Title: Exploring data conditions to improve business performance

Abstract:

Past researches drew from the industrial organization perspective have examined the role of the data to generate competitive advantage. Their analysis show data is a valuable resource that can leverage business partnerships, vertical integration or diversification. The emergence of data science has created new opportunities to understand better clients' needs and to manage more efficiently the organizations processes. Nevertheless, if data analytics represent an enormous potential, many organizations are still looking the conditions to obtain value from them. Our study contributes to this topical subject analysing the relationship between different combinations of data conditions and the company performance that we measure through the Customer management and Provider operations efficiency. Our methodology is novel compared to previous researches which are based in linear algebra. It is based on the use of a fuzzy-set qualitative comparative analysis (fsQCA) which allows to reveal multiple paths to achieve the possible outcomes. Our results show that the consistency, completeness and protection of the data along with a data-driven company profile are different possible solutions to a better Customer management and Provider operations efficiency. Our conclusions allow practitioners to uncover the strength of the data in the hopes of solving many of their business performance concerns.

Keywords: data science, competitive analytics, qualitative comparative analysis, business performance

1. Introduction

Industry competition has been the focus of research for many strategy scholars. In the origin, Schumpeter (1934) describes the dynamic market process through which firms compete in a 'perennial gale of creative destruction' and depicts a disequilibrium leading to shifts in market status quo. The capacity to obtain data on the markets or customers can improve organizations to adapt themselves to the environment changes and improve their position against competitors who may be less informed or unable to adapt rapidly to the required changes. Consequently, practitioners and scholars have started to develop theories that permit the acquisition of data and propose models of

firm strategy (Barney, Wright, & Ketchen, 2001). Very-known models such as resource based value (Barney, 1991), strategic groups (Newman, 1978), five forces (Porter, 1991) have had a strong influence on strategy providing the guidelines for data collection and use. Their logical justification is based that the ownership of a superior resource is the critical source of competitive advantage (Barney, 1986). They aim to analyse the industry, the different stakeholders involved in the marketplace and their respective bargaining power and to look for how the organization can develop resources to successfully compete in its external environment (Lim, Stratopoulos & Wirjanto, 2011).

Data-centric approaches have become increasingly popular in a wide variety of seemingly unrelated research fields such as materials science, medicine, astronomy, chemistry or even business management as it was reported in one of the leading journals Science (Reed, 2011). Recent progress in the technology of experimentation and measurement makes it possible to obtain a huge amount of high-dimensional data (Igarashi et al., 2016). Data sets have so exploded in the number of observations and dimensionality which have led to investigate in the modelling of high dimensional data. Effective use of high-dimensional data requires sparse data sets which arise from many important areas involving human-computer interaction e.g. the patient electronic health record in Health care (Wang, Zhou, & Hu, 2014) or human-human interaction e.g. social networks (Purdy, 2012). The utility of sparse models (SpM) have been demonstrated in this context as a key technology of data-driven science (Igarashi et al., 2016).

Nevertheless, we decide to use the term of data science as it is coined in the research field of Strategic Management i.e. the extensive use of data in the aim of company innovation, competition, and productivity (Davenport & Harris, 2007; Grimaldi,

Fernandez & Carrasco, 2018; Harris & Craig, 2011; Waller & Fawcett, 2013). McAfee & Brynjolfsson (2012) and Evans (2013) confirm that executives across many industries are allocating resources into data science projects with the aim to better monitor, measure and manage their organizations and in the hopes of solving many of their long-standing operational concerns (Forbes, 2014). Nevertheless, if data science represents an enormous potential and essence of firm capabilities, many organizations are still looking how to obtain value from them and get sources of competitive advantage (Woerner & Wixom, 2015). This study analyses the relationship between the use of data science and the firm performance.

The literature is developing toward a perspective of firms as complex systems of interdependent characteristics and choices in which competitive advantage frequently does not rest on a single attribute but, instead, resides in the relationships and complementaries between multiple characteristics (P. C. Fiss, 2007; Miller, 1986). An understanding of drivers of firm performance requires the acknowledgment and the approach of the complexity of firms and their environment. The notion of organizational configurations stresses this idea by suggesting that "organizational structures and management patterns are best understood in terms of overall patterns rather than in terms of analyses of narrowly drawn sets of organizational properties" (Meyer, Allen, & Smith, 1993)

Accordingly, the method of analysis is different from those used in previous research in supposing that different combinations of casual conditions may be individually linked to firm performance (Davenport & Harris, 2007). Therefore, instead of structural equation modelling based on partial least squares analysis (Ren, Wamba, Akter, Dubey & Childe, 2017; Kwon, Lee, & Shin, 2014), our study uses configurational comparative methods to reveal multiple paths to explain the benefits perceived in the management of the

company processes. In the configurational analysis, therefore, the focus shifts from the net effect of a single characteristic on performance to the analysis of multiple configurations associated with high performance. Traditional multivariable analytical methods are frequently less adept at capturing complex systems of interdependencies among the elements of a configuration and outcome variables. However, the development of a theory of configurational approaches is scanty in research on firm performance and this study aims at filling this gap.

Contrary to previous statistical researches, our study analyses the relationship between different combinations of conditions of data and two different outcomes related to the value of this use. The structure of our study is as follows. Section 2 presents the theoretical background while section 3 shows the methodology used. Section 4 presents and analyses the results. Section 5, finally, develops a discussion and suggests future researches.

2. Theoretical background

2.1 Business Performance

Lavalle, Lesser, Shockley, Hopkins & Kruschwitz (2011) by running in 2011 a survey to business executives across the globe show that companies who consider themselves as top performers use analytics twice than those who consider themselves as lower ones. By data analytics, we employ Davenport & Harris (2007) 's definition, i.e., 'the extensive use of data, statistical and quantitative analysis, explanatory and predictive models, and fact-based management to drive decisions and actions. Analytics are a subset of what has come to be called business intelligence'. Lavalle et al. (2011) conclude that analytics capabilities and business performance are correlated. Their study uses regression model for data analysis analysing which variables account for deviations from the ideal organizational configuration displayed by top-performing firms.

Although such correlation-based approaches are useful for examining the relative contribution of different elements, they face considerable challenges in modelling the ways that cause may combine rather than compete in providing with the expected outcomes (P. C. Fiss, 2007; Ragin & Strand, 2008). We believe that shifting to a configurational understanding of business performance will show that the individual components behave differently under different conditions.

Chiang (2017) adds that the business impact of the data science is different when we examine company front-office (Customer) or back-office processes (Provider). On one side, by increasing the data capture at multiple points of the Customer management, analytics find customers' knowledge, improve the analysis of the customer journey, support human decision making with automated algorithms and finally allow to deliver a personalized and differentiated customer experience (Chiang, 2017; Fosso Wamba et al., 2015; Waller & Fawcett, 2013) . On the other side, data science leverages a direct communication with providers or distributors and permits real-time management of the supply chain orienting the business value in that case to the efficiency of the operations (Addo-Tenkorang & Helo, 2016). In the transport industry for instance, recent studies show that data science provides a better management of the container flows between ports (Tsai & Huang, 2017) or a decrease of the bullwhip effect between providers of a supply chain (Hofmann, 2017) or detects problem root cause before an incident occurs and stops the production (Chien, Liu, & Chuang, 2017). Our empirical study contributes to exploring the different combinations of causal conditions that may be linked to the improvement of the processes related to the Customer (front-office) and Provider (back-office) management.

2.2 Data Maturity Model

Davenport & Prusak (2000) and Otto & Hüner (2009) analyse the organizational barriers that fail to improve firm performance. They show a main impediment is the incapacity to handle a huge amount of unstructured and structured data that the applications and connected sensors collect. Wegener (2008) adds that data needs to comply with a minimum of quality to provide value. He illustrates it through examples coming from different industries. Firms operating in pharmaceutical and transport need accurate and opportune data to comply with strict regulations on the traceability of events and accountability. Companies of the consumer product sector need real-time and accurate data to improve efficiency and agility of the business ecosystem actors that participate in their production lines. Even if their main business remains B2B (Nike, Adidas, Procter & Gamble, L'Oréal…), they modify their commercial strategy to directly understand the needs of their final consumers through their web and social media applications. They get in touch with their final users, but, immerged in an ocean of data, they struggle to get the correct information, i.e. the trends that characterize their consumer markets. Ecommerce, finance and trading companies who are per se data-intense business companies develop also large efforts to achieve an accurate data set able to make them develop new strategy and launch new products.

According to Becker, Knackstedt, & Pöppelbuß (2009) a maturity data model (MDM) is an artefact that aims at solving the problem of data management of an organization providing with a status and identifying actions for improvements. Spruit & Pietzka (2014) after a thorough analysis of the literature of existing MDM models choose 4 areas to cover all aspects of data management: data consistency, data completeness, data usage and data protection. The condition of data usage consists on defining who uses the data in which systems, which employee has read/write access and if it is clear why

people are granted or denied access to certain data, and if the organization can find out if there are data ownership concepts implemented and see whether the historical grown way still displays the needs. It is generally agreed throughout academia and practitioners that divided responsibilities shared with different people are not effective (EDUCAUSE, 2009; Loshin, 2010).

Furthermore, due to data privacy and data protection reasons, data have to be distributed to appropriate users and not be made available for users without access rights and data availability of course must be ensured at all times (Anderson & Moore, 1990). The condition of data protection is a secured data against possible incidents. Incidents can be of different kinds; either failure of components, software bugs or steered by people on purpose, like sabotage, hacking, fraud or theft. To ensure a good running of the business, Shaw, Chen, Harris, & Huang (2009) recommend to conduct it via physical measures and software precautions. This present study aims at exploring the different combinations of these 4 causal variables to the efficiency of the back-office and front-office company processes.

2.3 Data-driven profile

Echoing the works of Davenport & Prusak (2000), Bonabeau (2003) asserts that very few senior executives take currently their decisions based on data. Woerner & Wixom (2015) add that this practice remains still marginal to the success of the business. They raise the question why if most large organizations decided to implement analytical applications and business intelligence software few years ago (McAfee & Brynjolfsson, 2012), they still don't use them now to take decisions and prefer instead to continue using their gut feel. Davenport & Harris (2007) name analytical competitor: "an organization that uses analytics extensively and systematically to outthink and outexecute the competition". For our paper, we define the variable 'data-driven' to

describe the organizations which prefer to leverage data to take decisions instead of using instinct or professional experience.

3. Methodology

Qualitative Comparative Analysis (QCA) is a research methodology to conceptualizing and analysing causality that definitively differs from statistics based on linear algebra such as structural equation modelling (Ren, Wamba, Akter, Dubey & Childe, 2017; Kwon, Lee, & Shin, 2014). The latter seeks to estimate the separate contribution of each cause (independent variable) in explaining variation in the outcome (dependent variable) in order to determine a possible correlation between them. QCA is a research methodology for small sampling of results (e.g. between 10 and 50 cases) (P. C. Fiss, 2007) that incorporates Boolean logic for a comparison of principles. QCA focuses on identifying the relationships of necessity and sufficiency between the causes and the outcome (Ragin, 2008). QCA application follows two approaches. The first one is the crisp-set qualitative comparative analysis (csQCA), which is suitable for variables with binary values (0 o 1), where a value 1 indicates the presence of a condition and 0 its absence or negation. The second one is the fuzzy-set qualitative comparative analysis (fsQCA) which is used to analyse variables with continuous values. Our study used FsQCA through the computer software R and the package QCA (Dusa, 2019).

The advantages of QCA in comparison with correlational techniques are double: (a) equifinality, which means that different paths can lead to the same outcome; (b) asymmetry, meaning the presence and the absence of the outcome, respectively may lead to different explanations. The first step of the method is called the calibration of the values in order to determine the different thresholds within the values. According to Fiss (2011), the calibration of the variables reduces the sample dependence, because

membership to a set depends on the knowledge instead of the arithmetic mean, which reduces representativeness.

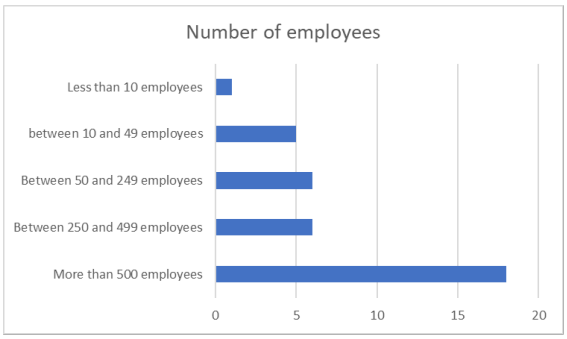
After the calibration, the second step is the analysis of the data which follows three steps: (1) creating the truth table, which comprises all the possible combinations of conditions (the table has 2 k rows, where k is the number of causal conditions used in our study); (2) reducing the table according to the minimum number of cases required to obtain the outcome (also called frequency) and the level of minimum consistency (also called inclusion) that, according to Ragin (2008) is 0.75; and (3) transforming (through a Boolean algorithm) the truth table into the combinations of variables that produce the outcome.

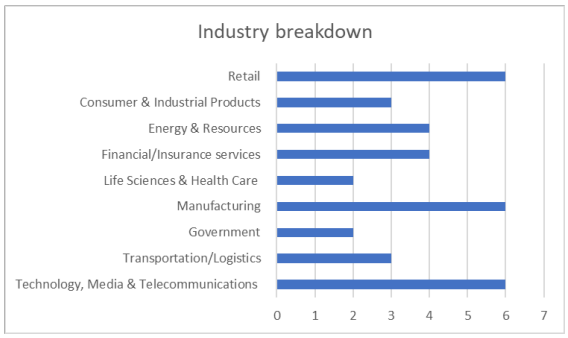
3.1 Data collection and description of the conditions

We address the 9 major industries in Spain: Retail, Consumer & Industrial Products, Energy & Resources, Financial/Insurance, Life Science & Health care, Manufacturing, Government, Transportation/Logistics and finally Technology, Media and Telecommunications. This questionnaire is addressed to Chief Data Scientist, Chief Technology Officer or Chief Information Officer with company headquarter based in Spain. It was also important that this professional has a transversal view across his organization, usually reporting to the CEO and sitting in the Executive Board. The Camerdata institution[1] yearly issues a report of the Spanish economic activities including the 1st level organizational structure. We extracted from the 2017 report an initial list of 153 Senior Executives who received the questionnaire. We got 47 responses for a 30.7% response rate. After the removal of missing data and outliers, a sampling of 37 responses keeps for QCA analysis (see figure 1).

Figure 1:

[1] https://www.camerdata.es





The answers represent the 9 major industries. In terms of size, half of the sampling are companies with more than 500 employees (see figure 1). In summary, we observe the composition of the sampling represents the Spanish market engaged in analytics projects[2].

We use Likert-type scales to measure latent constructs which are thought of as unobservable characteristics, feelings, opinions such as data-driven behaviour or brand image improvement. A 5-point response scale fits with the nature of the statements presented in the survey. The reliability of the Likert-type scales is showed using Cronbach's Alpha with values superior to 0.7 for good internal consistency. The Appendix lists the 31 measurement items of the questionnaire where 1 = "completely disagree" and 5 = "completely agree". The survey is based on a google form which includes three sections. First, the respondents position themselves according to different statements aimed at measuring the maturity of the data across 4 different axes: data consistency, data completeness, data usage and data protection. Then, they indicate in which extent the data of the company provide them benefits in the management of their Customers and Providers. Then, the survey raises questions to evaluate the data-driven propensity/behaviour of the company (to take a decision based on data instead of instinct/experience). Finally, it concludes with questions about demographic information, the organization's age, and the number of employees.

[2] https://www.camerdata.es

For each statement, we check the reliability of Likert-type scales using Cronbach's Alpha (internal consistency). All the values are superior to 0.7 which is synonym of good internal consistency.

3.2 Feature selection technique

Feature selection methods aim to create a more accurate predictive model. Our conditions and outcomes are latent constructs that we measure through 31 survey items/inputs which have each of them 5 points Likert scale. To find the most influential combinations of inputs, we conduct a factor analysis to reduce the number of variables. Fewer attributes are desirable because they reduce the complexity of the model, and a simpler model is easier to understand and explain. Our objective is to transform the interpretation of a 31-question survey to the study of seven factors i.e. five conditions and two outcomes. In our case, we apply together filter and wrapper methods (Guyon, 2003).

For each factor, we calculate the internal consistency reliability (ICR) as measured by Cronbach alpha, the composite reliability (CR) and the convergent validity as measured by the Average Variance Extracted (AVE) (see table 1). ICR is a measure based on the average inter-correlation between different items while convergent validity refers to the degree to which a measure is correlated with other measures that it is theoretically predicted to correlate with. To calculate the AVE of the latent constructs, we take the loadings of the different items on the construct and calculate the average of squared loadings. We calculate the composite reliability (CR) which may differ from the Cronbach alpha's if the factor loadings of the items are not the same.

By evaluating the scores, we determine which items are to be kept or removed from the dataset to satisfy the conventional thresholds of 0.5 for AVE, 0.7 for either Cronbach alpha or composite reliability (Peterson, 1994).

Our results help us to identify and remove the inconsistent item M10 from the dataset which decreased the accuracy of the data completeness construct in our model. However, we validate the possible combinations of variables as described in the table 1.

Table 1: Feature Selection technique

Condition and outcomes	Abbreviation	Items combinations	Description	Factor analysis
Data-driven profile	Dat-drv	M1 to M3	Preference to use data to take decision	ICR = .780 CR = .855 AVE = .542
Data consistency	Dat_con	From M4 to M7	Expressing the data in the same way	ICR = .840 CR = .892 AVE = .674
Data completeness	Dat_cmp	M8 and M9. M10 excluded	Expressing no data are missing	ICR = .77 CR = .844 AVE = .73
Data usage	Dat_usg	M11 and M12	Access of data is defined	ICR = .760 CR = .893 AVE = .806
Data protection	Dat_pro	M13 only	Access of data is managed and controlled	Not applicable
Front-office/Customer management (OUTCOME)	Custom	A1 to A4	Improvement of the Customer management efficiency	ICR = .830 CR = .895 AVE = .680
Back-office/Provider management (OUTCOME)	Prom	A5 to A10	Increase of Provider operations efficiency	ICR = 0,800 CR = .883 AVE = .716

It is worthy to notice that we consider and evaluate if the company size affects the results. After including in the model, we decided to remove it because we found that it doesn't influence the two outcomes. Our objective was to avoid including noise in our solution.

3.3 Factors calibration

The calibration is established according to three different thresholds: full membership, full non-membership and cross-over. The full membership corresponds to a response equivalent to 5 = "completely agree" for all the items aggregating the variable. A full non-membership corresponds to a response equivalent to 1 = "completely disagree" or 2 = "disagree" for all the items aggregating the variable. The cross-over is given considering all the answers equal to the choice 3 = "neither agree nor disagree". The result for cross-over is 0.5. For the rest of the answers, we decided to apply a linear function between the cross-over and the full membership.

The resulting equation is:

$$\text{Calibration} = (0{,}5 / 2n) * \sum_{1 \leq k \leq n} sum(Answers(k)) - 0.25$$

Where n is the number of questions that aggregate the variable and,
Answer (k) the result given by the respondent based on Likert-scale of the statement k

Table 2 shows the threshold values used for the calibration of the conditions and the outcomes.

Table 2: Calibration of the variables

Variables	Full membership	Cross over	Full Non-membership
	1	0.5	0
Data consistency	20	12	4
Data completeness	10	6	2
Data usage	10	6	2
Data protection	5	3	1
Data-driven profile	25	15	5
Front-office/Customer management (OUTCOME)	20	12	4
Back-office/Provider management (OUTCOME)	15	9	3

4. Results

We analyse the conditions that lead to an improvement or a deterioration of the Customer and Provider management processes. Then, we present an analysis of necessity and sufficiency conditions.

4.1 Analysis of necessary conditions

Table 3 examines the relationship between the five conditions and the two outcomes: Customer and Provider Management. Our study analyses both the presence of the condition and its absence (asymmetry). The analysis in Table 3 shows that none of the conditions is either necessary for improving the Customer management and Provider operations (consistency lower than 0.9) or for decreasing them. Thus, the increase or decrease of these outcomes leads therefore to a combination of different conditions.

Table 3 - Analysis of necessary conditions

Conditions tested	Customer Mngt		~ Customer Mngt		Provider Mngt		~ Provider Mngt	
	Cons.	**Cov.**	**Cons.**	**Cov.**	**Cons.**	**Cov.**	**Cons.**	**Cov.**
Data consistency	0.480	0.898	0.343	0.413	0.471	0.798	0.426	0.588
~Data consistency	0.686	0.619	0.815	0.531	0.757	0.619	0.854	0.568
Data completeness	0.477	0.823	0.526	0.583	0.462	0.723	0.549	0.698
~ Data completeness	0.758	0.713	0.840	0.509	0.806	0.687	0.782	0.542
Data usage	0.704	0.844	0.662	0.511	0.725	0.788	0.653	0.577
~ Data usage	0.592	0.731	0.798	0.634	0.611	0.684	0.760	0.692
Data protection	0.825	0.784	0.686	0.420	0.853	0.736	0.699	0.490
~ Data protection	0.390	0.658	0.648	0.704	0.409	0.626	0.623	0.775
Data-driven profile	0.546	0.846	0.487	0.486	0.634	0.890	0.449	0.513
~ Data-driven profile	0.668	0.670	0.845	0.545	0.653	0.593	0.804	0.667

4.2 Analysis of sufficiency

The sufficiency analysis explains which combination of conditions is sufficient to obtain the outcome (Ragin, 2008). These solutions incorporate all the logical remainders, providing a solution that is easier to analyse. We present in the tables 4 and 5 the solutions after the Boolean minimization (Thiem & Dușa, 2013).

4.2.1 Customer Management (Custom)

The model that gives conditions to improve the Customer management presents 4 causal configurations (see Table 4.1). These 4 patterns show a consistency over 0.919, which is sufficient to produce the outcome (Fiss, 2011). First of all, even if there is no condition present in all the configurations, each condition is present at least in half of the configurations. Moreover, the condition ‘data protection’, present in 3 out of 4, is

the more relevant. Following Ragin (2008), our study analyses the 3 causal configurations with the highest unique and raw coverages, because high coverage values are synonym of greater empirical relevance. Configuration 1 (dat_con * dat_usg * dat_pro) being (*) the logical operator AND, shows the presence of data with high consistency, high usage and high protection all together lead to an improvement of the Customer management. Configuration 2 (dat_usg * dat_pro * ~ dat_drv), being (~) the logical operator NO or ABSENCE, reflects a situation that also leads to an improvement of the Customer management. Finally, configuration 3 (dat_con * dat_cmp * dat_pro * dat_drv) indicates that a high data consistency, completeness, protection and data-driven profile all together improve the Customer management.

However, the model that analyses the reduction of the Customer management shows 3 causal configurations (see Table 5.1). This model shows the diversity of existing paths leading to the outcome (~Customer management). We highlight that data consistency and data protection are present in all the configurations showing as in the previous model that data protection is a relevant condition.

4.2.2 Provider management (Prom)

The model that gives conditions to increase the provider operations efficiency presents 3 causal configurations (see Table 4.2). These three patterns show a consistency over 0.936, which is sufficient to produce the outcome (Fiss, 2011). First of all, data-driven is a condition present in all configurations. According to Ragin (2008), our study analyses the 3 causal configurations with the highest unique and raw coverages. Configuration 1 (~dat_con * ~dat_cmp * dat_drv) and configuration 2 (~dat_con * ~dat_usg * ~dat_pro * dat_drv) show that the presence of high data-driven condition leads to an increase of the Provider efficiency even if the rest of conditions are absent, ambiguous or negative. Finally, configuration 3 (dat_con * dat_cmp * dat_pro *

dat_drv) indicates that a high data consistency, completeness, protection and a data-driven profile all together improve the provider operations, it is an ideal situation, but the lowest coverage of this pattern shows its coverage limitation.

The model that analyses the reduction of the Provider operations efficiency shows four causal configurations (see Table 5.2). This model shows again the data-driven variable is a condition present in all configurations (~Prom). Moreover, we note that data consistency is also present in all the configurations showing that a lower data consistency is a relevant condition for lower provider operations efficiency.

Table 4.1 Analysis of sufficiency conditions (Custom)

Conf.	Conditions					Coverage		Consistency
	Data consistency	Data completeness	Data usage	Data protection	Data-driven	Raw	Unique	
1	●		●	●		0.423	0.033	0.927
2			●	●	○	0.502	0.138	0.916
3	●	●		●	●	0.284	0.030	1.000
4	○	●	○	○	●	0.112	0.012	0.903

Solution coverage: 0.628. Solution consistency: 0.919.
Frequency threshold = 1. Consistency threshold = 0.9

The black circles indicate the presence of antecedent conditions while the white circles show the absence or negation of antecedent conditions. The blank cells represent ambiguous conditions.

Table 4.2 Analysis of sufficiency conditions (Prom)

Conf.	Conditions					Coverage		Consistency
	Data consistency	Data completeness	Data usage	Data protection	Data-driven	Raw	Unique	
1	○	○			●	0.456	0.040	0.924
2	○		○	○	●	0.217	0.000	0.910
3	●	●		●	●	0.311	0.021	0.993

Solution coverage: 0.6583. Solution consistency: 0.936.
Frequency threshold = 1. Consistency threshold = 0.9

The black circles indicate the presence of antecedent conditions while the white circles show the absence or negation of antecedent conditions. The blank cells represent ambiguous conditions.

Table 5.1 Analysis of sufficiency conditions (~Custom)

Conf.	Conditions					Coverage		Consistency
	Data consistency	Data completeness	Data usage	Data protection	Data-driven	Raw	Unique	
1	○		○	○		0.563	0.188	0.736

2	○	○		○	●	0.296	0.010	0.764
3	○	●		○	○	0.316	0.047	0.836

Solution coverage: 0.643. Solution consistency: 0.744.
Frequency threshold = 1. Consistency threshold = 0.75

The black circles indicate the presence of antecedent conditions while the white circles show the absence or negation of antecedent conditions. The blank cells represent ambiguous conditions.

Table 5.2 Analysis of sufficiency conditions (~Prom)

Conf.	**Conditions**					**Coverage**		**Consistency**
	Data consistency	Data completeness	Data usage	Data protection	Data-driven	Raw	Unique	
1	○	○	○		○	0.544	0.025	0.814
2	○	○		○	○	0.563	0.078	0.758
3	○	●	●		○	0.285	0.015	0.845
4	○	●		○	○	0.313	0.025	0.947

Solution coverage: 0.884. Solution consistency: 0.731.
Frequency threshold = 1. Consistency threshold = 0.75

The black circles indicate the presence of antecedent conditions while the white circles show the absence or negation of antecedent conditions. The blank cells represent ambiguous conditions.

4.3 Multiple regression analysis (MRA) – Classical linear algebra method

With the objective to compare the previous results with those obtained with a more traditional solution (based on correlations), we run a classical linear algebra method i.e. a multiple regression analysis (MRA). The results are presented in the Table 6. The model explains 8% of the variance of Customer Management and 30% of Provider Management (adjusted R-squared) which are respectively very weak and weak correlation rates and low results. Contrary to net effects analyses (e.g., structural equation modelling, multiple regression, analyses of variance) that examine direct effects of individual independent variables on outcome variable (dependent variable), our method fsQCA identifies combinations of causal conditions that lead to an outcome of interest in the real business world. This technique stresses that combinations of conditions result in an outcome, rather than individual variables. Moreover, fsQCA overcomes the limitation of net effects analyses, which assume symmetrical relationships between variables. Indeed, fsQCA can identify different configurations of conditions that predict both the presence and the absence of an outcome.

Table 6 – Multiple regression results

	Custom			Prom		
Model	Estimate	Std Error	Pr(>|t|)	Estimate	Std Error	Pr(>|t|)
Data consistency	0.336	0.194	0.094	0.069	0.165	0.67806
Data completeness	-0.241	0.196	0.230	-0.353	0.167	0.04403*
Data usage	-0.022	0.196	0.913	0.052	0.167	0.76016
Data protection	0.113	0.176	0.525	0.184	0.149	0.22982
Data driven	0.117	0.152	0.449	0.404	0.129	0.00423**
Ajusted R-squared		0.079			0.2993	
F		1.57			3.82**	

Signif. codes: ‘ ‘ <1, *p < .05, **p < .01

5 Discussion

Our paper aims at understanding the configurations associated to better performance of the company and we decide to analyse the processes related to Customer management and Provider management (operations efficiency). We find that an analytical competitor who is oriented to take decision based on the use of data (Davenport & Harris, 2007) yields the improvement of the supplier operations efficiency in three main configurations that have the greater empirical relevance and summed up, reached 65% of the solution. It shows that the data play a more important role. Indeed, we observe that company decision makers with higher providers efficiency results take actions for the control and the management of their providers based on the data they have, and our study highlights it is still true even if a room exists regarding the maturity of the data they use.

However, this data-driven behaviour is not always applied as far as it concerns the management of the Customer processes. In that case, company decision makers with best customer management need an additional condition that their data satisfy at least three conditions (out of 4 based on our model). Data usage and Data protect are the two more frequent conditions amongst the three solutions analysed. Looking ahead, the

maturity of the data entails the next question about its source: external or internal, local, regional or international (Becker et al., 2016). We suggest as further lines of research to include these variables as new conditions of the study and analyse the contribution factor on the firm performance.

We propose also that future studies investigate if the improvement of the Customer and Provider management processes have a direct impact not only on the back and front-office processes but also on the financial and non-financial Key Performance Indicators (KPIs) of the company such as the profitability, the market response, the market position value, and the new product success rate. These results could be collected extending the current survey prepared for the companies. Indeed, the recent works of Becherer, Helms, & McDonald (2012) corroborate that the measures of business performance dimensions may rely on respondents' subjective assessments and validate this approach. It will be also interesting in this extended survey to ask for the future initiatives that companies wish to start and analyse if those who receive more benefits (better Customer management or Provider operations efficiency) plan to acquire more business analytics capabilities.

6 Conclusion

The conclusion of the study is different combinations of casual conditions (mainly related to 3 items of the data maturity along with the data-driven company profile) drive to a better Customer management and Provider operations efficiency (outcomes). The results show one solution does not fill all and further studies should understand better the equifinality of the outcomes. This study has some limitations. Firstly, it measures conditions according to the answers given by the respondents of technical departments and with a technological profile (CTO and CIO). The important length of our questionnaire with 31 statements and the C-level position of the respondent make us

anticipated that the population of cases would be too low for statistical standard techniques and FsQCA would be best applied.

But, if we increase the size of the sampling (medium or large N) by including responses from representative of sales, marketing, production and procurement departments, we believe, based on the works realized by (Fiss, 2007; Piñeiro-Chousa, López-Cabarcos & Pérez-Pico, 2016; Ragin, 1992; Woodside & Zhang, 2012) that FsQCA can be applied whenever complex causality is present. Moreover, the number of variables we can include in a QCA analysis does not depend on the number of cases, but the higher the diversity index (the ratio of the number of observed configurations to that of all logically possible configurations), and the more cases per configuration relative to our target population, the more credible our results. For example, Berg-Schlosser & Meur (2009) mention in 'an intermediate-N analysis (10 to 40 cases) would be to select from 4 to 6-7 variables. Consequently, we find appropriate to extend the number of variables or cases as a future line of research if the current diversity index does not decrease (7/37).

Secondly, we analyse the management of Customer or Provider processes without distinguishing the type of channels used (offline, online, mixed). Understanding the relationship between data and firm performance requires more specific questions to seize if the analytics improve the physical or digital marketing campaigns or alternatively, if the decrease in supply chain costs are related to a physical decrease of the inventory or a digital integration of the systems between provider and customer.

Acknowledgements:

The authors thank Dr. Marc Torrent, Director at Big Data Center of Excellence Barcelona and Circe Serra, Senior project manager at Eurecat for their support in providing data, information and contacts throughout this project. Eurecat is the most

important technology centre in Catalonia, offering innovation solutions to boost competitiveness of industrial and services Catalan companies.

Appendix

Construct measurements. The respondent has asked to evaluate in a Likert-scale between 1-5 each of the following statements:

First section – Data Maturity adopted from (Kwon et al., 2014)

Block	items	Statements	Strongly disagree	Disagree	Neither agree nor disagree	Agree	Strongly agree
Data-driven behaviour	M1	Our company takes into account the data analytics skills in the hiring process					
	M2	A relevant number of our employees have a science or technology academic background					
	M3	Our company delivers training classes related to data analytics & visualization					
Data consistency	M4	A common definition of the data sources (templates, order forms) allows to share data inside the organization					
	M5	All data (ordering details, material inventory, etc.) are managed in the same way throughout the organization					
	M6	An automatic method of maintaining data consistency is being used					
	M7	There is no input error in all the data (e.g. information manually entered incorrect, machine calibration outlier, etc.)					
Data completeness	M8	All sources (data) have been inputted by our company with no omissions. New variables are included if required					
	M9	All sources (data) have been inputted by our suppliers and clients with no omissions. New variables are included if required					
	M10	Problems due to incomplete data are hard to be found					
Data Usage	M11	The employees of functional departments (business users) find the source and have access the data they need					
	M12	There is a dialogue between the IT department and the functional departments that permits the perfect exploitation of the data					
Data Protection	M13	Data access (especially sensitive data) is restricted according to the rules defined by the user profile					

Second section – Outcomes

Business Processes	items	Statements	Strongly disagree	Disagree	Neither agree nor disagree	Agree	Strongly agree
1. Customer management	A1	Our data analytics allow our organization to improve its marketing campaigns					
	A2	Our data analytics allow to improve our brand image					
	A3	Our data analytics allow the organisation to have a 360º vision of our client and develop specific actions by customer segment (Coupons, loyalty program, special offer...)					
	A4	Our data analytics allow to make the best offer for the client in response to market conditions or competition behaviour					
2. Supplier and distributors management and internal processes		Our data analytics allow the organization to					
	A6	a. Cooperate better with suppliers/distributors, predicting demand					
	A7	b. Improve the traceability of the supply chain, manufacturing and logistics operations					
	A8	c. Decrease operation and management costs for material procurement and sales/distribution (for instance: inventory reduction)					
	A9	d. Prevent from Fraud detection and Cyber intelligence					
	A10	e. Prevent from financial risks (increase of materials price, monetary risks, etc.)					
	A11	e. Analyse the risks related to compliance, ethics and corporate responsibility					

Third section: Demographic information

C0	Company name:				
C1	Position of the respondent in the company				
C2	Email of the person responding to this questionnaire				
C3	Where are the headquarters of the company?				
C4	In how many countries does the company have operations	1. one country 2. between 2 and 5 3. more than 5			
C5	In which sector does the company operates?				
C8	What is the size of the company?	1. Fewer than 10 employees 2. 10 to 49 employees 3. 50 to 249 employees 4. 250 to 499 employees 5. More than 500 employees			
C9	Age of the company (#years)	below 3 years	3-5 years	5-10 years	more than 10 years

Bibliography

Addo-Tenkorang, R., & Helo, P. T. (2016). Big data applications in operations/supply-chain management: A literature review. *Computers and Industrial Engineering*, *101*, 528–543. https://doi.org/10.1016/j.cie.2016.09.023

Anderson, B. D. O., & Moore, J. B. (1990). *Optimal Control: Linear Quadratic Methods*. Upper Saddle River, NJ, USA: Prentice-Hall, Inc.

Barney. (1991). Firm Resources and Sustained Competitive Advantage. *Journal of Management*, *17*(1), 99–120.

Barney, J. (1986). Strategic Factor Markets: Expectations, Luck, and Business Strategy. *Management Science*, *32*(10), 1231–1241. https://doi.org/10.1287/mnsc.32.10.1231

Barney, J., Wright, M., & Ketchen, D. J. (2001). The resource-based view of the firm: Ten years after 1991. *Journal of Management*, *27*(6), 625–641. https://doi.org/10.1177/014920630102700601

Becherer, R. C., Helms, M. M., & McDonald, J. P. (2012). The Effect of Entrepreneurial Marketing on Outcome Goals in SMEs. *New England Journal of Entrepreneurship*, *13*(2), 57–70. https://doi.org/10.1108/NEJE-11-02-2008-B002

Becker, J., Knackstedt, R., & Pöppelbuß, J. (2009). Developing Maturity Models for IT Management. *Business & Information Systems Engineering*, *1*(3), 213–222. https://doi.org/10.1007/s12599-009-0044-5

Becker, T., Curry, E., Jentzsch, A., & Palmetshofer, W. (2016). *New horizons for a data-driven economy: Roadmaps and action plans for technology, businesses, policy, and society*. *New Horizons for a Data-Driven Economy: A Roadmap for Usage and Exploitation of Big Data in Europe*. https://doi.org/10.1007/978-3-319-21569-3_16

Berg-Schlosser, D., & Meur, G. (2009). Configurational Comparative Methods: Qualitative Comparative Analysis (QCA) and Related Techniques. Thousand Oaks, California: SAGE Publications, Inc. https://doi.org/10.4135/9781452226569 NV - 51

Bonabeau, E. (2003). *Don't trust your gut*. *Harvard business review* (Vol. 81). Graduate School of Business Administration, Harvard University. Retrieved from http://cat.inist.fr/?aModele=afficheN&cpsidt=14791484

Braumoeller, B. (2008). *Fuzzy-Set Social Science. Journal of Politics* (Vol. 70). https://doi.org/10.1017/S0022381607080309

Chiang, W.-Y. (2017). Discovering customer value for marketing systems: an empirical case study. *International Journal of Production Research*, *55*(17), 5157–5167. https://doi.org/10.1080/00207543.2016.1231429

Chien, C. F., Liu, C. W., & Chuang, S. C. (2017). Analysing semiconductor manufacturing big data for root cause detection of excursion for yield enhancement. *International Journal of Production Research*, *55*(17), 5095–5107. https://doi.org/10.1080/00207543.2015.1109153

Davenport, T. H., & Harris, J. G. (2007). *Competing on Analytics: The New Science of Winning* (Harvard Bu).

Davenport, T. H., & Prusak, L. (2000). *Working Knowledge: How Organizations Manage What They Know*. Boston, MA, USA: Harvard Business School Press.

Dusa, A. (2019). *QCA with R. A Comprehensive Resource*. New York: Springer International

Publishing.

Educause Center For Applied research. (2009). *Educause Center For Applied research*. *Data stewardship, security and policies. ECAR Research Study 8*.

Evans, P. (2013). How data will transform business. available on-line at http://www. ted. com/talks/philip_evans_how_data_will_transform_business.

Fiss, P. (2011). *Building Better Causal Theories: A Fuzzy Set Approach to Typologies in Organization Research*. *Academy of Management Journal* (Vol. 54). https://doi.org/10.5465/AMJ.2011.60263120

Fiss, P. C. (2007). A SET-THEORETIC APPROACH TO ORGANIZATIONAL CONFIGURATIONS. *Academy of Management Review*, *32*(4), 1180–1198. https://doi.org/10.3109/10253890.2015.1121984

Forbes. (2014). Making Analytics accountable: 56% of Executives expect analytics to contribute to 10% or more growth in 2014. Retrieved from https://www.forbes.com/sites/louiscolumbus/2014/12/10/making-analytics-accountable-56-of-executives-expect-analytics-to-contribute-to-10-or-more-growth-in-2014/

Fosso Wamba, S., Akter, S., Edwards, A., Chopin, G., & Gnanzou, D. (2015). How "big data" can make big impact: Findings from a systematic review and a longitudinal case study. *International Journal of Production Economics*, *165*, 234–246. https://doi.org/10.1016/j.ijpe.2014.12.031

Grimaldi, D., Fernandez, V., & Carrasco, C. (2018). Heuristic for the localization of new shops based on business and social criteria. *Technological Forecasting and Social Change*, (September 2017), 1–9. https://doi.org/10.1016/j.techfore.2018.07.034

Guyon, I. (2003). AnIntroductionToVariableAndFeautureSelection.pdf, *3*, 1157–1182. https://doi.org/10.1016/j.aca.2011.07.027

Harris, J. G., & Craig, E. (2011). Developing analytical leadership. *Strategic HR Review*, *11*(1), 25–30. https://doi.org/10.1108/14754391211186287

Hofmann, E. (2017). Big data and supply chain decisions: the impact of volume, variety and velocity properties on the bullwhip effect. *International Journal of Production Research*, *55*(17), 5108–5126. https://doi.org/10.1080/00207543.2015.1061222

Igarashi, Y., Nagata, K., Kuwatani, T., Omori, T., Nakanishi-Ohno, Y., & Okada, M. (2016). Three levels of data-driven science. *Journal of Physics: Conference Series*, *699*(1). https://doi.org/10.1088/1742-6596/699/1/012001

Ji-fan Ren, S., Fosso Wamba, S., Akter, S., Dubey, R., & Childe, S. J. (2017). Modelling quality dynamics, business value and firm performance in a big data analytics environment. *International Journal of Production Research*, *55*(17), 5011–5026. https://doi.org/10.1080/00207543.2016.1154209

Kwon, O., Lee, N., & Shin, B. (2014). Data quality management, data usage experience and acquisition intention of big data analytics. *International Journal of Information Management*, *34*(3), 387–394. https://doi.org/10.1016/j.ijinfomgt.2014.02.002

Lavalle, S., Lesser, E., Shockley, R., Hopkins, M. S., & Kruschwitz, N. (2011). Big Data, Analytics and the Path From Insights to Value. *MIT Sloan Management Review*, *52*(2), 21–32. https://doi.org/10.0000/PMID57750728

Lim, J.-H., Stratopoulos, T. C., & Wirjanto, T. S. (2011). Path Dependence of Dynamic Information Technology Capability: An Empirical Investigation. *Journal of Management Information Systems*, *28*(3), 45–84. https://doi.org/10.2753/MIS0742-1222280302

Loshin, D. (2010). *MDM components and the maturity model. A Dataflux white paper*.

McAfee, A., & Brynjolfsson, E. (2012). Big Data's Management revolution. *Harward Business Review*.

Meyer, J. P., Allen, N. J., & Smith, C. A. (1993). Commitment to organizations and occupations: Extension and test of a three-component conceptualization. *Journal of Applied Psychology*, *78*(4), 538–551. https://doi.org/10.1037/0021-9010.78.4.538

Miller, D. (1986). Configurations of strategy and structure: Towards a synthesis. *Strategic Management Journal*, *7*, 233–249.

Newman, H. H. (1978). Strategic Groups and the Structure-Performance Relationship. *The Review of Economics and Statistics*, *60*(3), 417–427. Retrieved from https://econpapers.repec.org/RePEc:tpr:restat:v:60:y:1978:i:3:p:417-27

Otto, B., & Hüner, K. (2009). Funktionsarchitektur für unternehmensweites Stammdatenmanagement.

Peterson, R. A. (1994). A Meta-Analysis of Cronbach's Coefficient Alpha. *Journal of Consumer Research*, *21*(2), 381–391. Retrieved from http://www.jstor.org/stable/2489828

Piñeiro-Chousa, J. R., López-Cabarcos, M. Á., & Pérez-Pico, A. M. (2016). Examining the influence of stock market variables on microblogging sentiment. *Journal of Business Research*, *69*(6), 2087–2092. https://doi.org/10.1016/j.jbusres.2015.12.013

Porter, M. . (1991). Toward a dynamic theory of strategy. *Strategic Management Journal*, *12*(Winter Special Issue), 95–117.

Purdy, D. (2012). Sparse Models for Sparse Data: Methods.

Ragin, C. C. (1992). The Comparative Method: Moving Beyond Qualitative and Quantitative Strategies. By Charles C. Ragin (Berkeley, Los Angeles, and London: University of California Press, 1987. Paperback printing, 1989. xvii plus 185 pp.). *Journal of Social History*, *25*(3), 627–628. https://doi.org/10.1353/jsh/25.3.627

Ragin, C. C., & Strand, S. I. (2008). Using Qualitative Comparative Analysis to Study Causal Order: Comment on Caren and Panofsky (2005). *Sociological Methods & Research*, *36*(4), 431–441. https://doi.org/10.1177/0049124107313903

Reed, S. (2011). Is There an Astronomer in the House? *Science*, *331*(6018), 696–697. https://doi.org/10.1126/science.331.6018.696

Schumpeter, J. A. (1934). *The theory of economic development*. *Harvard University Press*. Cambridge, MA.

Shaw, R. S., Chen, C. C., Harris, A. L., & Huang, H.-J. (2009). The Impact of Information Richness on Information Security Awareness Training Effectiveness. *Computers & Education*, *52*(1), 92–100. Retrieved from https://www.learntechlib.org/p/67114

Spruit, M., & Pietzka, K. (2014). MD3M: The master data management maturity model. *Computers in Human Behavior*. https://doi.org/10.1016/j.chb.2014.09.030

Thiem, A., & Duşa, A. (2013). *Qualitative Comparative Analysis with R: A User's Guide* (Vol. 5). https://doi.org/10.1007/978-1-4614-4584-5

Tsai, F. M., & Huang, L. J. W. (2017). Using artificial neural networks to predict container flows between the major ports of Asia. *International Journal of Production Research*, *55*(17), 5001–5010. https://doi.org/10.1080/00207543.2015.1112046

Waller, M. A., & Fawcett, S. E. (2013). Data Science, Predictive Analytics, and Big Data: A

Revolution That Will Transform Supply Chain Design and Management. *Journal of Business Logistics*, *34*(2), 77–84. https://doi.org/10.1111/jbl.12010

Wang, F., Zhou, J., & Hu, J. (2014). DensityTransfer: A data driven approach for imputing electronic health records. *Proceedings - International Conference on Pattern Recognition*, 2763–2768. https://doi.org/10.1109/ICPR.2014.476

Wegener, H. (2008). Metadaten, Referenzdaten, Stammdaten. In *Integrierte Informationslogistik* (pp. 189–209). Springer.

Woerner, S. L., & Wixom, B. H. (2015). Big data: Extending the business strategy toolbox. *Journal of Information Technology*, *30*(1), 60–62. https://doi.org/10.1057/jit.2014.31

Woodside, A. G., & Zhang, M. (2012). Identifying X-Consumers Using Causal Recipes: ``Whales'' and ``Jumbo Shrimps'' Casino Gamblers. *Journal of Gambling Studies*, *28*(1), 13–26. https://doi.org/10.1007/s10899-011-9241-5