.

[23] M. Fleckenstein, L. Fellows, Modern Data Strategy, 1st ed., Cham, Switzerland, p.133, 2018.

[24] A. Gandomi, M. Haider, Beyond the hype: Big data concepts, methods, and analytics, International Journal of Information Management, Vol. 35, N°2, p.137-144, 2015.

[25] V.N. Gudivada, M.T. Irfan, E. Fathi, D.L. Rao, Cognitive Analytics: Going Beyond Big Data Analytics and Machine Learning, Handbook of Statistics, p.169-205, 2016.

[26] I. Guellil, K. Boukhalfa, Social big data mining: A survey focused on opinion mining and sentiments analysis, 12th International Symposium on Programming and Systems (ISPS), 2015.

[27] J. Hendler, Broad Data: Exploring the Emerging Web of Data, Big Data, Vol. 2, N°1, p.18-20, 2013.

[28] H. Ishikawa, "Social big data mining", CRC Press, Boca Raton, Florida, 2015.

[29] D Laney, 3D Data Management: Controlling Data Volume, Velocity, and Variety, Gartner, file No.949, 6 February 2001.

[30] D. Lazer and al, Computational Social Science, Science, Vol. 323, N°5915, p.721-723, 2009.

[31] J. Manyika, J. Chui, M. Brown, et al, Big Data: The Next Frontier for Innovation, Competition, and Productivity, McKinsey Global Institute, 2011.

[32] V. Mayer-Schönberger, K. Cukier, "Big Data: A Revolution that Will Transform how We Live, Work, and Think", Houghton Mifflin Harcourt, Boston, 2013.

[33] P. Mikalef, I. Pappas, J. Krogstie, M. Giannakos, Big data analytics capabilities: a systematic literature review and research agenda, Information Systems and e-Business Management, Vol. 16, N°3, p.547-578, 2017.

[34] R.R, Mukkamala, A. Hussain, R. Vatrapu, Fuzzy-Set Based Sentiment Analysis of Big Social Data, 2014 IEEE 18th International Enterprise Distributed Object Computing Conference, 2014.

[35] T. Nasser, RS. Tariq, Big Data Challenges, Computer Engineering & Information Technology, Vol. 04, N°03, 2015.

[36] D.T. Nguyen, D. Hwang, J.J. Jung, Time-Frequency Social Data Analytics for Understanding Social Big Data, Intelligent Distributed Computing VIII, p.223-228, 2015.

[37] E. Olshannikova et al, Conceptualizing Big Social Data, Journal of Big Data, Vol. 4, N°1, 2017.

[38] M.E. Porter, "Competitive advantage", Free Press, New York, 1985.

[39] F. Provost, T. Fawcett, Data Science and its Relationship to Big Data and Data-Driven Decision Making, Big Data, Vol. 1, N°1, p.51-59, 2013.

[40] M. Schroeck, R. Shockley, J. Smart, et al, Analytics: The realworld use of big data. New York, NY: IBM Institute for Business Value, Said Business School, 2012.

[41] H. Snyder, Literature review as a research methodology: An overview and guidelines, Journal of Business Research, Vol. 104; p.333-339, 2019.

[42] S. Stieglitz et al, Social media analytics – Challenges in topic discovery, data collection, and data preparation, International Journal of Information Management, Vol. 39, p.156-168, 2018.

[43] V. Storey, I.Y. Song, Big data technologies and Management: What conceptual modeling can do, Data & Knowledge Engineering, Vol. 108, p.50-67, 2017.

[44] J. Tang, Y. Chang, H. Liu, Mining social media with social theories, ACM SIGKDD Explorations Newsletter, Vol. 15, N°2, p.20-29, 2014.

[45] S. Warburton, "Digital Identity Matters", London: King's College London, 2010.

[46] Wong, G., Greenhalgh, T., Westhorp, G. et al. RAMESES publication standards: meta-narrative reviews. BMC Med 11, 20, 2013.

[47] P. Zikopoulos, R.B. Melnyk, "Harness the power of big data", New York, 2013.

Feedback protocols for anti-angiogenic therapy in the treatment of cancer tumors by chemotherapy

1st Amine Moustafid
Department of Mathematics and Computer Science, University Hassan II, Faculty of Sciences Aïn Chock, Casablanca, Morocco.
a.moustafid@gmail.com 0000-0002-9121-0745

***Abstract*—In this paper we deal with cancer control via a treatment model. A set-valued control method is used, to design the procedures leading to the formulation of model protocols, by which cancer cells are eradicated. These protocols are provided as selections of specific set-valued maps, depend upon the initial stage of cancer, and divide into two inconsistent types responding to two concerns: smoothness of treatment or minimal-dose therapy.**

***Index Terms*—Anti-angiogenic therapy, Chemotherapy, Set-valued analysis, Viability theory**

I. Introduction

Mathematical modelling of cancer, whether by ordinary differential equations (ODEs), delay differential equations [24], [31], partial derivatives equations (PDEs) [15], or even stochastic differential equations [18], [19], [23], is useful not only to gain a broad understanding of the tumor dynamics, but also to investigate the problems associated to control cancer.

Several studies address the subject of cancer control, by evoking different approaches, depending upon the model type. All of them aim at seeking protocol laws to destroy cancer cells, taking into account patient quality of life that can be described by constraints on both healthy cells and the administered drug doses.

For ODE models, as considered in this paper, numerous studies use optimal control techniques, which consist of designing a suitable criterion that involves minimizing cancer cells, for examples:

- [4] controls the model in [3], in order to attain the goal of reducing the amount of cancer cells, with acceptable consequence on other states, by two different optimal strategies: closed-loop SDRE (State-Dependent Ricatti Equation), and open-loop method utilizing steepest descent technique, the results of each are compared together, to figure out which one is the better choice.
- [40] presents optimal control method results for four different cancer models based on two sets of ODEs, and contain either chemotherapy, immunotherapy, anti-angiogenic therapy or combinations of these therapies, the optimal control problems in questions are solved numerically with Bock's direct multiple shooting approach.
- [17] uses Pontryagin's Maximum Principle to give an optimal control of bang-bang form.
- [16] illustrates and discusses three different approach for the control function: continuous, impulsive and hybrid.
- [29] uses for a well known model of the tumor-immune system interaction [5], an optimization algorithm to find optimal protocols as impulse-like drug administrations.

Much less frequently, there are studies using various different approaches:

- ◦ [39] uses an asymptotic approach to give sufficient conditions on both model and treatment parameters [5], under which all trajectories in the positive orthant tend to the tumor-free equilibrium point.
- ◦ [35] performs a quantitative analysis to discusses the impact of delay in immunotherapy with interleukin-2, at different antigenicity levels.
- ◦ [7] studies by the qualitative theory of differential equations the immunotherapeutic models of Kuznetsov and Stepanova, in meta-models form (i.e., family of models), and gives a general result on global eradication of cancer under immunotherapy.
- ◦ [34] proposes a generalization of d'Onofrio's background model, where one parameter is assumed to be time dependent, keeping all other parameters constant, and done a qualitative-asymptotic analysis for the second-order Taylor expansion of the model.

The approach using techniques of viability theory and set-valued analysis has been initiated since [2], by providing feedback protocols for immunotherapy model of [5], over a finite horizon. This set-valued approach is performed in [1], with respect to a wide class of ODEs. In the following, we give a concise overview along with main facts that is noteworthy to mention:

- ∗ The provided protocols aim at asymptotically eliminate cancer cells, taking into account constraints that may involve both injected drugs and cells in competition with cancer cells.
- ∗ They are given in feedback form as selections of adequately designed multifunctions, leading to two incompatible sets of protocol laws: one set is constituted of continuous laws, for a smooth therapy, and another set involves the minimal law for minimum drug doses, to minimize their side-effects. The former laws can easily be derived through universal formulas, while the latters need solving a convex quadratic program. Both laws have to be used depending on medical preferences.

- Further the therapy protocol design, the approach allows for a slight representation of *cancer staging* as known in medicine, providing mathematical criteria to categorize initial cancer stages given both their curability and the kind of protocols that can be formulated. Thus by this vision, a real medical strategy can be followed after diagnosis of the disease. In the case of stage 1, the applied medical protocols lead to decreasing cancer cells in respect of patient quality of life. Stage 2 describes cancers that can be amenable to stage 1 in a finite horizon, only after what decreasing cancer cells are guaranteed. While stage 3 characterizes cancers that are incurable, corresponding to late-stage cancers. See technical details in Section II.

In [6] we employ the previous set-valued method to address two cancer therapies: namely anti-angiogenic therapy, through its descriptive model established in [8], and chemo-immuno therapy, as treated in [9].

Newer approaches that combine anti-angiogenic drugs with chemotherapy, other targeted drugs, or radiation may work better than using them alone (mono-therapy). Coupled anti-angiogenic and chemotherapy promote a larger reduction of the tumor than use chemotherapy alone [3]. Along the same lines: [32] optimized and rationalized the association between anti-angiogenic and cytotoxic drugs in the treatment, to improve the anti-tumor efficacy, and [41] founded that metronomic schedules are more effective in eliminating tumour cells mainly due to their chemotherapeutic action on endothelial cells and that more frequent, low drug doses also entail outcomes in which the survival time of patient is increased, and [42] used Hamilton-Jacobi-Isaacs (HJI) partial differential equation, to derive a robust state feedback control of the combined therapy of cancer using chemotherapy and angiogenic inhibition, which guaranteed tumor contraction maps as a function of the initial state of tumor and the vasculature capacity. Other mixed therapies prove their capacity to eliminate the entire tumor than either single modality treatment, so it is in [11] which extracts situations where neither chemotherapy nor immunotherapy alone are efficient to control tumor growth. As well in [37] where such combination treatments have been suggested as a promising alternative to mono-therapy. In the study [36] both treatments concurrently is favorable, due to the lower toxicity and greater immune stability. For this combination therapy, as mentioned earlier, we already approach it by means of chemo-immuno model [9], so in article [6]. While [43] investigates how virotherapy could enhance chemotherapy. A comprehensive review of mathematical models for combination cancer therapy has been carried out in [44]. A Multi-Objective approach identifies the effective combination of therapeutic targets in cancer cells [45].

For instance we choose as subject for the method by [1], the model developed in [3], which combines between anti-angiogenic and chemotherapy. Based on the competition between normal cells and cancer cells as well as endothelial cells associated to angiogenic process, which helps cancer cells to growth. The effects of anti-angiogenic and chemotherapy agents are obviously included in the model.

The current paper is organized as follows: Section II is an overview of the set-valued control method developed by [1], in Subsection III-A, a detailed description of model of [3] is given, Subsection III-B is an application of the method to the model, this is followed in Subsection III-C with numerical examples and interpretations. We conclude in Section IV with a discussions.

II. The set-valued control method

This section is dedicated to summarize the general method developed in [1], and which we are going to apply on Pinho's model. It deals with the specific class of ODEs

$$\dot{x} = f(x,\tau) + B(x,\tau)u, \tag{1a}$$
$$\dot{\tau} = \tau\psi(x,\tau), \tag{1b}$$

conditioned initially at $t = 0$ by

$$x(0) = x_0, \tag{1c}$$
$$\tau(0) = \tau_0. \tag{1d}$$

- The state (x,τ) evolves in $\mathbb{R}^n_+ \times \mathbb{R}_+$:

$$\forall t \in [0,\infty[, (x(t),\tau(t)) \in \mathbb{R}^n_+ \times \mathbb{R}_+. \tag{2}$$

  This positivity condition represents the biological feasibility of interactive cells densities x_i, and tumor cell burden τ. Note that (1b) involves that whenever $\tau_0 = 0$ then there are no cancer cells in the tissue for all time. This is due to formula,

$$\bar{\tau}(t) = \tau_0 \mathrm{e}^{\int_0^t \psi(\bar{x}(s),\bar{\tau}(s))\mathrm{d}s}, \quad \text{for all } t \geq 0,$$

  for any solution $(\bar{x},\bar{\tau})$ of system (1).
- The control $u = u(t)$ evolves in the constraints subset of $\mathbb{R}^p$, defined by

$$K_p = [0, u_1^{\max}] \times \cdots \times [0, u_p^{\max}], \tag{3}$$

  where the components u_i stand for the doses of treatments, that can be administered continuously by some kind of portable pumps or straps to the patient's blood circulation, with the maximal tolerated doses $u_i^{\max}$.
- The functions f and ψ map $\mathbb{R}^n_+ \times \mathbb{R}_+$ into $\mathbb{R}^n$ and $\mathbb{R}$ respectively, ψ is supposed to be of class $\mathcal{C}^1$ on $\mathbb{R}^n_+ \times \mathbb{R}_+$, while $B(x,\tau) \in \mathbb{R}^{n\times p}$, and $B'(x,\tau) \in \mathbb{R}^{p\times n}$ denotes its transpose.

We remark that the right-hand side of (1a) is linear with respect to the control term u, while in the tumor dynamics (1b), the tumor cell burden τ can be extracted, this structure of the prototype (1) is well justified by a vast majority of ODE models in the literature [10]–[14], [22], [27], [28], [30], [33], [38], including the one of [3].

But there exists another varieties of models, like those using a direct control on cancer cells, in [20] for example, cancer cells can be killed by an external injection of ACI (Adoptive Cellular Immunotherapy), depicted by the control function $\epsilon_2 = \epsilon_2(t)$, at a rate proportional to the product of the

tumor cell burden, nevertheless, to approach this type of ODE models, we can augment the general model 1 by a suitable differential equation (for e.g., $\dot{\epsilon}_2(t) = -\epsilon_2(t) + v(t)$, where $v(t)$ is the auxiliary control).

Definition 2.1: We call a protocol any control $\bar{u}\colon [0,\infty[\to K_p$, such that system 1 has a solution $(\bar{x},\bar{\tau})\colon [0,\infty[\to \mathbb{R}^n_+ \times \mathbb{R}_+$, which satisfies

$$\lim_{t\to\infty} \bar{\tau}(t) = 0.$$

Definition 2.2: To each $\beta > 0$, we define the subset

$$D_\beta = \{(x,\tau) \in \mathbb{R}^n \times \mathbb{R} \mid \psi(x,\tau) \leq -\beta\}. \tag{4}$$

A solution $(\bar{x},\bar{\tau})$ of system 1 is said to be viable in the subset D_β on an interval $[0,\bar{t}[$, if

$$\forall t \in [0,\bar{t}[,\ (\bar{x}(t),\bar{\tau}(t)) \in D_\beta.$$

Proposition 2.3: Whenever a control $\bar{u}$ produces a viable solution $(\bar{x},\bar{\tau})$ on $[0,\infty[$ in the subset D_β, for an appropriate β, then it follows that $\bar{u}$ is a protocol in the sense of Definition 2.1.

Proof 2.4: For $t \geq 0$, (1b) is written

$$\dot{\bar{\tau}}(t) = \bar{\tau}(t)\psi(\bar{x}(t),\bar{\tau}(t)),$$

by viability of $(\bar{x},\bar{\tau})$ in D_β we get the differential inequality

$$\dot{\bar{\tau}}(t) \leq -\beta\bar{\tau}(t), \tag{5}$$

and by applying Gronwall's lemma we get the exponential estimate

$$0 < \bar{\tau}(t) \leq \tau_0 \mathrm{e}^{-\beta t},$$

then

$$\lim_{t\to\infty} \bar{\tau}(t) = 0.$$

Remark 2.5: Inequality (5) implies that $\dot{\bar{\tau}}(t) \leq 0$ for all t, i.e., tumor cell burden $\bar{\tau}$ will be on the decreasing, which is beneficial to the patient's quality of life during the treatment. The idea was first to give protocols to control tumor cell burden τ to 0, but it turns out that such protocols keep τ on decreasing sens overall $[0,\infty[$, with the exponential estimate $\tau_0\mathrm{e}^{-\beta t}$, where τ_0 is the initial tumor cell burden and $(-\beta)$ is the decay rate.

Remark 2.6: The viability of the solution $(\bar{x},\bar{\tau})$ in the subset D_β, requires the initial condition $\psi(x_0,\tau_0) < 0$, which is not always satisfied, but we will deal with this necessary condition in the next.

According to [1], such protocol may be characterized as a selection of the set-valued map $C_\beta(\cdot)$ defined on the subset D_β by

$$C_\beta(x,\tau) = \{u \in K_p \mid \langle \hbar(x,\tau), u\rangle_p \geq \ell(x,\tau)\}, \tag{6}$$

where functions $\hbar$ and ℓ are given by

$$\hbar(x,\tau) = -B'(x,\tau)\nabla_x\psi(x,\tau), \tag{7a}$$

$$\ell(x,\tau) = \langle \nabla_x\psi(x,\tau), f(x,\tau)\rangle_n + \tau\psi(x,\tau)\frac{\partial\psi}{\partial\tau}(x,\tau), \tag{7b}$$

with

- $\langle\cdot,\cdot\rangle_i$ is the euclidean inner product in $\mathbb{R}^i$, for $i = p, n$.
- $\nabla_x\psi(x,\tau)$ is the gradient vector at point x:

$$\nabla_x\psi(x,\tau) = \left(\frac{\partial\psi}{\partial x_1}(x,\tau),\cdots,\frac{\partial\psi}{\partial x_n}(x,\tau)\right)'.$$

To neatly give the set-valued characterization of protocol, we consider prior the following notation and hypothesis, and launch the ensuing theorem.

$$\Omega_- = \left\{(x,\tau) \in \mathbb{R}^n_+ \times \mathbb{R}_+ \mid \psi(x,\tau) < 0\right\}. \tag{8}$$

Hypothesis 2.7: The functions $\hbar$ and ℓ respectively given by formulas (7a) and (7b) satisfy the statement below

$$\forall (x,\tau) \in D_\beta, \exists u \in K_p\colon \langle \hbar(x,\tau), u\rangle_p > \ell(x,\tau).$$

Theorem 2.8: Let $(x_0,\tau_0) \in \Omega_-$ and $\beta_0 = -\psi(x_0,\tau_0) > 0$. Assume that there exists $\beta \in]0,\beta_0]$ for which Hypothesis 2.7 is checked. The selection s_β of the set-valued map $C_\beta(\cdot)$, having expression

$$s_\beta(x,\tau) = u \in C_\beta(x,\tau),\ \text{such as } \|u\| \longrightarrow \min, \tag{9}$$

provides a viable solution $(\bar{x},\bar{\tau})$ in the subset D_β on a maximal interval $[0,t^{\max}[$.

If in addition

Hypothesis 2.9: $\limsup_{\substack{t\to t^{\max}\\ t<t^{\max}}} \|(\bar{x}(t),\bar{\tau}(t))\| < \infty$,

then $t^{\max} = \infty$, and s_β becomes a protocol.

Remark 2.10: The Hypothesis 2.9 will prolongs the local viability of the solution $(\bar{x},\bar{\tau})$ over $t^{\max}$, to be global on the infinite interval $[0,\infty[$, while Hypothesis 2.7 is introduced so that the selection s_β will be well defined on D_β, but it turned out that Hypothesis 2.7 ensures in addition the continuity of s_β on D_β.

Now to do without the initial condition $\psi(x_0,\tau_0) < 0$ of Remark 2.6, we consider the following subset

$$\Omega_+ = \left\{(x,\tau) \in \mathbb{R}^n_+ \times \mathbb{R}_+ \mid \psi(x,\tau) \geq 0\right\}, \tag{10}$$

and we associate to each $\kappa > 0$ the set-valued map $C_\kappa(\cdot)$, having expression

$$C_\kappa(x,\tau) = \{u \in K_p \mid \langle \hbar(x,\tau), u\rangle_p \geq \ell(x,\tau) + \kappa\}, \tag{11}$$

where the functions $\hbar$ and ℓ still given by formula (7a) and (7b), then we set the ensuing theorem under the following hypothesis

Hypothesis 2.11:

$$\forall (x,\tau) \in \mathbb{R}^n_+ \times \mathbb{R}_+, \exists u \in K_p\colon \langle \hbar(x,\tau), u\rangle_p > \ell(x,\tau) + \kappa.$$

Theorem 2.12: Assume that $(x_0,\tau_0) \in \Omega_+$. The minimal selection s_κ of the set-valued map $C_\kappa(\cdot)$, having expression

$$s_\kappa(x,\tau) = u \in C_\kappa(x,\tau),\ \text{such as } \|u\| \longrightarrow \min, \tag{12}$$

steers the system 1 from Ω_+ to Ω_- on the interval $[0,t_1]$, provided that $\kappa t_1 > \psi(x_0,\tau_0)$. i.e., system 1 admits a solution $(\bar{x},\bar{\tau})$ on $[0,t_1]$, such that $(\bar{x}(t_1),\bar{\tau}(t_1)) \in \Omega_-$.

Proof 2.13: For $t \geq 0$,

$$\frac{\mathrm{d}}{\mathrm{d}t}\psi(\bar{x}(t),\bar{\tau}(t)) = \langle \nabla_x \psi(\bar{x}(t),\bar{\tau}(t)), \dot{\bar{x}}(t)\rangle_n + \dot{\bar{\tau}}(t)\frac{\partial \psi}{\partial \tau}(\bar{x}(t),\bar{\tau}(t)),$$

next, we use the dynamics (1a) and (1b), and the formulas (7a) and (7b) to write this differentiating in terms of the functions $\hbar$ and ℓ, and the selection s_κ:

$$\frac{\mathrm{d}}{\mathrm{d}t}\psi(\bar{x}(t),\bar{\tau}(t)) = \ell(\bar{x}(t),\bar{\tau}(t)) - \langle \hbar(\bar{x}(t),\bar{\tau}(t)), s_\kappa(\bar{x}(t),\bar{\tau}(t))\rangle_p,$$

or the selection s_κ is continuous (due to Hypothesis 2.11), then by integrating from $t = 0$ to t_1 we get

$$\psi(\bar{x}(t_1),\bar{\tau}(t_1)) = \psi(x_0,\tau_0) - \int_0^{t_1} [\langle \hbar(\bar{x}(t),\bar{\tau}(t)), s_\kappa(\bar{x}(t),\bar{\tau}(t))\rangle_p - \ell(\bar{x}(t),\bar{\tau}(t))]\,\mathrm{d}t,$$

since s_κ is a selection of the map $C_\kappa(\cdot)$ then we have

$$\psi(\bar{x}(t_1),\bar{\tau}(t_1)) \leq \psi(x_0,\tau_0) - \kappa t_1,$$

as

$$\kappa t_1 > \psi(x_0,\tau_0),$$

it follows that

$$\psi(\bar{x}(t_1),\bar{\tau}(t_1)) < 0.$$

Remark 2.14: The existence and the continuity of the selection s_κ require the Hypothesis 2.11.

Corollary 2.15: If the final state $(\bar{x}(t_1),\bar{\tau}(t_1))$ satisfies hypotheses of Theorem 2.8 as an initial one of system (1) at time $t_0 = t_1$, then

$$\left| \begin{array}{l} s_\kappa(\bar{x}(t),\bar{\tau}(t)) \text{ if } t \in [0,t_1], \\ s_\beta(\bar{x}(t),\bar{\tau}(t)) \text{ if } t \in [t_1,\infty[, \end{array} \right. \tag{13}$$

is a protocol.

Remark 2.16: The protocol (13) will increase the corresponding tumor cell burden $\bar{\tau}$ on an neighborhood of $t_0 = 0$, relatively to the interval $[0,t_1]$, which may be disadvantageous to the patient's quality of life in the beginning of treatment.
The condition $\kappa t_1 > \psi(x_0,\tau_0)$ can be regarded as a conflict between the continuity of protocol (13) at $(\bar{x}(t_1),\bar{\tau}(t_1))$, and the minimality of period t_1.
The protocol's existence is independent of any condition on initial state (x_0,τ_0), contrary to the decreasing of the corresponding tumor cell burden $\bar{\tau}$ which closely depends on sign of $\psi(x_0,\tau_0)$. This leads to stage the cancer as follow:

- $\psi(x_0,\tau_0) < 0$: As pointed out in Remark 2.5, the protocol (9) will decrease the cancer cells $\bar{\tau}$ on $[0,\infty[$. We then say that the cancer is non-advanced, or it is in stage I.
- $\psi(x_0,\tau_0) \geq 0$: As Remark 2.16, the protocol (13) cannot ensures the decreasing of cancer cells $\bar{\tau}$ on $[0,t_1]$. In this case we say that the cancer is advanced, or it is in stage II.

The initial state (x_0,τ_0) is deterministic to distinguish treatment strategies I and II. To cope with the advanced stage: $\psi(x_0,\tau_0) \geq 0$, we must transfer (x_0,τ_0) to a state (x_1,τ_1) of better stage, ie. $\psi(x_1,\tau_1) < 0$ at a time $t_1 > 0$, before joining the interval of decreasing $[t_1,\infty[$. Ultimately, the study reveals the key role that play the function ψ in cancer staging and designing protocols (9) and (13).

III. Application to the Model

A. The model

The model [3] under study is a normalized system of five ODEs, modeling continual interplay between normal cells: NCs, endothelial cells: ECs, chemotherapy agent: CA, anti-angiogenic agent: AA of concentration x_i, respectively, for $i = 1,2,3,4$ (here the integer $n = 4$); and cancer cells: CCs of concentration τ.

$$\dot{x}_1 = f_1(x,\tau), \tag{14a}$$
$$\dot{x}_2 = f_2(x,\tau), \tag{14b}$$
$$\dot{x}_3 = \delta + f_3(x,\tau), \tag{14c}$$
$$\dot{x}_4 = \phi + f_4(x,\tau), \tag{14d}$$
$$\dot{\tau} = \tau\psi(x,\tau), \tag{14e}$$

initial conditions are given by

$$x_i(t=0) = x_0^i \geq 0, \text{ for } i = 1,2,3,4; \tag{14f}$$
$$\tau(t=0) = \tau_0 \geq 0. \tag{14g}$$

Parameters $\delta = \delta(t)$ and $\phi = \phi(t)$ are the respective instantaneous infusion doses of CA and AA. (δ,ϕ) is almost continuous within numerical values of the reduced constraint subset (here the number of therapeutic agents $p = 2$)

$$K_2 = \left[0,\delta^{\max}\right] \times \left[0,\phi^{\max}\right], \tag{15}$$

defined by

$$\{(\delta,\phi) \in \mathbb{R}^2 \mid 0 \leq \delta \leq \delta^{\max} \text{ and } 0 \leq \phi \leq \phi^{\max}\}.$$

Infusion doses CA(δ) and AA(ϕ) are based on cyclophosphamide and TNP-470 respectively.

The model variables x_i and τ are assumed to interact as follows:

1) Normal cells.
   a) NCs grow logistically in the absence of CCs and AA,
   b) CCs destroy NCs at a rate proportional to the product of CCs,
   c) CA kills NCs at a rate governed by Michaelis-Menton kinetics with parameters p_3 and a_3.
2) Endothelial cells.
   a) ECs increase logistically to normalized carrying capacity,
   b) CCs stimulate ECs to grow at a rate directly proportional to CCs,
   c) AA kills ECs at a rate again governed by Michaelis-Menton term.

3) Chemotherapy agent.
   a) CA decays at a constant rate,
   b) CA decreases due to its action on NCs and CCs at a rates modeled by Michaelis-Menton dynamics.
4) Anti-angiogenic agent.
   a) AA degrades at a constant rate,
   b) AA decreases because of its action on ECs at a rate again modeled by Michaelis-Menton factor.
5) Cancer cells.
   a) CCs exhibit logistic proliferation rate with varying carrying capacity depending on ECs: The ECs increase the carrying capacity of the CCs,
   b) NCs eliminate CCs at a rate proportional to the product of NCs,
   c) CA kills CCs at a rate in Michaelis-Menton form.

Under the above considerations, functions f_i and ψ are expressed as

$$f_1(x,\tau) = \alpha_1 x_1(1-x_1) - q_1 x_1 \tau - p_1(x_2,x_4)\frac{x_1 x_3}{a_1+x_1}, \tag{16a}$$

$$f_2(x,\tau) = \beta\tau + \alpha_3 x_2(1-x_2) - \frac{p_3 x_2 x_4}{a_3+x_2}, \tag{16b}$$

$$f_3(x,\tau) = -\left[\xi + d_1\frac{x_1}{a_1+x_1} + d_2\frac{\tau}{a_2+\tau}\right] x_3, \tag{16c}$$

$$f_4(x,\tau) = -\left[\eta + d_3\frac{x_2}{a_3+x_2}\right] x_4, \tag{16d}$$

$$\psi(x,\tau) = \alpha_2\left[1-\frac{\tau}{1+\gamma x_2}\right] - q_2 x_1 - p_2(x_2,x_4)\frac{x_3}{a_2+\tau}, \tag{16e}$$

with

$$p_i(x_2,x_4) = p_{i0} + p_{i1}x_2 + p_{i2}x_4, \text{ for } i=1,2. \tag{16f}$$

Table I describes the model terms, and Table II lists the values of the positive constants parameters $\alpha_i, q_i, a_i, d_i, p_{ij}, \gamma, \beta, p_3, \xi$, and η (units are in days^{-1} except for the a_i's, and γ whose units are volume).

1 shows model interactions between the different interveners x_1, x_2, τ, and with therapy agents x_3, x_4.

B. Explicit expressions of protocols

According to [3], for all non-negative initial state (14f) and (14g), all solutions $(\bar{x},\bar{\tau})$ of model (14) remain in the non-negative orthant

$$\mathbb{R}^5_+ = \{(x_1,x_2,x_3,x_4,\tau)' \in \mathbb{R}^5 \mid x_i \geq 0, \text{ for } i=1,2,3,4; \text{ and } \tau \geq 0\},$$

so the model fulfills the positivity condition (2).

To give useful expressions of protocols (9) and (13) for the model (14), we have to explicit functions $\hbar$ and ℓ, given by (7a) and (7b), so we have to calculate partial derivatives of function ψ as expressed by (16e).

Entity	Term	Description
x_1	$\alpha_1 x_1(1-x_1)$	Logistic NCs growth
	$-q_1 x_1 \tau$	NCs death by CCs
	$-p_1(x_2,x_4)\frac{x_1 x_3}{a_1+x_1}$	NCs death by CA per concentration of ECs and AA
x_2	$\beta\tau$	CCs creation due to ECs
	$\alpha_3 x_2(1-x_2)$	Logistic ECs growth
	$-\frac{p_3 x_2 x_4}{a_3+x_2}$	ECs death by AA
x_3	$-\xi x_3$	CA washout
	$-d_1\frac{x_1 x_3}{a_1+x_1}$	CA combination with NCs
	$-d_2\frac{\tau x_3}{a_2+\tau}$	CA combination with CCs
x_4	$-\eta x_4$	AA washout
	$-d_3\frac{x_2 x_4}{a_3+x_2}$	AA combination with ECs
τ	$\alpha_2\tau\left[1-\frac{\tau}{1+\gamma x_2}\right]$	Logistic CCs growth
	$-q_2 x_1 \tau$	CCs death by NCs
	$-p_2(x_2,x_4)\frac{\tau x_3}{a_2+\tau}$	CCs death by CA per concentration of ECs and AA

TABLE I
TERMS OF MODEL (14).

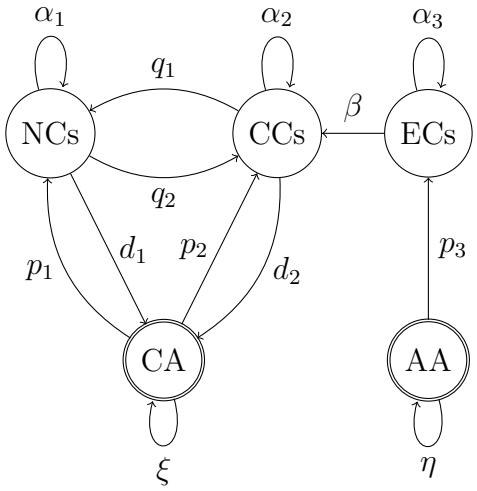


Fig. 1. Interactions in model (14) between compartments of cells, NCs, CCs and ECs, and compartments of agents, CA and AA.

$$\left|\begin{aligned}
\frac{\partial\psi}{\partial x_1}(x,\tau) &= -q_2,\\
\frac{\partial\psi}{\partial x_2}(x,\tau) &= \frac{\alpha_2\gamma\tau}{(1+\gamma x_2)^2} - \frac{p_{21}x_3}{a_2+\tau},\\
\frac{\partial\psi}{\partial x_3}(x,\tau) &= -\frac{p_2(x_2,x_4)}{a_2+\tau},\\
\frac{\partial\psi}{\partial x_4}(x,\tau) &= -\frac{p_{22}}{a_2+\tau},\\
\frac{\partial\psi}{\partial\tau}(x,\tau) &= \frac{x_3 p_2(x_2,x_4)}{(a_2+\tau)^2} - \frac{\alpha_2}{1+\gamma x_2},
\end{aligned}\right.$$

whence

$$\begin{aligned}\ell(x,\tau) &= -q_2 f_1(x,\tau)\\ &+f_2(x,\tau)\left[\frac{\alpha_2\gamma\tau}{(1+\gamma x_2)^2}-\frac{p_{21}x_3}{a_2+\tau}\right]\\ &-\frac{p_2(x_2,x_4)}{a_2+\tau}f_3(x,\tau)-\frac{p_{22}}{a_2+\tau}f_4(x,\tau)\\ &+\tau\psi(\tau,x)\left[\frac{x_3p_2(x_2,x_4)}{(a_2+\tau)^2}-\frac{\alpha_2}{1+\gamma x_2}\right],\end{aligned}$$

the operator $B(x,\tau)$ is represented by the matrix

$$B=\begin{pmatrix}0&0&1&0\\0&0&0&1\end{pmatrix}',$$

then

$$\hbar(x,\tau)=\left(\frac{p_2(x_2,x_4)}{a_2+\tau},\frac{p_{22}}{a_2+\tau}\right)',$$

and so, set-valued maps $C_\beta(\cdot)$ and $C_\kappa(\cdot)$ given by (6) and (11) are expressed as follows

$$C_\beta(x,\tau)=\left\{u\in K_2 \mid \frac{p_2(x_2,x_4)}{a_2+\tau}\delta+\frac{p_{22}}{a_2+\tau}\phi\geq\ell(x,\tau)\right\},$$

and

$$C_\kappa(x,\tau)=\left\{u\in K_2 \mid \frac{p_2(x_2,x_4)}{a_2+\tau}\delta+\frac{p_{22}}{a_2+\tau}\phi\geq\ell(x,\tau)+\kappa\right\},$$

then, selections s_β (9) and s_κ (12) are defined on $\mathbb{R}^4_+\times\mathbb{R}_+$ by

$$s_\beta(x,\tau)=(\delta,\phi)\in C_\beta(x,\tau),\ \text{such as}\ \delta^2+\phi^2\longrightarrow\min, \tag{17}$$

and

$$s_\kappa(x,\tau)=(\delta,\phi)\in C_\kappa(x,\tau),\ \text{such as}\ \delta^2+\phi^2\longrightarrow\min. \tag{18}$$

To complete our analysis we have to check the Hypothesis 2.9 of Theorem 2.8, for that we will use the following estimations of $\bar{x}_i$ and $\bar{\tau}$, taken from the proof of [3]

$$\begin{cases}\bar{x}_1(t)\leq 1,\\ \bar{x}_3(t)\leq\dfrac{\delta^{\max}}{\xi},\\ \bar{x}_4(t)\leq\dfrac{\phi^{\max}}{\eta},\\ \bar{x}_2(t)\leq M,\\ \bar{\tau}(t)\leq 1+\gamma M,\end{cases}$$

where the constant M is given by

$$M=\frac{1}{2}\left(1+\frac{\gamma\beta}{\alpha_3}\right)+\frac{1}{2}\sqrt{\left(1+\frac{\gamma\beta}{\alpha_3}\right)^2+\frac{4\beta}{\alpha_3}},$$

then

$$\|(\bar{x}(t),\bar{\tau}(t))\|\leq\max\left(\frac{\delta^{\max}}{\xi},\frac{\phi^{\max}}{\eta},1+\gamma M\right),$$

it follows that Hypothesis 2.9 is well checked.

Remark 3.1: For $\bar{\tau}$ we can also use the estimation

$$\forall t\in[0,t^{\max}[,\ \bar{\tau}(t)\leq\tau_0,$$

then Hypothesis 2.9 can be reduced to

$$\limsup_{\substack{t\to t^{\max}\\ t<t^{\max}}}\|\bar{x}(t)\|<\infty.$$

Hypotheses 2.7 and 2.11 are kept to the next section to be checked numerically.

C. Numerical simulations

In order to illustrate analytical results of previous section, we propose to state model (14) at stages I and II. We use the ordinary differential equations solver *ode45* of *matlab* to integrate model (14), concurrently with the quadratic programming solver *Quadprog* to return numerical approximations of selections (17) and (18).

The parameter values in Table II are taken into account of the simulations.

Parameter	Value
α_1	$0.0068\ \text{day}^{-1}$
α_2	$0.01\ \text{day}^{-1}$
α_3	$0.002\ \text{day}^{-1}$
q_1	$0.00702\ \text{day}^{-1}$
q_2	$0.00072\ \text{day}^{-1}$
γ	0.1615
β	$0.00371\ \text{day}^{-1}$
a_1	1.1
a_2	4.6205
a_3	4.6666
d_1	$0.0002\ \text{day}^{-1}$
d_2	$0.032\ \text{day}^{-1}$
d_3	$0.032\ \text{day}^{-1}$
p_{10}	$1.2\times10^{-7}\ \text{day}^{-1}$
p_{20}	$0.20581\ \text{day}^{-1}$
p_3	$1.7143\ \text{day}^{-1}$
p_{11}	$4.2\times10^{-8}\ \text{day}^{-1}$
p_{12}	$1.0\times10^{-7}\ \text{day}^{-1}$
p_{21}	$0.00431\ \text{day}^{-1}$
p_{22}	$19.4872\ \text{day}^{-1}$
ξ	$0.01813\ \text{day}^{-1}$
η	$0.136\ \text{day}^{-1}$

TABLE II
PARAMETERS OF MODEL (14).

The used numerical values of infusion rates of parameters δ and ϕ are based on their actual doses per tumor mass: $mg/(kg)(day)$.

Respectively, the numerical results seen in Figures 2 and 3 correspond to the following particular initial stages I and II, which belong subsets Ω_- and Ω_+, defined by (8) and (10).

$$\text{I:}\begin{cases}x_0^1=0.999998, & x_0^2=0.728255,\\ x_0^3=0.181034, & x_0^4=0.00171038,\\ \tau_0=0.6,\end{cases} \tag{19a}$$

and

$$\text{II:}\begin{cases}x_0^1=1, & x_0^2=1,\\ x_0^3=0.001, & x_0^4=0.1,\\ \tau_0=1.\end{cases} \tag{19b}$$

1) The numerical approach in Figure 2 illustrates the effects of the protocol on CCs, NCs, and ECs, which:
   a) Reverses the growth of CCs and make it on strict decreasing.
   b) Keeps NCs which is viewed as indicator of the patient's health on acceptable levels.
   c) Deprives CCs of useful ECs densities.
2) As Figure 3, it confirm Remark 2.16, CCs have a slight growth at the start of therapy, precisely on the time interval $[0, 10]$.

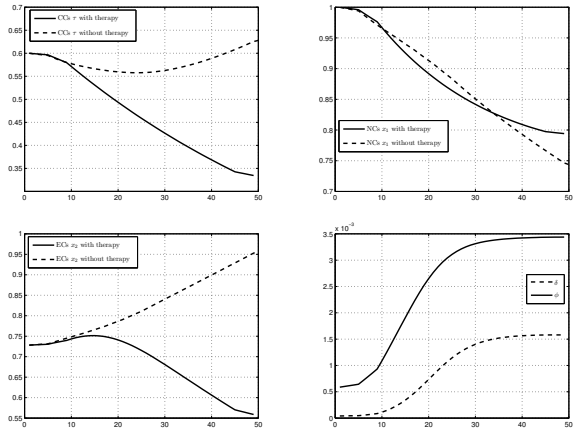


Fig. 2. Time evolution of CCs, NCs, and ECs, in absence of therapy: $(\delta, \phi) \equiv 0$ and presence of therapy: $(\delta, \phi) \equiv s_\beta(\bar{x}(t), \bar{\tau}(t))$, with $\beta = 0.00584$. (β is the parameter introduced in Definition 2.2).

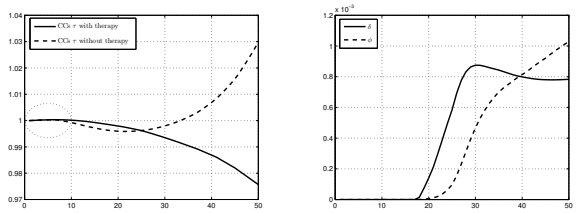


Fig. 3. Behavior of cancer cells τ, in absence of therapy: $(\delta, \phi) \equiv 0$ and presence of therapy: $(\delta, \phi) \equiv (13)$, with $t_1 = 30$ and $\kappa = 10^{-5} \approx 0$. ($\kappa$ is very small that (13) be continuous, as in Remark 2.16).

IV. Conclusion

The set-valued method [1], [6] is adapted to control Pinho's model (14). Numerical simulations in Figures 2 and 3 agreed with the theoretical characterization of the protocols (9) and (13). The main advantages of this method over others is that it:

1) Does not request complicated conditions to models subject of study, just
   a) The smoothness propriety of the function ψ, that it be of class $\mathcal{C}^1$ on $\mathbb{R}^n_+ \times \mathbb{R}_+$.
   b) The upper limit Hypothesis 2.9, which can be given by using comparison method, like the standard Kamke comparison theory used in the proof of [3].
2) Solves the setting control problem in Definition 2.1, for any value taken by the initial state (x_0, τ_0) in $\mathbb{R}^n_+ \times \mathbb{R}_+$, even the system 1 is in advanced stage I: $\psi(x_0, \tau_0) \geq 0$, we manage to build a protocol.
3) Gives continuous protocols on state feedback, with minimal norms (in the frame-work of set-valued analysis), which constitute lively advantages to the patient over the therapy session.
4) Provides protocols by doing simple selections, contrary to those given by the optimal control theory, which uses the result developed by Fleming and Rishel to establish the existence of an optimal control, and the classical Pontryagin's Maximum Principle to characterize optimal control, and some propositions to prove the uniqueness of the optimality system, in which state system is coupled with co-state system [9], [20], [21].
5) Does not require an analysis that puts into account the parameter values, unlike the case if we seek to stabilize the model (14) around its equilibrium points.
   In [3] for example, the equilibrium points of sub-models (in the absence of CA or AA) and the full-model (in the presence of both CA and AA), depend on constant parameters of Table II, and on variable parameters δ and ϕ, and these parameters also determine the nature of the equilibria.
6) Is applicable not perforce around the equilibrium points, but to the entire dynamics of the system (free state x).
7) Is a qualitative analysis with non-specific system parameters, that can vary greatly from patient to another [11], or can depend on time during the therapy for a given patient [23].
8) Is useful to approach another problem situations, apart from controlling cancer, inspired by the work carried out in [1], and [6], [25] adapts a unified set-valued method, to approach the asymptotic null-controllability with mixed constraints on state and control.

References

[1] K. Kassara, *A unified set-valued approach to control immunotherapy*, SIAM J. Contr. Optim., 48(2009), 909-924.
[2] K. Kassara, *A set-valued approach to control immunotherapy*, Mathematical and Computer Modelling, 44(2006), 1114-1125.
[3] S.T.R. Pinho and F.S. Bacelar and R.F.S. Andrade and H.I. Freedman, *A mathematical model for the effect of anti-angiogenic therapy in the treatment of cancer tumours by chemotherapy*, Nonlinear Analysis: Real World Applications, 14(2013) 815-828.
[4] Khalili, Pariya and Zolatash, Sareh and Vatankhah, Ramin and Taghvaei, Sajjad, *Optimal control methods for drug delivery in cancerous tumour by anti-angiogenic therapy and chemotherapy*, IET Systems Biology, 5(2021) 14-25.
[5] Kirchner, D. and Panetta, J.C, *Modeling immunotherapy of the tumor-immune interaction*, J. Math. Biol, 37(1998), 235-252.
[6] K. Kassara and A. Moustafid, *Angiogenesis inhibition and tumor-immune interactions with chemotherapy by a control set-valued method*, Mathematical Biosciences, 231(2011), 135-143.
[7] Alberto d'Onofrio, *Metamodeling tumor-immune system interaction, tumor evasion and immunotherapy*, Mathematical and Computer Modelling, 47(2008), 614-637.
[8] Hahnfeldt, P. and Panigrahy, D and Folkman, J. and Hlatky, L., *Tumor development under angiogenic signaling: A dynamical theory of tumor growth, treatment response, and postvascular dormancy*, Cancer Res, 59(1999), 4770-4775.

[9] L.G. de Pillis and W. Gu and K.R. Fister and T. Head and K. Maples and A. Murugan and T. Neal and K. Yoshida, *Chemotherapy for tumors: An analysis of the dynamics and a study of quadratic and linear optimal controls*, Mathematical Biosciences, 209(2007), 292-315.

[10] L.G De Pillis and A Radunskaya, *The dynamics of an optimally controlled tumor model: A case study* , Mathematical and Computer Modelling, 37(2003), 1221-1244.

[11] L.G. de Pillis and W. Gu and A.E. Radunskaya, *Mixed immunotherapy and chemotherapy of tumors: modeling, applications and biological interpretations*, Journal of Theoretical Biology, 238(2006), 841-862.

[12] de Pillis, L.G. and Radunskaya, A.E. and Wiseman, C.L., *A validated mathematical model of cell-mediated immune response to tumor growth*, Cancer Res, 61(2005), 7950-7958.

[13] Sotolongo-Costa, O. and Morales-Molina, L. and Rodriguez-Perez, D. and Antonranz, J.C. and Chacon-Reyes, M., *Behavior of tumors under non stationary therapy*, Cancer Res, 61(2003), 242-253.

[14] Swarnali Sharma and Samanta, G. P., *Dynamical Behavior of a Tumor-Immune System with Chemotherapy and Optimal Control*, Journal of Nonlinear Dynamics, (2013).

[15] *Mathematical modelling of the spatio-temporal response of cytotoxic T-lymphocytes to a solid tumor*, Matzavinos, A. and Chaplain, M.A. and Kuznetsov, J.V.A., Math. Med. Biol., 21(2004), 1-34.

[16] F. Castiglione and B. Piccoli, *Cancer immunotherapy, mathematical modeling and optimal control* , Journal of Theoretical Biology, 247(2007), 723-732.

[17] Werner Krabs and Stefan Pickl, *An optimal control problem in cancer chemotherapy*, Applied Mathematics and Computation, 217(2010), 1117-1124.

[18] Miljana Jovanović and Marija Krstić, *Analysis of non-autonomous stochastic Gompertz model with delay*, Applied Mathematics and Computation, 242(2014), 101-108.

[19] El Kettani Moummou and R. Gutiérrez-Sanchez and M.C. Melchor and E. Ramos-Ábalos, *A stochastic Gompertz model highlighting internal and external therapy function for tumour growth*, Applied Mathematics and Computation, 246(2014), 1-11.

[20] Subhas Khajanchi and Dibakar Ghosh, *The combined effects of optimal control in cancer remission*, Applied Mathematics and Computation, 271(2015), 375-388.

[21] Burden, T. and Ernstberger, J. and Fister, K. R., *Optimal control applied to immunotherapy*, Dis. Cont. Dyn. B, 4(2004), 135-146.

[22] Kanchi Lakshmi Kiran and S. Lakshminarayanan, *Optimization of chemotherapy and immunotherapy: In silico analysis using pharmacokinetic-pharmacodynamic and tumor growth models*, Journal of Process Control, 23(2013), 396-403.

[23] Mazen Alamir, *On probabilistic certification of combined cancer therapies using strongly uncertain models*, Journal of Theoretical Biology, 384(2015), 59-69.

[24] Marek Bodnar and Monika Joanna Piotrowska, *Stability analysis of the family of tumour angiogenesis models with distributed time delays*, Communications in Nonlinear Science and Numerical Simulation, 31(2016), 124-142.

[25] Boujallal, L. and Kassara, K., *State-input constrained asymptotic null-controllability by a set-valued approach*, Control Theory Applications, IET, 9(2015), 2211-2221.

[26] Monika Joanna Piotrowska, *An immune system-tumour interactions model with discrete time delay: Model analysis and validation*, Communications in Nonlinear Science and Numerical Simulation, 34(2016), 185-198.

[27] Kanchi Lakshmi Kiran and S. Lakshminarayanan, *Global sensitivity analysis and model-based reactive scheduling of targeted cancer immunotherapy* , Biosystems, 101(2010), 117-126.

[28] Louise Viger and Fabrice Denis and Martin Rosalie and Christophe Letellier, *A cancer model for the angiogenic switch* , Journal of Theoretical Biology, 360(2014), 21-33.

[29] Antonio Cappuccio and Filippo Castiglione and Benedetto Piccoli, *Determination of the optimal therapeutic protocols in cancer immunotherapy* , Mathematical Biosciences, 209(2007), 1-13.

[30] Hsiu-Chuan Wei, *A modified numerical method for bifurcations of fixed points of ODE systems with periodically pulsed inputs*, Applied Mathematics and Computation, 236(2014), 373-383.

[31] Sandip Banerjee and Ram Rup Sarkar, *Delay-induced model for tumor-immune interaction and control of malignant tumor growth*, Biosystems, 91(2008), 268-288.

[32] Sébastien Benzekry and Guillemette Chapuisat and Joseph Ciccolini and Alice Erlinger and Florence Hubert, *A new mathematical model for optimizing the combination between antiangiogenic and cytotoxic drugs in oncology*, Comptes Rendus Mathematique, 350(2012), 23-28.

[33] Harold P. de Vladar and Jorge A. González, *Dynamic response of cancer under the influence of immunological activity and therapy*, Journal of Theoretical Biology, 227(2004), 335-348.

[34] C. Cattani and A. Ciancio, *Qualitative analysis of second-order models of tumor-immune system competition*, Mathematical and Computer Modelling, 47(2008), 1339-1355.

[35] Sandip Banerjee and Alexei Tsygvintsev, *Stability and bifurcations of equilibria in a delayed Kirschner-Panetta model*, Applied Mathematics Letters, 40(2015), 65-71.

[36] Rachel Walker and Heiko Enderling *From concept to clinic: Mathematically informed immunotherapy*, Current Problems in Cancer, (2015).

[37] Urszula Ledzewicz and Mozhdeh Sadat Faraji Mosalman and Heinz Schättler, *Optimal controls for a mathematical model of tumor-immune interactions under targeted chemotherapy with immune boost*, Discrete and Continuous Dynamical Systems - Series B, 18(2013), 1031-1051.

[38] Konstantin E. Starkov and Alexander P. Krishchenko, *On the global dynamics of one cancer tumour growth model*, Communications in Nonlinear Science and Numerical Simulation, 19(2014), 1486-1495.

[39] Konstantin E. Starkov and Luis N. Coria, *Global dynamics of the Kirschner-Panetta model for the tumor immunotherapy*, Nonlinear Analysis: Real World Applications, 14(2013), 1425-1433.

[40] Michael Engelhart and Dirk Lebiedz and Sebastian Sager, *Optimal control for selected cancer chemotherapy ODE models: A view on the potential of optimal schedules and choice of objective function*, Mathematical Biosciences , 229(2011), 123-134.

[41] Diego S. Rodrigues and Paulo F.A. Mancera and Suani T.R. Pinho, *Understanding the antiangiogenic effect of metronomic chemotherapy through a simple mathematical model*, Physica A: Statistical Mechanics and its Applications, 464(2016), 251-266.

[42] Alamir, Mazen, *Robust feedback design for combined therapy of cancer*, Optimal Control Applications and Methods, 35(2014), 77-88.

[43] Joseph Malinzi and Rachid Ouifki and Amina Eladdadi and Delfim F. M. Torres and K. A. Jane White, *Enhancement of chemotherapy using oncolytic virotherapy: Mathematical and optimal control analysis*, Mathematical Biosciences & Engineering, 15(2018), 1435-1436.

[44] Joseph Malinzi and Kevin Bosire Basita and Sara Padidar and Henry Ademola Adeola, *Prospect for application of mathematical models in combination cancer treatments*, Informatics in Medicine Unlocked, 23(2021), 100534.

[45] Spolaor, Simone and Scheve, Martijn and Firat, Murat and Cazzaniga, Paolo and Besozzi, Daniela and Nobile, Marco S., *Screening for Combination Cancer Therapies With Dynamic Fuzzy Modeling and Multi-Objective Optimization*, Frontiers in Genetics, 12(2021), 449.

Principal Component Analysis applied to survey data: Methodological aspects and application

Bourass Hind

Université Mohammed V Rabat, FSJES Agdal

bourasshind1@gmail.com

Soussi Noufail Outmane

Université Mohammed V Rabat, FSJES Salé

soussioutmane@gmail.com

ORCID : 0000-0002-0269-7935

***Abstract*—**

The objective of this paper is to highlight the methodology of principal component analysis. It is a branch of multivariate descriptive statistics that allows for the simultaneous processing of any number of quantitative data. Firstly, it is interesting to review the different methods used for the construction of the questionnaire and the necessary steps for its elaboration. Secondly, the techniques used can be generalised to two complementary analyses: reliability analysis and dimensionality analysis. Secondly, the focus will be on assessing the internal consistency of the scales. It is therefore appropriate to consider principal component analysis as an extremely powerful tool for synthesising the information contained in the various data to have a representation that allows easier interpretation.

***Keywords*—**

Factor analysis, Survey, Principal component analysis, Multivariate statistics

I. INTRODUCTION

In management science, there are several methodological choices in applied research. Surveys are one of the most widely used research methods in this discipline, allowing the analysis of different types of variables, both qualitative and quantitative.

The methodological choice based on quantitative studies makes it possible to obtain numerical

relational information that serves to better understand the chosen theme or subject. Moreover, recent developments in quantitative tools make it possible to quantify qualitative variables, and one can, conversely, make interpretations and draw qualitative conclusions from quantitative data (Barbet, 1988).

The aim of this paper is to highlight a methodology often used by management science researchers whose main research instrument is data drawn from questionnaires. This is the methodology of exploratory analysis of measurement instruments. Exploratory factor analysis aims to explore a number of items in a given population. The techniques used can be generalised to two complementary analyses: reliability analysis and dimensionality analysis. For this purpose, a statistical analysis is used: factor analysis.

This paper is divided into four points. The first point describes the steps involved in designing the questionnaire. The second point describes the choice of a scale to measure a latent variable from a set of items. The methodology for the evaluation of these measurement instruments is discussed with a numerical application. The last point is reserved for the methodology of the application of the principal component analysis method and its statistical evaluation

II. METHODOLOGY FOR THE CONSTRUCTION OF THE QUESTIONNAIRE

A. Definition and construction of the research instrument

The questionnaire is a document on which the answers or reactions of one or more individuals are noted. Indeed, it is applied to a set (sample) that must allow statistical inferences to be made, in order to measure and evaluate behavioural phenomena in a population. Therefore, the systematic implementation of pre-test procedures and scientific design of questionnaires is essential for the quality of the data collected, in particular to minimise so-called measurement errors.

The questionnaire survey is considered as a data collection tool to understand and explain phenomena on the one hand, and to study the characteristics of a population on the other. In addition, the questionnaire may allow us to verify, confirm or refute hypotheses related to a research.

Thus, an essential step in market research. Firstly, it provides the data needed to create scales for the brief as its structured form makes it easier to analyse the hypotheses. (Malhotra, 2008).

Secondly, it is a measurement tool. Its main purpose is to operationalise the user's request for information in a format that allows statistical measurement. The concepts of 'reality' must be operationalised in such a way that specialists and users can perform the necessary analyses, which the questionnaire designer can implement in the questionnaire, and which the respondents can understand and answer correctly.

However, in order for the questionnaire to fulfil its functions, it was necessary to have a clear and precise definition of the questionnaire. The first phase of the development of the questionnaire was a very crucial phase for us in order to be in line with the scientific approach. For this purpose, we were largely inspired by the steps proposed by (Churchill, 1998):

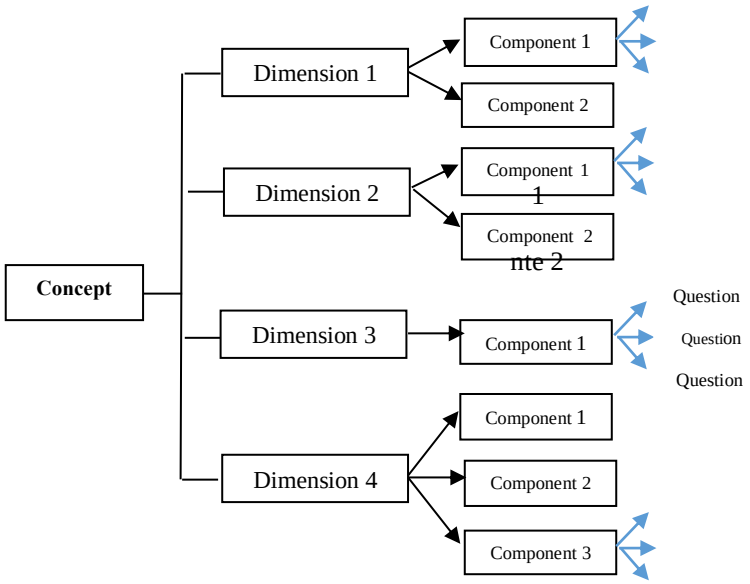
- Defining the purpose of the research ;
- The specification of the information sought, i.e. the objectives and hypotheses to be tested;
- The type of questionnaire and its mode of administration ;
- The content of individual questions ;
- The form of response to each question ;
- The wording of each question ;
- The sequence of questions ;
- The physical characteristics of the questionnaire;
- Pre-testing of the questionnaire, and drafting of the final version.

The latter plays a key role in the data collection process, as well as influencing the image of the statistical agency that uses it. It also. It has a significant influence on respondent behaviour, interviewer performance, collection costs and respondent relations, and therefore has a considerable impact on data quality.

The process of designing the questionnaire consists of several successive steps:

- The development of a conceptual framework,
- The writing and sequencing of questions,
- The correct use of visual design elements and the technical implementation of electronic questionnaires.

Figure 1: Construction of a questionnaire



Therefore, we will now deal with the essential steps related to the construction and implementation of the questionnaire.

B. Forms and phases for the construction of the questionnaire

The questionnaire survey is a reliable measurement instrument that is easy to[1] administer and analyse (Malhotra, 2008).
Questionnaires can be administered by interviewers in person or by telephone, or they can be self-administered on paper or in another medium, such as an audio cassette or the internet. Respondents may be asked to report on themselves, others in their household or other entities, such as businesses.

The design of the questionnaire, in accordance with the Code of Practice [2] . Therefore, the wording, structure and presentation of all questionnaires must lead to valid and reliable results. There are two types of questionnaires:

- Direct-administration questionnaires: in this case, the respondent writes his/her answers on the questionnaire. The person conducting the survey (the interviewer) may or may not be present; if he or she is present, he or she may clarify the content of an answer if the respondent so requests.
- Indirect administration questionnaires: in this case,the interviewer records the answers provided by the subject. The interviewer is therefore necessarily present.

Indirect administration gives the best results, but requires greater resources; when the number of people to be interviewed is very large or the wording extremely precise, direct administration should be used.

[1] Surveys are one of the most affordable ways to collect quantitative data.

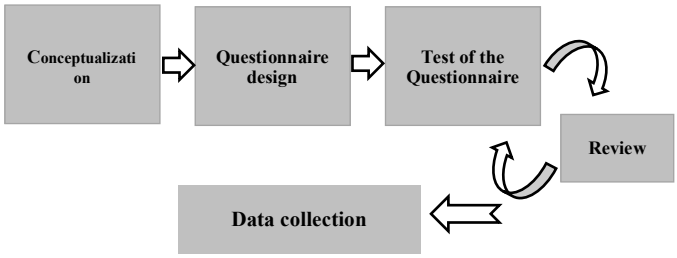
[2] This is principle number 12 of the European Statistics Code of Practice, which is intended to ensure that statistics accurately and reliably describe reality.

The questionnaires can be administered in three ways:

- *By mail*: questionnaires are mailed to individuals in the selected sample.
- *By interview*: the questionnaire is completed during or after a face-to-face meeting between an interviewer and the respondent.
- *By telephone*: only a very short questionnaire can be submitted.

The risk in the first method is that there may be a greater or lesser percentage of non-responses, which distorts the representativeness of the sample. In contrast to interviews, the composition of the sample is easier to respect in this case, but the possibility of refusal or absence of respondents should not be forgotten.
When developing a questionnaire, the whole cycle of questionnaire design and testing should be covered. Five main steps should be distinguished (see Figure 2).

Figure 2: The five stages to be covered by a questionnaire



a. Conceptualisation phase:

This is the most important phase that comes before we even start thinking about the formulation of the questions, the conceptual basis of the questionnaire must be operationalised from the literature review.

The complexity of theoretical concepts entirely requires a strict selection of empirical features (indicators) that can be observed in a survey. These indicators are considered an appropriate representation of the concept. On the other hand, the development of the conceptual framework is obviously more important for completely new surveys, whereas in existing surveys the concepts are well established. In this case, less attention is required at this level.

The main output of this step is the list of target variables (in addition to basic decisions on the target population, data collection mode, etc.).

b. Questionnaire design phase

The questionnaire design phase starts after the specification of the conceptual basis and the determination of the data collection method. Based on the content of the questionnaire and the requirements of the data collection method, the sequence of the thematic sections of the questionnaire is decided.
The form and choice of answers proposed will be specified according to three criteria:

- Clarity[3] (understanding);
- Neutrality[4] (authenticity of answers)
- Adequac [5] which refers to the ability of the interviewees to answer the questionnaire. Three elements are important at this level: the length of the questionnaire, the order and the orientation of the questions.

Furthermore, the strategy for designing a questionnaire is relatively simple. It involves ensuring four important things:

- That respondents understand the questions.
- That they are able to respond.
- Let them agree to respond.
- That the answer is formulated authentically and not influenced

c.Questionnaire testing phase

The third phase consists of testing the questionnaire on a representative sample of the panel in order to check the order of the questions and their comprehension, and then to correct the questionnaire if necessary in the light of any problems encountered. The objective of this test is to evaluate the ease of understanding, the degree of acceptance and the ease of interpretation. The test is therefore an absolutely necessary step that must be carried out with rigour.
In this phase two things need to be checked:

- That the answers to each question are consistent with the expected outcome;
- This can be done by analysing the testers' answers, or by discussing them better once they have answered.

As making a good questionnaire is much more difficult than you might think, this step is essential, even for the most experienced, to ensure that there are no ambiguities that you might have missed.

d. Review phase

Generally, two or more rounds of questionnaire testing are recommended. If a questionnaire has been modified as a result of the test results, a new round of testing is normally required. This involves testing the questionnaire at an early stage of its development, revising the questionnaire according to the test results, and then testing the revised questionnaire. This process may be repeated for two, three or even more test phases.

[3] The set of people to be interviewed, the 'sample', is drawn from a larger population, the 'parent population' (also called the 'reference population' or 'parent population').
[4] The information to be collected must be requested (via the questionnaire) in an objective manner. This means ensuring the authenticity of the answers. A questionnaire is considered neutral when it prejudges the possible answers as little as possible.
[5] This criterion therefore refers to the fit of the questions to the characteristics of the questions to the characteristics of the respondents.

Different methods of testing the questionnaire can be used during each phase of the test. In ongoing surveys, the evaluation of previous waves of the survey can make an important contribution to the revision of the questionnaire.

C. Choice of Measurement Scale

The method most adapted by researchers in management science is the scale of measurement which consists of expressing the intensity of its approval and adheres directly to the respondents, proposals to be judged according to a predetermined scale, it includes a statement expressing the state favorable or unfavorable to the object of interest. Each answer reflects, therefore, a degree of satisfaction with the existence of the elements of the theoretical model. A component of the scale is called a modality. The number of modalities can naturally vary, but generally there are 5 or more.

The score for each type of response is calculated by adding up the propositions that are favourable to the dimension being measured. However, this method implies assuming that each item has a weight similar to the average weight.

A pre-test generally ensures that the items are consistent with the variable to be measured. However, it is necessary to supplement this reflection with statistical (factorial) analyses in order to verify the dimensionality of the measurement scale (coherence) and to remove items relating to outliers.

Therefore, we adopted the five-point scales (ranging from 1 to 5 according to the following table), to be used in quantitative studies to achieve consistent measures and to seek a fit between the abstract ideas used to understand the social world, and what happens in the real (empirical) world (W. L. Neuman & Kreuger, 2003). The advantage of such a design lies in its ability to make a mental map of the respondent's evaluation obvious to the researcher (Schindler & Cooper, 2001).

Table 1: Five-point Likert Scale

Likert scale	**Scale coding**
Not at all in agreement	1
No agreement	2
Neither agree nor disagree	3
I agree.	4
Totally agree	5

D. Other forms of items

Generally, a question can be defined as an interrogative sentence that seeks an answer to assist in information gathering, testing and research. Good questions produce accurate answers and help to collect usable quantitative and qualitative data.

Historically, questions have evolved well into other types of questions for gathering information. Although the types of questions used in the design of a research study, these are determined by the information required, the nature of the study, the time needed to answer them and the budgetary constraints of a study. Thus the quality of the questions, i.e. the art of asking the questions, allows for in-depth knowledge to be gained, informed decisions to be made and effective solutions to be developed.

There are several types of questions. The most frequently answered are the dichotomous questions. The latter is usually a closed "Yes/No" question and is used for basic validation. In the context of this thesis work, it was essential to focus on these types of questions, for example this question is used to find out the gender of the respondent, this can allow us to group the data into two groups in order to test moderation for example.

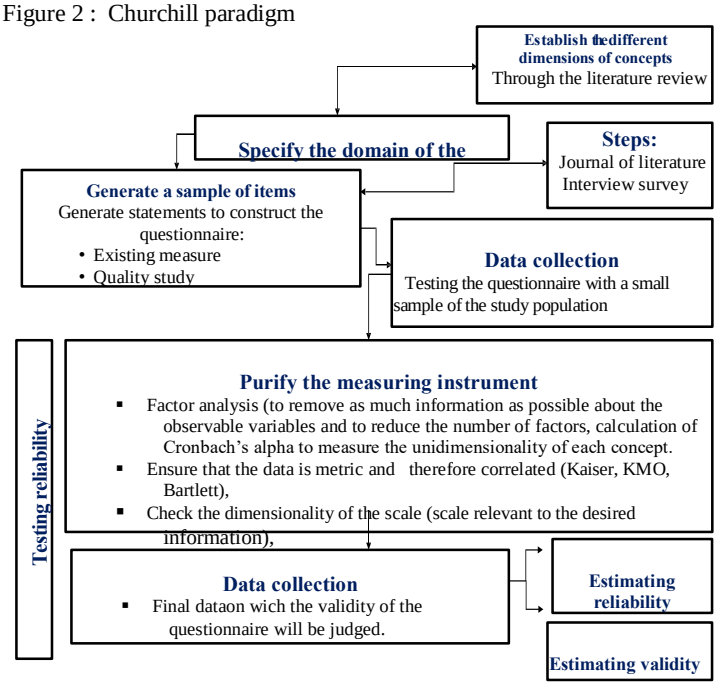
Another type of question is proposed. This type refers to multiple-choice questions. These questions are a type of question in which respondents are asked to select one (single selection multiple-choice question) or more answers (multiple selection multiple-choice question) from a given list of options. The multiple-choice question consists of either a single correct answer, one or more correct answers, incorrect answers, etc. For example, the questionnaire has several single-selection multiple-choice questions such as "What is your highest degree? Here the respondent is asked to provide only one answer from a list of six proposals.

III .METHODOLOGY FOR THE EXPLORATORY EVALUATION OF MEASUREMENT INSTRUMENTS

Based on the methodology presented by the Churchill paradigm, the construction of multi-scale questionnaire-type measurement instruments is based on two distinct but complementary phases:

- An exploratory analysis to explore the measurement instruments (this is what we will see in this work).
- A confirmatory analysis to confirm or refute the research hypotheses.

Figure 2 : Churchill paradigm

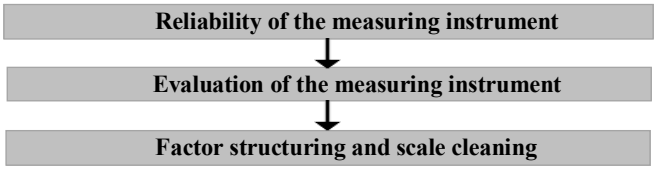


The idea is to first test the questionnaire with a small starting sample. The test consists of purifying the measurement instruments by eliminating items that do not contribute

significantly to the measurement of indicators measuring the concepts.
The exploration therefore consists of :

- Focus on the literature review to identify and specify the domain of the construct.
- The interviews and practical examples enable the researcher to generate indicators for measuring the concept.
- The collection of the first data is used to purify the measurement instrument in order to develop a second version of the questionnaire.

Figure 1: Protocol for the exploratory examination of instruments



In the following paragraphs, it will be interesting to examine the dimensionality and factoring of the measurement scales by applying the principal components method without and with rotation. The objective of this step in the analysis is to further refine the measurement instruments and to reconstruct the variables in such a way that orthogonal relationships between the factors selected for the four variables of this dimension are known. The tests are performed with the *Stata* software package.

A. Reliability analysis

Reliability analysis is used to characterise measurement scales made up of various items - in the case of a questionnaire, questions. The procedure used calculates several measures that allow the reliability of the scale to be assessed and also provides information on the relationships between the different items. In our research, we often speak of internal reliability for measures with several items.
The reliability of a measuring instrument depends on the randomness of the measurement error (Evrard et al., 2003). The lower the randomness, the more reliable the instrument is. To assess the reliability of the measurement scales we have mobilised in this doctoral research, we will use the most widespread indicator in management science research.

Cronbach's alpha coefficient (1951), known as the internal consistency coefficient (measures the correlation between the scale and its various items). It is calculated on the basis of the average variances and covariances of the items, according to the following formula

$$\rho^2(y,t) = \frac{k}{k-1}\left[1 - \frac{\sum_{i=1}^{k} \sigma^2 y_i}{\sigma^2 y}\right] \qquad \text{(Eq.1)}$$

With : $\boldsymbol{yi}$ For each i represent the scores of $\boldsymbol{k}$: Items of the questionnaire. $\boldsymbol{\sigma 2}$: is the variance of the sum of the items equal to the sum of the elements of the variance covariance matrix.
The Cronbach's alpha coefficient[6] can take on several values, from 0 to 1[7]. This index reflects a degree of homogeneity that is all the higher the closer the value is to one. Conversely, a Cronbach's close to 0 indicates that the measurement scale is unreliable. Al though there is no consensus in the literature on research methodology, many authors consider that a Cronbach's value of 0.6 [8] is acceptable for exploratory research (Malhotra, 2004; Mak, 1989; Evrard et al., 1993; Usunier et al., 1993, Peterson, 1995). Evrard et al (2003) consider that an alpha value between 0.6 and 0.8 is acceptable in an exploratory study. On the other hand, Nunnally (1967) and Peter (1979), reach more or less the same conclusion by setting the acceptance interval between 0.50 and 0.60.

However, it is a measure of the internal consistency of the questionnaire. In the first place, a high coefficient can then be interpreted as a good overall consistency of the items within the questionnaire. But a good general coherence does not mean that all the items, taken individually, are coherent with the others. It will be interesting to know the individual consequence of each item, in practice this step is primordial since it is important to know the effect of each item in the group of items constituting the construct or the dimension measured.

Table 3: Decision rule for the reliability of a measurement scale

Value of α	$\boldsymbol{\alpha > 0,8}$	$\boldsymbol{0.6<\alpha<0,8}$	$\boldsymbol{0,5<\alpha<0,6}$	$\boldsymbol{\alpha < 0,5}$
Internal consistence	High	Average	Low	Very low
Decision	***Acceptable***		acceptable in the case of an exploratory study	***Unacceptable***

Various software and packages facilitate the calculations. In particular, STATA, SPSS, R, etc. offer a fairly simple handling of the estimation of *Cronbach*'s alpha, and an easy examination of inter-item correlations and covariances.

B. Analysis of factoring and dimensionality

[6] This coefficient can be interpreted as an estimate of the reliability coefficient of the questionnaire.
[7] In practice, the closer Cronbach's alpha is to 1, the more correlated the items are with each other and the more internally consistent the scale is. In this case, the items can be added together to form an overall score for this scale, as they are supposed to measure the same phenomenon.
[8] Cronbach's Alpha up to levels of 0.7 - 0.8 indicating good internal consistency of the scale and the range 0.8 - 0.9 reflecting excellent consistency between items.

There are many types of factor analysis, but principal component analysis[9] is perhaps the simplest and most widely used. Principal components are synonymous with factors, and generally the components in principal component analysis are often components are often referred to as factors. We will use the two terms interchangeably here. interchangeably here.
Principal component analysis (PCA) is therefore a multivariate technique known as interdependence analysis, which consists of rotating the axes of the factors around the point of origin in order to redistribute the variance to be explained more equitably. This is one of the most commonly used factorial methods. It is particularly suitable for quantitative, continuous variables that are a priori correlated with each other. Once the data are projected into different planes, the proximities between variables are interpreted in terms of correlations, while the proximities between individuals are interpreted in terms of the overall similarity of the observed values.

a. Scaling: PCA factoring without rotation

smaller factors we should ignore, as they explain the least amount of total variance. One of the main criteria used is the *Kaiser* or *Kaiser Guttman* criterion, which consists of ignoring factors whose eigenvalues are equal to or less than one. Since the maximum amount of variance that can be explained by a variable is **one**, these factors are really only equivalent to the variance of a variable.
The first factor will always explain the largest proportion of the overall variance, the second second factor will explain the second largest proportion of variance not explained by the first factor, and so on. factor, etc., with the last factor explaining the smallest proportion of the overall variance. variance. Each variable is correlated with each factor. As the first factor explains the largest proportion of the overall variance, the correlations or loadings of the variables will, on average, be the highest. will, on average, be highest for the first factor, lowest for the second factor, and so on.

In practice, the following steps should be followed:

- PCA without rotation is applied to all the items forming the variable under study.
- We detect the values of "*Eigen value"* which are higher than 1[10] . And make sure by a simulation that its values have not been modified. To decide on the number of principal components with p components or $p-1$ principal components so that the unexplained variability does not increase.
- Retain the principal components for the latent variable.

b. *Dimensionality of measurement scales: PCA with rotation*

The initial (retained) principal components that explain most of the variance in the measurement variables are of the measurement variables are rotated to clarify their significance. The rotation of factors can be done in different ways. We will discuss two of these approaches. of them. The most common form of rotation is called VARIMAX, in which the factors are not related or orthogonal to each other, i.e. the factors are are not related or orthogonal to each other, i.e. the scores of one factor are not correlated with the scores of the other factors. Are not correlated with the scores of the other factors. The VARIMAX rotation attempts to maximise the variance explained by the factors by increasing the correlation of variables that are highly correlated with them and decreasing the correlation of variables that are not. correlated with them and decreasing the correlation of variables that are weakly correlated with them.

The main objective of this method is to obtain a clear structure of factor weights, i.e. factors that are clearly marked by strong correlations with some variables and weak correlations with other variables. In an analysis with many variables, it is necessary to rank the items in terms of decreasing size for each factor in order to see more clearly which variables contribute most strongly to each factor.
In addition, the value of the variance explained by the two rotated VARIMAX factors is the sum or eigenvalue of the squared loadings for each factor, divided by the number of variables. These proportions are naturally different from those of the original principal components without rotation because of the change in the weights of the variables in relation to these factors.

In addition, the value of the variance explained by the two rotated VARIMAX factors is the sum or eigenvalue of the squared loadings for each factor, divided by the number of variables. These proportions are naturally different from those of the original principal components without rotation because of the change in the weights of the variables in relation to these factors.

Ultimately, the following steps should be followed:

- Perform a Principal Component Analysis with rotation: VARIMAX with Kaiser normalisation. (Orthogonal rotation).
- Deduce the factor structure of the latent variable according to the selected components.

C. Structural evaluations of measurement scales

Structural evaluation of measurement scales where the KMO[11] (Kaiser-Meyer-Olkin) test is used. This is a factorial solution adequacy index that measures the inter-correlation of items. The closer the value is to 1, the more *factorizable* the items are, and PCA (Principal Component Analysis) can therefore be applied.
This measure gives an overall picture of the quality of inter-item correlations. The KMO index11 varies between 0 and 1

[9] PCA for principal component analysis
[10] According to the Kaiser rule, only factors with an eigenvalue greater than 1 are retained.
[11] The KMO index is sometimes referred to as the MSA "Measure of Sampling Adequacy" in Anglo-Saxon software. But the French translation "mesure d'adéquation de l'échantillon"

and gives additional information to the examination of the correlation matrix. Its interpretation is as follows:

Table 3: Interpretation rule for the KMO test

Valeur du test KMO	**< 0,5**	**[0,5 ; 0,6]**	**[0,6 ; 0,7]**	**[0,7 ; 0,8]**	**[0,8 ;0,9]**	**> 0,9**
Interprétation	Inacceptable	Insuffisient	Médiocre	Excellent	Méritoire	Excellent

Next, the relationships between the original variables and the components are examined. The stronger this relationship, the more the variable is "explained" by the factor. This relationship, which is expressed by a number between -1 and +1, is called the factor *loading* of the variable on the factor. A variable is only considered to be associated with a factor if its factor loading exceeds **0.3** in absolute value.

In order to improve the structuring of the factors of each of the measurement instruments used, we will carry out a PCA with orthogonal rotation of the VARIMAX type with Kaiser normalisation, generally used when it comes to measurement scales for which independent components are expected (Evrard et al., 2003).

This method makes it possible to obtain a clear factorial structuring by determining for each factor its weight on the different factors (axes). The items are then assigned to the different factors according to their weight. For each factor, only the items with a strong contribution are retained, while ensuring the significance of the percentage of variance explained and Cronbach's alpha coefficient.

IV. APPLICATION

We consider a questionnaire with several questions (items) grouped into theoretical variables. Our application concerns a latent variable named "*Decision evaluations of the line manager"*. And it is measured by six separate questions and structured according to a five-point *Likert* scheme as shown in the following table.

Table 2: Measurement scale for the latent variable "Decision evaluations".

Question/ Secondary issues	Scale of measurement				
Your decisions ...	*Not at all in agreement*	*No agreement*	*Neither disagree nor agree*	*I agree*	*Totally agree*
Are taken ***routinely*** (Q70)	☐	☐	☐	☐	☐
Are taken ***repeatedly*** (Q71)	☐	☐	☐	☐	☐
Are taken ***without involving*** my colleagues or superiors (Q72)	☐	☐	☐	☐	☐
Could be ***automated*** (Q73)	☐	☐	☐	☐	☐
Need ***experience*** and hindsight (Q74)	☐	☐	☐	☐	☐
Are so ***complex*** that they require computer use (Q75)	☐	☐	☐	☐	☐

As a reminder, the approach used for cleaning is the reduction of the random error of the measurement of the concepts studied (Examination of the internal consistency of the items). It is achieved by applying the following two steps: The first step is therefore to :

- ✓ Calculate the Cronbach's alpha coefficient. It is the usual measure of the internal consistency of a set of measurement indicators. It allows to estimate the extent to which an item can weaken the internal consistency[12] of a multiple scale.
- ✓ Remove items that weaken the Cronbach's alpha value below 0.60.

A. Analysis of internal reliability for the variable "Evaluation of decisions".

The objective of this application is to determine the reliability of the measurement instrument we have used to conduct this research. In our case, we use so-called Likert measurement scales.

The aim of this test is to improve its overall quality in order to be able to select *the* "*best*" items that reflect the constructs under study.

```
. alpha  Q70 Q71 Q72 Q73 Q74 Q75 , item

Test scale = mean(unstandardized items)
```

Item	Obs	Sign	item-test correlation	item-rest correlation	average interitem covariance	alpha
Q70	42	+	0.6369	0.4252	.2567364	0.5515
Q71	42	+	0.5815	0.3870	.2848432	0.5703
Q72	42	+	0.4888	0.1828	.3313589	0.6559
Q73	42	+	0.7402	0.5318	.2009872	0.4969
Q74	42	+	0.3929	0.1842	.3553426	0.6328
Q75	42	+	0.6742	0.4496	.2360046	0.5380
Test scale					.2775455	0.6229

The reliability analysis applied to the items of this variable shows that five items are retained instead of six. The value of Cronbach's alpha in this case indicates a value that is close to 0.632 after eliminating item number 74 (Q74), which affirms that the scale presents a satisfactory internal consistency for the constructs.

```
. alpha  Q70 Q71 Q72 Q73  Q75 , item

Test scale = mean(unstandardized items)
```

Item	Obs	Sign	item-test correlation	item-rest correlation	average interitem covariance	alpha
Q70	42	+	0.6393	0.4101	.3547232	0.5687
Q71	42	+	0.5809	0.3701	.3975029	0.5901
Q72	42	+	0.5423	0.2242	.4343786	0.6660
Q73	42	+	0.7707	0.5602	.2496129	0.4808
Q75	42	+	0.6561	0.4036	.3404955	0.5703
Test scale					.3553426	0.6328

[12] The most frequently used method of reliability estimation in psychometrics is the internal consistency method, also known as homogeneity, unidimensionality and internal consistency.

B. Application of the CPA

The evaluation of decisions is a variable defined from two items following the reliability analysis. First, it is interesting to apply a principal component analysis without rotation to the variable "Decision evaluation". This analysis will allow us to identify a clear factor structure with reliable internal consistency.

```
. pca Q70 Q71 Q72 Q73 Q75
Principal components/correlation                  Number of obs    =         42
                                                  Number of comp.  =          5
                                                  Trace            =          5
    Rotation: (unrotated = principal)             Rho              =     1.0000
```

Component	Eigenvalue	Difference	Proportion	Cumulative
Comp1	2.106	.778279	0.4212	0.4212
Comp2	1.32772	.586856	0.2655	0.6867
Comp3	.740861	.280383	0.1482	0.8349
Comp4	.460478	.0955314	0.0921	0.9270
Comp5	.364946	.	0.0730	1.0000

Principal components (eigenvectors)

Variable	Comp1	Comp2	Comp3	Comp4	Comp5	Unexplained
Q70	0.5010	-0.3157	0.4471	0.5310	-0.4092	0
Q71	0.4895	-0.4673	0.0565	-0.2577	0.6873	0
Q72	0.2419	0.7248	0.1696	0.4163	0.4626	0
Q73	0.4970	0.3942	0.1884	-0.6630	-0.3500	0
Q75	0.4514	0.0347	-0.8559	0.1969	-0.1538	0

Indeed, principal component analysis without rotation allowed us to retain two main factors (*Engen-value* values greater than one). The two principal components capture about 68.67% of the overall variability.
The first factor alone captures 36.70% and has a trace value of 1.83. The second factor absorbs 31.97% of the observed variance with a trace value of around 1.59.

The structure retained for the "decision evaluation" variable is a two-dimensional structure. Indeed, principal component analysis without rotation allowed us to retain two main factors (*Engen-value* values greater than one). The two principal components capture approximately 68.67% of the overall variability.
The principal component is exploratory factor analysis is to arrive at a parsimonious conceptualisation of latent traits, by determining the number and nature of a small set of factors explaining the correlation networks among a set of variables (Fabrigar et al, 1999). This is based on what Pohlmann (2004) considers to be the fundamental theorem of factor analysis that the correlation between two variables depends on the similarity of their relationship with the latent factors.

This technique, which aims to synthesise the information, consists firstly of verifying that the dimensions of the construct are clearly identified by common factors that are well distinguished and meet the criteria of convergent and discriminant validity (explaining the variability of the latent variable). And secondly to carry out a principal component analysis without rotation in order to determine the number of factors of the construct. The number of components is equal to the number of items.
After factoring, the results show that this factor is composed of all the items retained from the previous analysis (reliability of the measurement scales). Indeed, after rotation, items Q70, Q71, Q72, Q73 and Q75 are retained for better decision making. In addition, this analysis allowed us to retain a dimensional distribution of these items according to the estimated factors. In particular, items Q70, Q71, Q75 for the first factor and Q72, Q73 for the second factor.

```
. rotate, kaiser blanks(0.3)
Principal components/correlation                  Number of obs    =         42
                                                  Number of comp.  =          2
                                                  Trace            =          6
    Rotation: orthogonal varimax (Kaiser on)      Rho              =     0.5952
```

Component	Variance	Difference	Proportion	Cumulative
Comp1	1.9529	.334718	0.3255	0.3255
Comp2	1.61818	.	0.2697	0.5952

Rotated components (blanks are abs(loading)<.3)

Variable	Comp1	Comp2	Unexplained
Q70	0.5256		.4124
Q71	0.6012		.3084
Q72		0.7192	.1797
Q73		0.6219	.2692
Q74	0.3876		.719
Q75	0.3674		.5403

```
. pca Q70 Q71 Q72 Q73 Q75 , component(2)
Principal components/correlation                  Number of obs    =         42
                                                  Number of comp.  =          2
                                                  Trace            =          5
    Rotation: (unrotated = principal)             Rho              =     0.6867
```

Component	Eigenvalue	Difference	Proportion	Cumulative
Comp1	2.106	.778279	0.4212	0.4212
Comp2	1.32772	.586856	0.2655	0.6867
Comp3	.740861	.280383	0.1482	0.8349
Comp4	.460478	.0955314	0.0921	0.9270
Comp5	.364946	.	0.0730	1.0000

Principal components (eigenvectors)

Variable	Comp1	Comp2	Unexplained
Q70	0.5010	-0.3157	.3391
Q71	0.4895	-0.4673	.2053
Q72	0.2419	0.7248	.1792
Q73	0.4970	0.3942	.2734
Q75	0.4514	0.0347	.5692

C. Overall assessment: KMO test

The output above shows that the goodness-of-fit index for this factorial solution is 0.59. This value seems poor when referring to the acceptance thresholds of the KMO test. This value seems poor when referring to the acceptance thresholds of the KMO test. However, it seems that the set of items selected can constitute a coherent double structure allowing

```
. estat kmo
Kaiser-Meyer-Olkin measure of sampling adequacy
```

Variable	kmo
Q70	0.6404
Q71	0.5522
Q72	0.4452
Q73	0.6088
Q75	0.7513
Overall	0.5897

to approach the variable "Evaluation of decisions".

V. Conclusion

This methodological work par excellence outlines the different stages required to carry out an exploratory factor analysis, the importance of which lies in the choice of techniques and methodology to be adapted. Indeed, it is important to present the research tool that is based on a questionnaire survey. In developing the questionnaire, the whole cycle of designing and testing the questionnaire (pre-test) was covered. The questionnaire went through several stages including the questionnaire design phase which started after the specification of the conceptual basis and the determination of the data collection mode. Thus, the factor analysis approach allows the identification of latent factors from directly measured variables. That is, it defines each latent variable (construct) by associating a number of measured variables (item) with it.

The exploratory study involved another analysis that falls within the framework of factorial component analysis, which aims to find the factors that summarise the data or their characteristics. The approach consists of conducting the three main validity stages most commonly used: PCA without rotation, PCA with varimax rotation and the KMO test. At this level, the questionnaire is considered purified and valid for a second administration on a larger sample.

REFERENCES

[1] A. David. (2000), Logique épistémologie et méthodologie en sciences de gestion : trois hypothèses revisitées, David, A., Hatchuel. and Laufer, R. Les nouvelles fondations des sciences de gestion, Paris, Vuibert, p. 83-109.

[2] F. Pearson. (2012), Méthodologie de la recherche en sciences de gestion-Marie-Laure Gavard-Perret, David, G. Christophe, H. and Alain Joliber, Méthodologie de recherche, Raymond-Alain Thietart et al, 4th edition, DUNOD.

[3] J. Barbet. (1988), Faut-il encore parler d'approche qualitative et d'approche quantitative, Recherche et Applications En Marketing, 3(1), 75-89.

[4] J. Cronbach. Coefficient *alpha and the internal structure of tests*, Psychometrika, Springer Science+Business Media, vol. 16, September 1951, pp. 297-334 (ISSN 0033-3123 and 1860-0980, DOI 10.1007/BF023105559.

[5] J. Hair., R. Anderson., Tatham, R. and Black, W. (1998), Multivariate data analysis (5 th ed), New York: Prentice Hall.

[6] L. J, Cronbach. (1951), Coefficient alpha and the internal structure of tests, Psychometrika, 16, 297-334.

[7] J. R, Busemeyer., L. E, Jones. (1983), Analysis of multiplicative combination rules when the causal variables are measured with error, Psychological Bulletin, 93, 549-563.

[8] L. V, Campenhoudt., R, Quivy. (2011), 1st step: the starting question, in Handbook of Social Science Research (Vol. 4, pp. 24-37).

[9] P. Roussel. (2005), Méthodes de développement d'échelles pour questionnaires d'enquête, Management des ressources humaines, méthodes de recherches en sciences humaines et sociales, edited by Roussel, P., Wacheux, F., Brussels.

[10] R. Cudeck. (2001), Cronbach's Alpha on two-item scales, Journal of Consumer Psychology, 10 (1/2), 55-69.

[11] T. Jolliffe. (1986). Principal component analysis. Springer-verlag.

[12] T. W, Anderson. (1963). Asymptotic theory for principal component analysis. Annals of Mathematical Statistics 34, 122-148 - Hotelling, H. (1933). Analysis of a complex of statistical variables into principal components. Journal of educational psychology, 417-441.

linking the innovation efforts carried out in the university to regional development

Nabil BOUKHNIEF

Systems Engineering Laboratory
Sultan Moulay Slimane University
n.boukhnief@usms.ma

Youmna EL HISSI

Systems Engineering Laboratory
Sultan Moulay Slimane University
y.elhissi@usms.ma

Hanaa HACHIMI

Systems Engineering Laboratory
Sultan Moulay Slimane University
hanaa.hachimi@usms.ma

***Abstract*—** In an economic context characterized by rapid change and the appearance of new production models based on knowledge and knowledge, the links between the university and the regions must be strengthened. The modern economy based on knowledge mobilizes various knowledge developed by fundamental research. Since their foundation, Universities have been responsible for the creation and transmission of knowledge. It is therefore the university laboratories that prepare the competitiveness of tomorrow.

Keywords: Globalization, university research technology transfer

INTRODUCTION

Globalization has brought about a profound change in economic development throughout the world, which is leading organizations responsible for promoting development to adopt a regional approach based on the sharing of knowledge and know-how. Globalization is also increasing the focus on innovation, which is now the main driver of economic growth – both for companies and for regions and countries. However, universities are located at the intersection of these two deep trends: anchored at the regional level, they are undoubtedly also the main driver of innovation.

This article analyzes the relationship between the university and its regional environment by trying to answer the following question: How to link the innovation efforts carried out in the university to regional development?

IDENTIFYING COMPETITIVE ADVANTAGE

The identification of innovations essential to the economic prosperity of regions is based on the idea that regions can only develop by first identifying their own competitive advantage. And to do this, regions must develop a regional strategy that is the cornerstone of viable public policy and effective local action. To define this strategy, the regions face several obstacles. The first is the lack of skills and tools to identify own competitive advantage. To achieve this, regional leaders must first be provided with the capacities and skills required for this purpose. But the phase of identifying competitive advantage does not end once regional leaders have developed a regional strategy. The development of a regional strategy presupposes the existence of one of the "regional governance" stakeholders, capable of ensuring the supervision and management of this strategy. regional governance makes it possible to channel the initiatives carried out in these different sectors. The development and implementation of a regional strategy comes with many challenges for regions, which must not only identify the specific assets they have and study the markets they can exploit, but also ensure the convergence of investments. public and those from the private sector. The experience acquired in this area nevertheless suggests that regional governance is based on the appointment of a catalytic and unifying body capable of organizing a round table between the various regional actors, and on the broader commitment of the leaders of the public and to define so precisely the innovations that regions need to fully achieve economic prosperity. Universities can play this role by contributing to this process: indeed, they often have regional

economic expertise that can facilitate the identification of the competitive advantage of their region.

INNOVATIONS MADE BY THE UNIVERSITY

The importance of innovation as a driver of economic growth and wealth creation is widely recognized. As a result, governments' understanding of the role of universities has evolved. They now expect universities to make a major contribution to advancing innovation and economic development.

Currently, universities archive the results of their research activities for funding purposes. However, these research results are not always adapted to regional development strategies, nor easily accessible to people who do not belong to the research community. It would therefore be necessary to begin by reforming these catalogs, taking into account the needs of other players in the regional innovation market.

There is therefore a gap to be bridged between universities, able to provide innovation thanks to the results of research carried out, and the regions which need the innovations. Universities are often uninformed of regional development needs or competitive advantages. Universities build bridges to the regions. In most cases, these gateways do not adequately meet development needs. As we pointed out in the previous section, universities must first contribute to the development of the regional strategy. It is therefore necessary to set up a mechanism for the “centralization” of information. A market works because it offers buyers and sellers a neutral meeting place. The primary need is to enable universities and regions to exchange information. It can simply be a matter of creating a virtual meeting place, allowing universities to put their “catalogue” of innovations online, and regions to share their “innovation needs”. Such an initiative would certainly contribute to regional development.

ACADEMIC RESEARCH AND KNOWLEDGE TRANSFER

Since the early 1980s, public authorities and universities have multiplied initiatives to strengthen their knowledge transfer capacity, in particular the transfer of inventions and discoveries from the academic world to industry. “Research commercialization” refers to the process of transforming scientific discoveries and inventions into marketable products and services. Generally, the results of university research are commercialized by licensing patents to companies. The term "technology transfer" has a number of specialized meanings, but basically it refers to "the movement of know-how, technical knowledge or technology from one organization to another" (Bozeman, 2000, p. .629). This term is most commonly used in the context of the transfer of inventions and related know-how from research organizations (especially universities and public research institutions) to users. “Knowledge transfer” is a more recent term used mainly in Europe to refer to the transfer of knowledge and academic expertise to research users. Although the results of research are of crucial importance, the term "knowledge transfer" is often used in a broader sense, encompassing academic knowledge of all kinds, whether or not it has commercial value, and the transfer to larger user groups than just companies.

The commercialization of research is based on intellectual property rights (IPRs). IPRs reward investment in R&D by granting ownership to inventors, their employers, those who have funded the research, or a combination of these actors.

Invention licensing and business creation are not the only mechanisms for commercializing university research as both graduate students and teachers bring knowledge from universities to companies while the latter access developed knowledge by universities through sponsored research, conferences and academic journals (Sizer, 2002). But increasingly, licensing and company creation are seen as the two key mechanisms for commercializing academic research.

CONCLUSION

We have emphasized in this article the interaction between the university and the region as well as industry and on the regional role assigned to universities which has enabled them to forge stronger links with regional bodies and encouraged them to develop services that meet the region's needs in terms of innovation. It should be noted that those responsible for developing the regional strategy must create conditions that allow both universities and individuals to enjoy a great deal of autonomy, because they will thus be able to exercise their entrepreneurial spirit more easily. It is essential that officials, in addition to supporting access to information, support the strengthening of policies and initiatives in this area, the development of best practices and the creation and creation of an environment conducive to collaboration between the university and other regional players.

REFERENCES

Barkley, D., M. Henry et D. Lee (2006), « Innovative Activity in Rural Areas: The Importance of Local and Regional Characteristics », Community Development

Investment Review, vol. 2, n° 3, Banque de Réserve Fédérale de San Francisco.

Lyons, T.S. (2004), « Policies for Creating an Entrepreneurial Region », Main Streets of Tomorrow: Growing and Financing Rural Entrepreneurs, compte rendu de conférence, Banque de Réserve Fédérale de Kansas City, pp. 97-105.

Porter, M.E. (1999), « New Strategies for Inner-City Economic Development », in J. Blair et A. Resse (éd.), Approaches to Economic Development, Sage Publications, Thousand Oaks, Californie, pp. 32-47.

BOUCHER, G., C. CONWAY, et E. VAN DER MEER (2003), « Tiers of Engagement by Universities in their Region's Development », Regional Studies, Vol. 37, No. 9, pp. 887-897.

CLARK, B. (1998), Creating Entrepreneurial Universities: Organisational Pathways of Transition, Oxford: Pergamon Press.

LEYDESDORFF, L. et H. ETZKOWITZ (2001), « The Transformation of University- Industry-Government Relations », Electronic Journal of Sociology.

Impact of the COVID-19 pandemic on Financial Reporting via Internet - A Morocco case study

Imane Jabrane
Communication Research Team (ERC)
Laboratory, University of Sultan
Moulay Slimane - Beni Mellal,
Morocco.

Toufik Majdi
Communication Research Team (ERC)
Laboratory, University of Sultan
Moulay Slimane - Beni Mellal,
Morocco.

Issam Izza
Communication Research Team (ERC)
Laboratory, University of Sultan
Moulay Slimane - Beni Mellal,
Morocco.

***Abstract*—** This paper analyzes the impact of COVID-19 on the financial reporting quality of Moroccan companies and whether decision-making technologies have a mitigating effect. The pandemic has caused a distress around the world, with profound economic and social risks. Especially, Moroccan firms have been negatively impacted, the first and most obvious impact from the crisis was their stability and their financial reporting quality which has been lower during the pandemic. Our paper aims to observe the important role of the financial reporting quality which facilitates the efforts made by corporate governance in order to support Moroccan firms in this COVID-19 crisis. Also, we report that financial stability and Moroccan firm's performance are negatively affected by daily new confirmed COVID-19 cases. The paper documents evidence to support the role of financial communication reporting in order to help economic stakeholders and corporate governance to make strategic decision during the pandemic. To that end, this study investigates the time series and examines the role of financial communication reporting during the COVID-19 pandemic and to check whether there are changes in Moroccan firm's efficiency. The results of this study provide useful suggestions to the practice.

Keywords— Financial reporting quality, COVID-19, time series Analysis, Financial Stability, Morocco.

I. INTRODUCTION

The impact of coronavirus (COVID-19) pandemic is the reason why this period will not leave our memories for decades. The COVID-19 pandemic has caused a distress around the world, with profound economic and social risks (Khan, 2022)(Didier et al., 2021a). Economists have drawn attention to financial stability and firm's performance as an important factor in every economic development assessment. Companies are facing great difficulties in dealing with the impact of the pandemic and its effects have varied among many economic sectors (Didier et al., 2021b). More specifically, companies of varying sizes that are the mainstay of any economy, were the most affected, given their relatively weak position in times of health crises that turned into an economic crisis, especially their resort to bank financing or commercial loans (Hasanov et al., 2022).

Furthermore, the COVID-19 economic crisis has significantly presented new challenges for companies requiring them to adopt new strategic approaches and choices in the light of financial communication reporting. Thus, understanding the company policies and make informed strategic decisions for economic stakeholders(Gerth et al., 2021). Moroccan SMBs (Small and Medium-Sized Businesses) and SMEs (Small and Medium Enterprises), like the rest of the world, need to revise their existing financial reporting quality to take advantage of the power of decision-making technologies, to better connect with the various economic sectors' stakeholders. A great number of Moroccan firms have been negatively impacted, the first and most obvious impact from the crisis was their stability and their financial reporting quality which has been lower during the pandemic (bkam, n.d.). Since good financial reporting quality helps financial economic sectors' stakeholders to make proper

decisions and helps to improve the efficiency of financial markets. Also, it provides useful insight for policymakers and investors.

The financial communication quality is important to keep the efficiency and stability of Moroccan SMBs and SMEs, because of the national economy actors, such as investors, lenders and regulators who directly rely on financial information from financial reporting to take strategic decisions (Bajkó et al., 2022; Pu et al., 2021). Also, it should be reviewed on a daily, weekly, monthly, quarterly, half-yearly and yearly basis, with the most performance decision-making technologies, as it gives the ability to plan and make it possible for intervening in a timely manner and in the event of any crisis. A series of studies examined the impact of the 2008 financial crisis on the quality of financial information (Cimini, 2015; Eng et al., 2019). Like the 2008 financial crisis, COVID-19 has caused significant disruption to financial information and the global economy.

However, it is still unknown whether the financial reporting quality moderates the impact of COVID-19, which can identify the strengths and weaknesses of the Moroccan company through financial communication indicators. Moreover, it can actively participate in the efforts made by corporate governance in order to support Moroccan SMBs and SMEs. This study addresses the effect of the COVID-19 pandemic on financial reporting quality, it aims to contribute to the ongoing discussion of the pandemic and to discuss the role of corporate governance in the COVID-19 crisis.

In this context, the present study contributes to the growth literature on the economic effects of COVID-19 in general, and the Moroccan SMBs and SMEs-level consequences of the pandemic in particular. We analyze the impact of COVID-19 on financial stability and firms' performance of Moroccan SMBs and SMEs by drawing on data from the World Bank Enterprise Analysis conducted in many countries prior to the COVID-19 outbreak, since Morocco is one of them. More specifically, this study aims to observe the very important role of the financial reporting quality which facilitate the efforts made by corporate governance in order to support Moroccan SMBs and SMEs in this COVID-19 crisis. Also, we report that financial stability and firms' performance of Moroccan SMBs and SMEs are negatively affected by daily new confirmed COVID-19 cases. The study documents evidence to support the important role of financial communication reporting for helping economic stakeholders and corporate governance to make strategic decision during the pandemic. To that end, this study investigates the time series forecasting and regression econometric models examines the role of financial communication reporting during the COVID-19 pandemic and to check whether there are changes in Moroccan SMBs and SMEs efficiency. The results of this study provide useful suggestions to the practice.

The remainder of this paper is structured as follows. Section 2 reviews relevant related works and develops hypotheses. Section 3 describes the methodology and data. Section 4 presents our results and Section 5 concludes.

II. LITERATURE AND HYPOTHESES

Previous studies of public health crises have indicated that many pandemics such as MERS, Ebola, and Zika virus have damaged the economic growth and caused disruptions in many activities (Pagnottoni et al., 2021) (Peixoto et al., 2021) (Cimini, 2015). While studies of the economic growth to those epidemics may provide some insightful information, given the special nature of the current COVID-19 pandemic, these past experiences may have little compared to the unique challenges of the current pandemic (Barro et al., 2020).

Studies motivated by the huge impacts of COVID-19 have primarily focused on the financial reporting quality which becomes less value-relevant during the pandemic. In (Baek et al., 2020), the authors analyze the effects of COVID-19 on US stock market volatility at the industry level. The natural gas and oil, hotel, food service and lodging sectors have showed large increases in risk. They have used machine learning models to identify influential economic indicators. In (Chiah & Zhong, 2020), the authors examine the impact of COVID-19 on trading volume in stock markets around the world. They have shown that the increase in the volume of transactions is related to the national culture and the institutional environment of each country. Also, the intensity of trade is also associated with the gambling opportunities of many

countries. State-level data was used by the authors (Pham et al., 2021) to assess the relationship between COVID-19 outbreaks and stock returns. they have been found that daily increases in the number of infected cases, hospitalized cases, and deaths are negatively associated with the next day's stock market returns of companies headquartered in the same state. The association is less pronounced between states with high levels of medical resources and states likely to have federal government support.

The literature on the 2008 financial crisis shows that crises can lead to higher or lower quality of financial information. The same is true during the COVID-19 pandemic (Hsu & Liao, 2021). If companies exploit the higher information, they are more likely to manage their profits to influence their stakeholders and investors in order to mitigate the negative effect of COVID-19. Thus, we work up the following hypotheses:

Hypothesis 1: The quality of financial reporting is lower/higher during the COVID-19 pandemic.

Hypothesis 2: The decrease in cases of COVID-19 has a mitigating effect on financial reporting quality during the pandemic.

III. Methodology and Data Description

The deployed methodology is effectively used to analyze time series of Morocco's health situation with financial reporting quality. They investigated the time series and examines the role of financial communication reporting during the COVID-19 pandemic and to check how there are changes in Moroccan firms' efficiency. This could help decision makers to design strategies and policies to manage its spread.

In this study, we focus on Morocco because it is one of the emerging economies in Africa. Furthermore, Morocco is known as a developing regional economic power and our research will add to current literature.

We have used the World Bank Enterprise Surveys and COVID-19 Impact Follow-up Surveys as our data sources. They are standardized and representative surveys at the national and company level on a sample of non-financial companies in the private sector of a country. The methodology used is a random sampling methodology and is followed in all these surveys using a master questionnaire. The classes used to construct a representative sample in each country are economic region, industry and firm size. Since these surveys use a global approach based on a common questionnaire, the results are fully comparable by period which facilitates research, analysis, including comparisons. Another important feature is that they collect information on the business environment in addition to a wide range of business characteristics and financial results. The second data source is collected by many sources including the world health organization, worldometers and from data delivered by the Moroccan ministry of health. The dataset is maintained in a CSV format. The data have cleaned and normalized data and made it easy for further processing and analysis, arranging dates and consolidating several files into normalized time series.

IV. Results

The descriptive statistics, reported in Figure 1, daily cases of COVID-19 are likely stabilized at the beginning of March 2020 with a small margin due to the strict measures taken by the Moroccan authorities. In late July, as authorities began to ease these measures, cases began to rise exponentially this time due to increased population movement and travel during the summer. Faced with the exponential growth of cases and deaths, governments have taken emergency measures to reduce the spread of the disease. By flattening the epidemic curve, it was expected to reduce the risk of breaking the capacity of health care systems. As a result, these emergency measures have reduced economic flow through all kinds of businesses. The quarantine has also led to an increase in the unemployment rate and the closure of businesses, making social inequalities more evident.

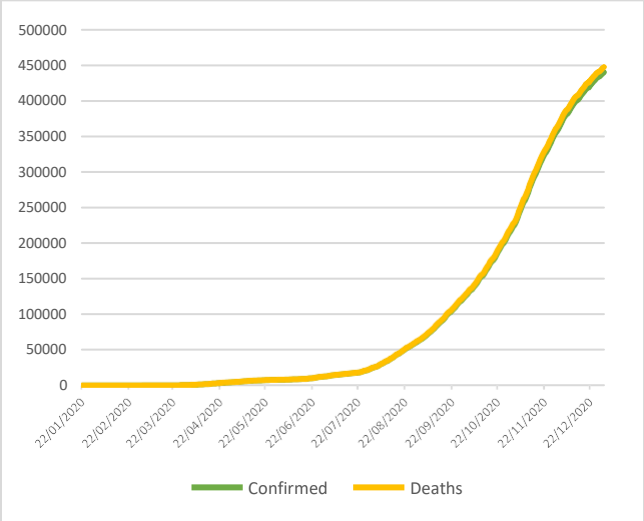


Figure 1: Total COVID-19 Cases in Morocco over time.

Figure 2 shows an important measure of the impact of the pandemic on companies that exited the market during the economic crisis. The figure shows estimates about businesses that closed altogether. It shows the share of Moroccan companies whose closure has been confirmed since the declaration of the pandemic in Morocco. Thus, the pandemic shock has been closed permanently firms of different sizes as shown in figure 2.

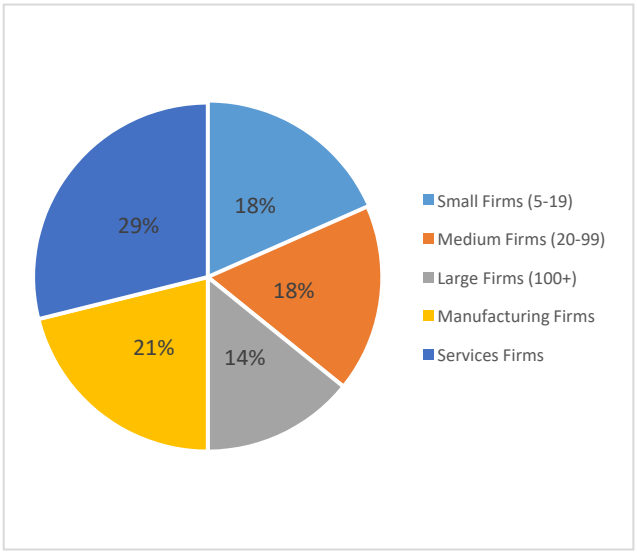


Figure 2: Percent of firms confirmed permanently closed.

As economies around the world were increasingly impacted by the COVID-19 pandemic, Moroccan companies experienced high financial distress. Figure 3 shows the share of Moroccan companies delaying payments to owners, suppliers or tax authorities for more than a week due to the covid-19 pandemic. This statistic gives an idea of the impact of the liquidity or solvency crisis induced by covid-19.

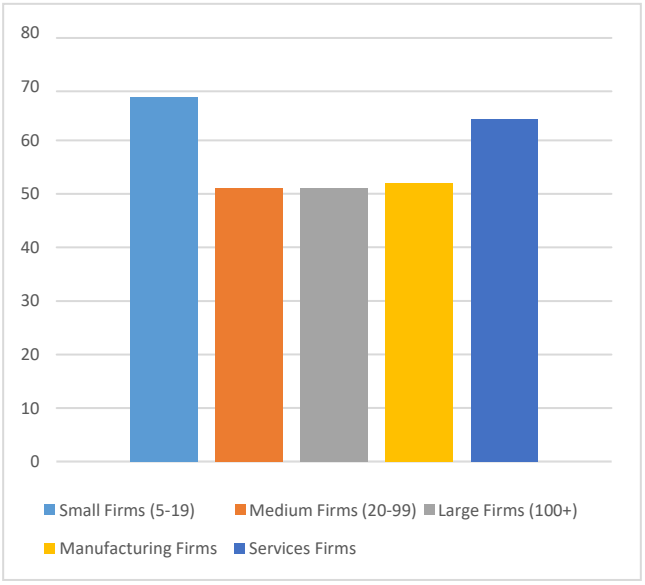


Figure 3: Share of firms delaying payments for more than one week due to COVID-19 (%).

V. Conclusion

Given the seriousness of the situation of recording thousands of daily cases of COVID-19 in Morocco, our paper adds to ongoing contributions on COVID-19 and shows important evidence which a larger board can help mitigate the negative impact of COVID-19 on the financial reporting quality. A good financial information quality can help financial market investors or stakeholders to make the right decisions and help improve the efficiency of financial markets during the pandemic, our results provide useful information for policy makers and investors. This work documents evidence to support the important role of financial communication reporting that can help economic stakeholders and corporate governance to make strategic decision in the pandemic. It investigates the time series and examines the role of financial communication reporting of Moroccan firms' efficiency during the COVID-19 pandemic.

References

Baek, S., Mohanty, S. K., & Glambosky, M. (2020). COVID-19 and stock market volatility: An industry level analysis. *Finance Research Letters*, *37*, 101748. https://doi.org/10.1016/j.frl.2020.101748

Bajkó, N., Fülöp, Z., & Pércsi, K. N. (2022). Changes in the Innovation- and Marketing-Habits of Family SMEs in

the Foodstuffs Industry, Caused by the Coronavirus Pandemic in Hungary. *Sustainability*, *14*(5), 2914. https://doi.org/10.3390/su14052914

Barro, R., Ursúa, J., & Weng, J. (2020). *The Coronavirus and the Great Influenza Pandemic: Lessons from the "Spanish Flu" for the Coronavirus's Potential Effects on Mortality and Economic Activity*. https://doi.org/10.3386/w26866

bkam. (n.d.). *Rapport annuel 2020*.

Chiah, M., & Zhong, A. (2020). Trading from home: The impact of COVID-19 on trading volume around the world. *Finance Research Letters*, *37*, 101784. https://doi.org/10.1016/j.frl.2020.101784

Cimini, R. (2015). How has the financial crisis affected earnings management? A European study. *Applied Economics*, *47*(3), 302–317. https://doi.org/10.1080/00036846.2014.969828

Didier, T., Huneeus, F., Larrain, M., & Schmukler, S. L. (2021a). Financing firms in hibernation during the COVID-19 pandemic. *Journal of Financial Stability*, *53*, 100837. https://doi.org/10.1016/j.jfs.2020.100837

Didier, T., Huneeus, F., Larrain, M., & Schmukler, S. L. (2021b). Financing firms in hibernation during the COVID-19 pandemic. *Journal of Financial Stability*, *53*, 100837. https://doi.org/10.1016/j.jfs.2020.100837

Eng, L. L., Fang, H., Tian, X., Yu, T. R., & Zhang, H. (2019). Financial crisis and real earnings management in family firms: A comparison between China and the United States. *Journal of International Financial Markets, Institutions and Money*, *59*, 184–201. https://doi.org/10.1016/j.intfin.2018.12.008

Gerth, F., Ramiah, V., Toufaily, E., & Muschert, G. (2021). Assessing the effectiveness of Covid-19 financial product innovations in supporting financially distressed firms and households in the UAE. *Journal of Financial Services Marketing*, *26*(4), 215–225. https://doi.org/10.1057/s41264-021-00098-w

Hasanov, F. J., Javid, M., & Joutz, F. L. (2022). Saudi Non-Oil Exports before and after COVID-19: Historical Impacts of Determinants and Scenario Analysis. *Sustainability*, *14*(4), 2379. https://doi.org/10.3390/su14042379

Hsu, Y.-L., & Liao, L.-K. (Connie). (2021). Corporate governance and stock performance: The case of COVID-19 crisis. *Journal of Accounting and Public Policy*, 106920. https://doi.org/10.1016/j.jaccpubpol.2021.106920

Khan, S. U. (2022). Financing constraints and firm-level responses to the COVID-19 pandemic: International evidence. *Research in International Business and Finance*, *59*, 101545. https://doi.org/10.1016/j.ribaf.2021.101545

Pagnottoni, P., Spelta, A., Pecora, N., Flori, A., & Pammolli, F. (2021). Financial earthquakes: SARS-CoV-2 news shock propagation in stock and sovereign bond markets. *Physica A: Statistical Mechanics and Its Applications*, *582*, 126240. https://doi.org/10.1016/j.physa.2021.126240

Peixoto, H. M., Gallo, L. G., Nascimento, G. L., & de Oliveira, M. R. F. (2021). Economic impact of Zika virus infection and associated conditions. In *Zika Virus Biology, Transmission, and Pathology* (pp. 63–74). Elsevier. https://doi.org/10.1016/B978-0-12-820268-5.00006-7

Pham, A. V., Adrian, C., Garg, M., Phang, S.-Y., & Truong, C. (2021). State-level COVID-19 outbreak and stock returns. *Finance Research Letters*, *43*, 102002. https://doi.org/10.1016/j.frl.2021.102002

Pu, G., Qamruzzaman, Md., Mehta, A. M., Naqvi, F. N., & Karim, S. (2021). Innovative Finance, Technological Adaptation and SMEs Sustainability: The Mediating Role of Government Support during COVID-19 Pandemic. *Sustainability*, *13*(16), 9218. https://doi.org/10.3390/su13169218

AUTHORS INDEX

ABOUT THE EDITOR IN CHIEF

Prof. Dr. Hanaa Hachimi, Ph.D in Applied Mathematics & Computer Science and a Ph.D in Mechanics & Systems Reliability, Secretary General of Sultan Moulay Slimane University in Beni Mellal. President of the Moroccan Society of Engineering Sciences and Technology (MSEST). I am Associate Professor at the Sultan Moulay Slimane University (USMS) of Beni Mellal, Morocco. I am Director of the Systems Engineering Laboratory (LGS) and IEEE Senior Member, precisely I am affiliated at the Big Data, Optimization, Service and Security (BOSS) team at USMS. I am Lecture and Keynote Speaker of the courses: Optimization & Operational Research, Graph Theory, Statistics, Probability, Reliability and Scientific Computing. I am Member of the Moroccan Society of Applied Mathematics (SM2A).

Previously Associate Professor at Ibn Tofail University, National School of Applied Sciences of Kenitra, Morocco.

for more information, visit our website : http://www.hanaahachimi.com/

EDITORIAL BOARD

Editorial bord:


IJOA
International Journal
on Optimization and Applications