Radboud Universiteit

MASTER THESIS: SOLVING MESSY INFORMATION PROBLEMS

BY A COMBINATION OF BIG DATA ANALYTICS AND SYSTEM DYNAMICS

20 Augustus 2018

Ruud Jilesen (s4706188)
First Supervisor: Prof. dr. A.M.A. van Deemen Second Supervisor: Dr. H. Korzilius

Master Thesis: Solving Messy Information Problems

By a combination of Big Data Analytics and System Dynamics

Author:

Name:	Ruud Jilesen
Program:	MSC Business Administration – Business Analysis and Modelling
Institute	Radboud University, School of Management
	Nijmegen, The Netherlands
Student number:	S4706188

Graduation Committee

First Supervisor:	Dhr. prof. dr. A.M.A. van Deemen
Second Supervisor:	Dhr. Dr. H. Korzilius

ACKNOWLEDGEMENTS

I would first like to thank my first thesis supervisor prof. dr. van Deemen of the School of Management at Radboud University. It was a process of ups and down and he helped me trough this process. When I ran into trouble or just wanted a confirmation if I was on the right track, his door was open.

I would also like to thank the experts, who were involve in the validation interviews for the present study. Without their cooperation and passion, the validation could not have been successfully conducted.

I would also like to thank dr. H. Korzilius as second reader of this thesis and for his flexibility at the end.

Finally, I would like to thank my friends and family for their support and encouragement throughout my Masters and this thesis. I probably would not have finished it without them.

Ruud Jilesen

ABSTRACT

The present study gives an answer on the question: Which possibilities exist for solving messy information problems when using a combination of system dynamics and big data analytics? Messy information problems exists of five main characteristics: ambiguity, incompleteness, biases and faults, building the dynamic complexity structure and understand the dynamic complexity structure. The first archetype is fixes that fails and consists of two balancing loops and two reinforcing loops. The two balancing loops in this structure showed the processes of data mining and deep learning. The two reinforcing loops represent the application of expert and theoretical knowledge. The second archetype is success to the successful and consists of two reinforcing loops. One loop about parameter specification using model structure characteristics and one loop about parameter specification using. However, these loops are limited by the effects of the consequences the project have on participants who work with expert and theoretical knowledge. Another important limitation is not all data will be suitable for the combinations. Furthermore, a good examination of the big data analytics results, can prevent rejecting them to early.

Index

1. INTRODUCTION

The last decades the world is globalizing and becomes increasingly more a village. A driver for this globalization is information and communication technology (ICT) (Sterman, 2000). ICT connects people and machines with each other, which causes an interconnected world. In addition, ICT causes a lot of dataflows through the network which is being stored in the applications as data warehouses and in the cloud. This results in a complex world, with a lot of data flowing through all the electronic channels (Sterman, 2000).

The last ten years the amount of digital data increased tremendously, because of social media and the internet of things (Lee, 2017). The amount of data has now out-speed most or maybe even all existing computer infrastructures (Sivarajah, Kamal, Irani & Weerakkody, 2016). That is the reason why for increasingly more problems a lot of data is available for solving problems. However, the growing amount of data has the limitation that is difficult to control and check. This leads to questions about validity and reliability. All digital data including written and numerical data, are just a small part of all available data (Forrester, 1992). A result is that for some problems not all information is available to solve them. According to Forrester (1992) and Vennix (1996) more information can be found in people's head, also called mental data. Furthermore, this limited information results in the fact that no optimal solution can be found (Vennix, 1996).

Besides this information perspective, another point of view of ICT is the communication perspective, so the connections they establish. ICT increases the speed of connections between persons and machines, which make that organizations are under a reading glass. This results in the fact that organizations, according to Bryson (2004), involve stakeholders in their strategic problem definitions and in solving the strategic problems with decisions that fulfill the needs of the stakeholders. Nowadays stakeholders have the possibility to have impact on organizations and people that determines their survival. The involvement of stakeholders generates multiple perspectives on problems and their solutions (Vennix, 1996). These multiple perspectives each have access to different sources of data. Interpretation of this data is based on their frame of reference, which cause different insights into problem definitions and solutions (Vennix, 1996). Different representations of reality are caused by the limited storage and processing capacity of humans, people construct a reality instead of giving the true reality (Vennix, 1996). This construction of reality causes biases. These connections between people and machines and biased representations of reality are almost impossible to oversee as a human. For example, if something is changed in department A, this will influence department B, which results again in an effect on department A. This phenomenon is called dynamic complexity (Sterman, 2000), which means a system reacts because of feedback-loops in an anticipated, unanticipated and non-linear way (Sterman, 2000). These characteristics of limited information and dynamic complexity characterize messy problems (Vennix, 1996). The present research will focus on big limited data sets, this led to the decision to label messy problems, in this study as messy information problems to make a difference between normal messy problems and this special type.

There is published a lot about solving messy problems. The most common way for solving messy problems in general is system dynamics (Vennix, 1996; Sterman, 2000) and a special form of system dynamics is building system dynamic models in groups, also known as Group model building (Vennix, 1996). All the problems that are described and solved with system dynamics in previous articles contain small sets of data. However, other ways of a more natural way of solving can be found, for example the framing contests (Kaplan, 2008). Framing contest are the process of realigning frames to establish a common frame (Kaplan, 2008). An example of solving messy problems with a framing contest can be found in the paper of Kaplan (2008). In the paper of Kaplan (2008) the Advanced Technologies Group (ATG) of CommCorp was designing a new strategy, because the current market was not profitable anymore. In their designing process they let some groups of employees participate. In addition, the company had many sources of data in and out of the company that they could use during the designing process. The result was a conflict including multiple problem definitions. Furthermore, the information available was limited, which was the cause of the different frames (Kaplan, 2008). In addition, each frame influenced the other groups and parts of the organization and their environment. Natural ways of solving messy problems are mostly inefficient (Kaplan, 2008; Sterman, 2000). A practical perspective is needed to gain insight in efficient ways of solving messy information problems.

Over 150 papers are written about big data sets, in journals like the Journal of Big Data Research. The researchers in these studies used big data analytics. These types of analytics do not really focus on dynamic complexity but are mainly based on econometrics. An example can be found in the paper of Callado, Kelner, Sadok, Kamienski and Fernandes (2010). In the paper off Callado et al (2010) the central topic is network traffic identification. In the years before the study, many studies focused on discovering the best optimization algorithms for network traffic identification. After a comparison between the different studies, the conclusion was that no algorithm excelled. A reason was that comparing the process and outcomes was difficult. Although two insights were found. The first insight is that a combination of algorithms gave better results. Second, bidirectional algorithms, that look like feedback loops, gave significantly better results. According to Callado et al (2010) more research is demanded to investigate other methodological combinations. In addition, Ulrich (2003) suggests that a combination of different methods could improve problem solving from a system perspective, which is the basis for solving messy problems. Especially there is suggested that combinations of soft system methodologies, like system dynamics, and hard system methodologies, like big data analytics will help solving messy problems. However, no previous research can be found on combining system dynamics and big data analytics to solve messy problems. This knowledge gap of using a combination of system dynamics with big data analytics to solve messy problems leads to the following research objective. The objective of this research is to gain insights into the possibilities of solving messy information problems by using a combination of system dynamics and big data analytics. The present study includes a literature review and builds a causal loop diagram about the interaction of system dynamics and big data analytics on messy information problems. Therefore, the main question of the present research is:

Which possibilities exist for solving messy information problems when using a combination of system dynamics and big data analytics?

To answer this main question, first an answer is needed on the following questions. The first three questions define the main concepts. The last question forms the basis for the conceptual model.

- *What are messy information problems?*
- *When is a messy information problem solved?*
- *What are the characteristics of system dynamics and big data analytics?*
- *How are the characteristics of messy information problems related to concepts system dynamics and big data analytics individually?*

When the main concepts are clearly defined, and a conceptual model is given. Further research by building a causal loop diagram via a literature review and validation of the model, can give an answer on the following two questions:

- *Which characteristics of system dynamics and big data analytics provide possibilities for using combinations in relation to messy information problems?*

- *Which of the characteristics found in the literature review on the combination of system dynamics and big data analytics in relation to solving messy information problems are responsible for the success of solving messy information problems according to practitioners of system dynamics and big data analytics?*

The first four questions are answered in the theoretical background, chapter 2. A research method description follows in chapter 3. The results of the literature review can be found in chapter 4. The results of the validation of the causal loop diagram can be found in chapter 5. The conclusion and discussion can be found chapter 6. Furthermore, in appendix A is an explanation of causal loop diagrams added.

2. THEORETICAL BACKGROUND

In this chapter the main concepts of the present study are explained. First an explanation of the definition of messy information problems will be given. The second paragraph gives an answer on when a messy information problem is solved. The third paragraph discusses the characteristics of system dynamics and big data analytics. The fourth paragraph gives an answer on the question how messy information problems, system dynamics and big data analytics are related to each other.

2.1 Messy information problems

The first chapter introduced messy problems and messy information problems. In previous literature messy problems have many definitions, which sometimes are - concise and sometimes extensive. For example, Inge Bleijenberg, van Engen and Rouwette (2013) define messy problems as people having different doubts about if a specific problem is actually the problem and what the problem is caused by. This definition is in line with the definition of Vennix (p13., 1996). Vennix believes that people have entirely different views on whether there is a problem and if they agree there is a problem, they have different views about what the problem exactly is. This type of definition resembles to the group model building definition. This type of system dynamics focusses on building an SD- model to reach commitment and consensus about the problem. However, there is a second group that defines a messy problem in a different way. Enserink, Koppejan and Mayer (2012) define messy more as unpredictable and irrational behavior by people and organizations and less as problems. Homer (1996) for example defines messy information as messy details in data on the problem. Campbell (2001) focusses in her definition more on the dynamics, which are complex and therefore messy. This definition is in agreement with the work of Sterman (2000). Sterman (2000) defines messy problems as problems with limited information and with dynamic complexity. Limited information refers to data that has been sampled, averaged and/or delayed. It causes questions about validity and reliability, because of biases, errors, and other imperfections. This all is caused by selection in information. Sterman (2000) does not define the size of the used data sets, but in most case studies in the journal ‘System Dynamics Review’ the written and numerical datasets are small. The dynamic complexity refers not to the details but to the complex and dysfunctional behavior. Complex and dysfunctional behavior arises from the interactions among the agents over time. These interactions are made complex by feedback between the agents and delays in the effects of their individual behavior (Sterman, 2000). This research follows the definition of Sterman (2000), who defines messy problems as problems with limited information and dynamic complexity. We agreed on the definition of

Sterman (2000) however we want to include the condition that messy problems need to have large data sets. That is the reason why the present research does not use the concept messy problems, but it uses the concept messy information problems. Messy information problems therefore are problems with limited information in large data sets, that are dynamically complex. An example of messy information problems is the climate problem. The climate problem has a lot of connections by different greenhouse gasses and other variables. These connections cause feedback and delays in a system representation. Information on the problem also is limited because not all information is available and the meaning about some data is not available.

2.2 Solving messy information problems

By having a clear definition of messy information problems, clear criteria for when a messy information problem is solved are needed. To formulate these criteria, first a better understanding is required of the characteristics of messy information problems.

2.2.1 Limited information

First, messy information problems contain limited information, also sometimes called imperfect information (Sterman, 2000). Limited information is a popular term in messy problems. Imperfect information however is known from game theory, also sometimes called incomplete information. However, these three verbs are used completely different. Incomplete information is related to partly unavailable information, that results in the use of unjustified assumptions by people to base their decisions on (Kreps & Wilson, 1981). Imperfect information is the lack of meaning of the data, which makes it sensitive for multiple interpretations (Osborne & Rubinstein, 1994). When a player in a game thinks that data has a certain meaning, this affects his behavior in playing the game. Nonetheless, this not connotes that the conclusion about the given data is correct. In conclusion, limited information exists of three different aspects. Incomplete information is the first dimension, that can be compared to the selective perception of Sterman (2000). However, incomplete information has a wider application than only mental models because messy information problems contain large databases of numerical and written data. Focusing only on selective perception could be confusing and can give shortcomings when only focusing on mental models. The ambiguity dimension can be compared to imperfect information of Osborne and Rubinstein (1994). Thus, limited information is a combination of incomplete and imperfect information. Sterman (2000) divides limited information in three dimensions, by adding a third dimension of biases. The dimension of biases focusses on the interpretation process of data, which is not explicitly mentioned in the concepts of incomplete and imperfect information. Incomplete and imperfect information focusses only on the criteria of the data, but not the effects during interpretation. That is why biases and faults are added as the third dimension of limited information.

Each creation of reality for problem solving can be based on three types of information: mental data, written data, and numerical data (Forrester, 1992). Mental data is formed by selection and storage by people and this information is stored in people's head. Because a person has a limited space to store data, he selects information based on a certain filter and stores this filtered information. This results in limited information (Sterman, 2000). This type of information is the richest form of information. However, with the rise of ICT, increasingly more data is written and numerical. This type of data also contains lots of information. However, it is again limited information, because it is even further filtered than mental data (Forrester, 1992).

With this limited storage, processing capacity and filtered information people, get certain believes how things work and will work. However, these believes, and following actions or decisions are based on incomplete information. This results in biases, which can be defined as systematic faults. These systematic faults result in wrong measurements and conclusions (Vennix, 1996). Each bias is based on a certain heuristic, which can be defined as a mental strategy (Goodwin & Wright, 2014). An example of a heuristic is reasoning based choice, which means that people construct reasons to resolve the problem and justify their choices. The biases caused by this heuristic focus on people that are sensitive for framing. The result is irrational decision making (Shafir, Simonson & Tversky, 1993). Heuristics are important in relation to messy information problems, because it explains why people find it difficult to solve messy information problems. Most heuristics relate to the recognition heuristic, which predicts that people will choose the option that they recognize (Gigerenzer, Todd & the research group, 1999). This results in biases that are based on a wrong correlation between cause effect relationships. Another important heuristic, in relation to messy information problems, is the availability heuristic, which predicts that people attach a high probability to anomalies they can easily remember from examples (Tversky & Kahneman, 1974). Another important heuristic is the representativeness heuristic, which predicts that recognized patterns appear typical, but they are random. Therefore, people interpret randomness as a pattern or correlation (Tversky & Kahneman, 1974). The last heuristic that will be explained is the anchoring and adjustment heuristic. This heuristic predicts that adjustments to an initial value are made in a wrong way, for example too big or small (Tversky & Kahneman, 1974). All these heuristics and associated biases cause a worse decision making, that have sometimes a big impact on the results of a solution.

At last, most information contains ambiguity, which effects the quality of problem understanding in a negative way. To receive a better understanding about these effects, first ambiguity of information is defined. Ambiguity can be found in different forms, like in preferences, relevance, intelligence/information and meaning (March 1987). In addition, ambiguity arises in the field of problem solving (Kaplan, 2008), such as making choices which relates to game theory (Yang, 2018). In the present research we focus on the ambiguity by using intelligence/information for problem solving. According to March (1987) ambiguity of intelligence/information is that people depending on their environment and experience define their own outcomes for a problem and solutions. In addition, they make their own calculations about the expected consequences of a solution. An example given by March (1987) is the income statement. Furthermore, according to Kaplan (2008) ambiguous information is the linchpin in strategic problems. A reason for this is the quantity and variety of data, related to many variables. As Sterman (p20., 2000) said "*Ambiguity arises because changes in the state of the system resulting from our own decisions are confounded with simultaneous change in a host of other variables. The number of variables that might affect the system vastly overwhelm the data available to rule out alternative theories and competing interpretations*.". This variety on data of variables results in multiple possible frames. The separate multiple perspectives lead to bad framing of the problem (Kaplan, 2008). These multiple perspectives generate different views about the question if there is a problem and second about what the problem is. Thus, there are no "objective" problems, but only problems that are defined by people (Vennix, 1996). Solving these types of problems is according to Vennix (1996) about consensus and commitment to a certain problem definition. Kaplan (2008) describes it as framing, and to process is called a framing contest. Each individual actor in this case has its own frame. Frames are the means by which a person sort the ambiguity of information. Then each actor will sometimes compare its frame with another actor in a subjective way. If there is a high degree of frame resonance, the individual frames merge together as a group with the same dominant frame. If there is a low degree framing practices take place to increase the frame resonance. The framing practices consist, establish or undermine legitimacy of a frame or claims-maker. Another practice is to realign the frame (bridging, amplify, extend or transform). In case of a high degree of frame. Resonance can be established and it will lead to a decision, in case of a bottleneck the decision will be deferred and maybe suspended at the end (Kaplan, 2010). In case of messy information problems this framing is a problem. The multiple perspectives cause different frames and different ways of both interpreting and solving the problems.

In conclusion, the limited information component of messy information problems is solved if important incompleteness of data can be identified and replaced with the right selection. Secondly, biases and heuristics that influence the solution negatively can be recognized and changed into no wrongly biased information and decisions. At last the ambiguity of information needs to be decreased, so the data fits the purpose. In messy information problems this process is almost impossible to be carried out by normal people, because data sets are too big to control and check. Thus, validity and reliability of the data must be secured in another way.

2.2.2 Dynamic complexity

The second characteristic of messy information problems is dynamic complexity. Complexity is used widely in academic literature; however, it mostly refers to complexity in details (Sterman, 2000). Complexity also exists in behavior (Sterman, 2000), also known as dynamic complexity. The explanation for this is feedback mechanisms and delays (Sterman, 2000), which decreases understanding of the behavior. Interconnections and interactions between different agents cause these feedback loops. An agent is comparable to a person or an object. Not all feedback loops are always clear, sometimes there are multiple agents in a chain, which are not noticed at first sight. Each agent mostly requires time take actions based on information. At this point the time horizon appears and delays occur. This is the reason why most messy problems have a longer time horizon and become dynamically complex.

To conclude solving the dynamic complexity component of messy information problems requires identifying the feedback loops in the problem structure. A second component is identifying the delays and the impact of the delays. A system dynamics system representation and analyzing this representation can give insights for a solution to the messy information problem. A system is a way of combining variables, connections and polarities of relations into structures, to construct feedback loops. In addition, delays can be included within a system dynamics system representation. Receiving insights in the system representation and analyzing system behavior is required to solve the dynamically complex component of messy information problems.

2.3 System dynamics and big data analytics as tools

This paragraph discusses the tools of system dynamics and big data analytics. The first subparagraph describes the characteristics of system dynamics. The second subparagraph describes the characteristics of big data analytics.

2.3.1 System dynamics

System dynamics is a method that has the goal to increase understanding of messy problems in general. System dynamics offers a way to show the connections between variables in a system by building a model and showing the behavior of the system by running the model (Vennix, 1996; Sterman, 2000). Furthermore, the strength of system dynamics is the way it can structure a problem, because stakeholders within the messy problem have different views on if there is a problem and what the problem is. The basis for solving a messy problem is to receive a common structure of the problem and an explanation about how it works. In the literature of group model building this is also called consensus and commitment of the problem (Vennix, 1996). The structure of the problem shows how different parts of the problem are interacting and what parts are responsible for the system behavior. These insights are valuable for further solving the problem and the effect of the solution on the problem. Furthermore, system dynamics offers the possibility for including data into the model. A weakness of system dynamics is hidden in the system model building and by including data into the model. Most relationships in the model relate to mental data, which is sometimes biased or ambiguous. The same applies for the large written and numerical data sets where the present research is focusing on. It is hard to check in these large databases if the data and extracted information are reliable and valid (Vennix, 1996; Forrester, 1992). However, if the behavior of the system is not affected too much by wrongly selected, biased or ambiguous information then still the results can be reliable (Sterman, 2000). Based on the results of the systems behavior, a scholar can decide whether the data is reliable enough. Relationships between variables based on mental data never can be fully validated, system dynamics models also cannot be fully validated. That is why many scholars not validate this type of models but build confidence in the model, by testing the model. The more tests the model passes, the more confidence. (Taylor, Ford & Ford, 2010; Sterman, 2000; Vennix, 1996).

System dynamics consists of two types of model analytics. If there are variables that are measurable then stock-and-flow diagrams will be used. If the data does not include measurable objectives, then causal loop diagrams will be used by scholars. The stock-and-flow diagrams are labelled as quantitative models and the causal loop diagrams are labelled as qualitative models (Vennix, 1996). Building qualitative and quantitative models consist of several steps, which have been identified by Martinez-Mayano and Richardson (2013), which is shown in figure 1. In the model in figure 1 the not underlined variables are process steps and the underlined variables are outcomes. This model of the system dynamics approach illustrates that model building is not a linear process. The main process in system dynamics is to understand the problem and system, what demand for an iterative process.

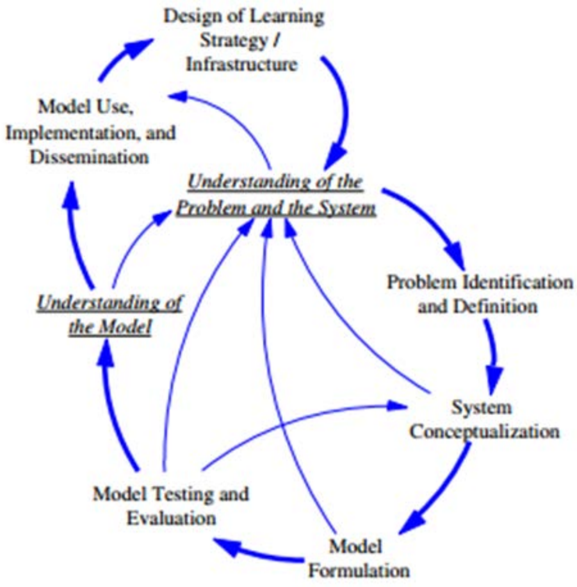
Design of Learning Strategy / Infrastructure
Model Use, Implementation, and Dissemination
Understanding of the Problem and the System
Problem Identification and Definition
Understanding of the Model
System Conceptualization
Model Testing and Evaluation
Model Formulation

Figure 1: Model of the system Dynamics Approach (p108, Martinez-Mayano & Richardson, 2013)

2.3.2 Big data analytics

Big data analytics is a method designed for specific types of problems. These problems are characterized by three to seven V's, based on the stream of literature applied. All problem definitions contain the following three V's. The first three V's represent volume, variety and velocity. Volume represents the size of data sets. Variety represents the diversity of data sources coming in different formats. At last velocity refers to the speed the data is generated (Sivarajah, Kamal, Irani & Weerakkody, 2016). The large data sets from different sources and in different formats raise questions about the meaning of data. A change of meaning of data arises if data is used in other contexts then it was collected for. This change of meaning relates to the validity of the results and the reliability of the data. If data is collected for a certain purpose and afterwards used in a different context, the validity of the result become questionable. The article of Sivarajah et al (2016) introduces several V's for this problem. The problems are related to variability, value and veracity. Variability represents a constant change of meaning, depending on the context it is used and the context it is collected. For example, the Gross Domestic Product (GDP) is 10.000. In the western world, this will be labeled as poor, however in the African content this might be labeled as relatively rich. Furthermore, written and numerical data that is used for big data analytics is a smaller selection of all possible data, compared to mental data. This has consequences for the veracity. Veracity represents imprecision and inconsistency in large data sets and is more about understanding the data and its integral discrepancies. Let's take the GDP example again, we have ninety-eight people earning 1000 euro a year, and 2 earning 500.000 a year. Then the average is around 10000. The only fact is that most people are extremely poor and two are very rich. However, if we only collected the average for each 100 inhabitants, we cannot make conclusions about the standard of living in a certain country. This value is lost because of the average, but we can reveal how rich a region is compared to other regions. This example exemplifies the V of value. Value refers to valuable information inside the data, because not all data or combinations of data are valuable or meaningful.

These six V's represents the type of problems big data analytics are applicable for. Big data analytics use special algorithms to manipulate the V's in an efficient manner. In addition, it has a special analytics process and infrastructure that fits the needs of the V's, for example cloud computing ensures maximum computing capacity and efficient storing (Sivarajah et al, 2016; Wang, Xu, Fujit, Liu, 2016). Furthermore, special management prescriptions are used to get organizations into big data analytics practices (Sivarajah et al, 2016; Janssen, Voort & Wahyudi, 2017). However, this last advantage is not interesting for the present research, because it is situation specific. The first two are important, because no other methods offer these advantages as a combination.

Big data analytics consists of five types of analytics. The two most used types are the descriptive and predictive analytics (Sivarajah et al, 2016). Descriptive analytics examine data and information to the current state of a situation in a way that developments, patterns and exceptions become valuable (Joseph & Johnson, 2013; Gandomi & Haider, 2015). Predictive analytics try to forecast and use statistical modelling to determine possibilities in the future (Waller & Fawcett, 2013).

2.4 Messy information problems, system dynamics and big data analytics

In previous paragraphs is described what messy information problems exactly are, how they are solved, and what the tools of system dynamics and big data analytics are. This paragraph combines all this information and illustrates the relationships between the three main concepts of this study. The result is a conceptual model, which will be used as the preliminary model for building the causal loop diagram in chapter four.

The present study defines messy information problems as problems with large limited data sets and dynamically complex. Limited data sets are characterized by incompleteness, which lead to biased information and ambiguous information. The bigger the characteristics are the messier the messy information problem. This relation forms our first part of the conceptual model. In figure 2 shows the model conceptualization of this explanation.

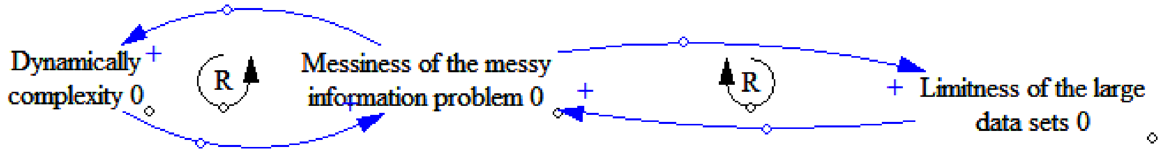
Dynamically complexity 0
+
R
Messiness of the messy information problem 0
+
R
+
Limitness of the large data sets 0

Figure 2: Messy information problem

Now the system dynamics part comes in. System dynamics uses mostly mental data for building a system dynamics model (SD-model). This SD-model helps to receive- insights into the system and its behavior. The use of mental data leads to a decrease of validity and reliability of the model, because mental data is incomplete, biased or ambiguous. This decrease of validity and reliability also leads to reduced quality in the insights of the model. Furthermore, system dynamics uses confidence tests to increase confidence in the model. However, some data sets are too large for normal confidence tools, they are not applicable in all situations. These relationships lead to the following addition of the conceptual model, which is shown in figure 3.

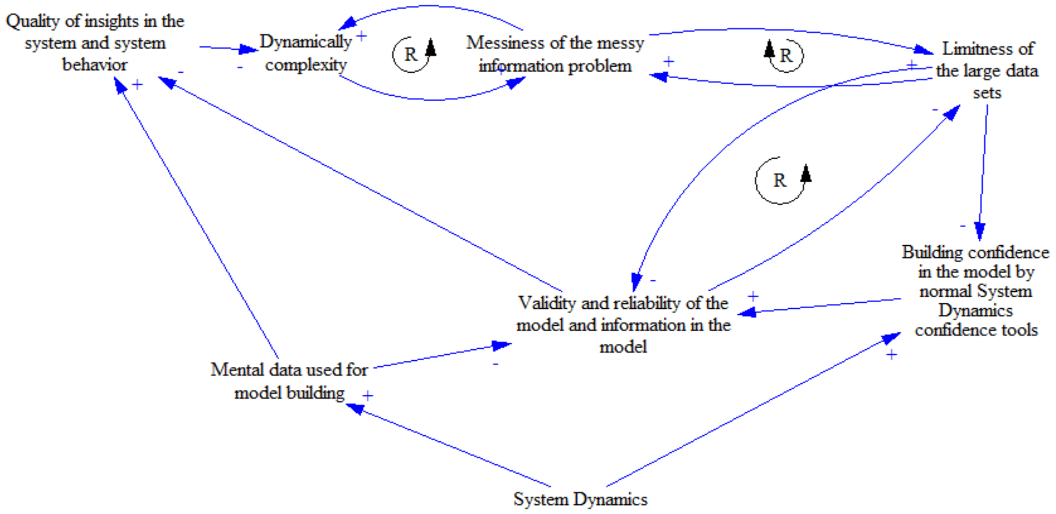
Quality of insights in the system and system behavior
Dynamically complexity
R
Messiness of the messy information problem
R
Limitness of the large data sets
R
Building confidence in the model by normal System Dynamics confidence tools
Validity and reliability of the model and information in the model
Mental data used for model building
System Dynamics
+
-
+
-
+
+
-
-
-
+
-
+
+

Figure 3: Dynamics of messy information problem solving by use of system dynamics

The last part consists of adding big data analytics to the model. Big data analytics can support to describe model elements like relationships. However, the limited data sets limited the practicability of the descriptive analytics. A reason for this is missing, ambiguous or faulty/biased data. This addition of big data analytics into the conceptual model is illustrated in figure 4.

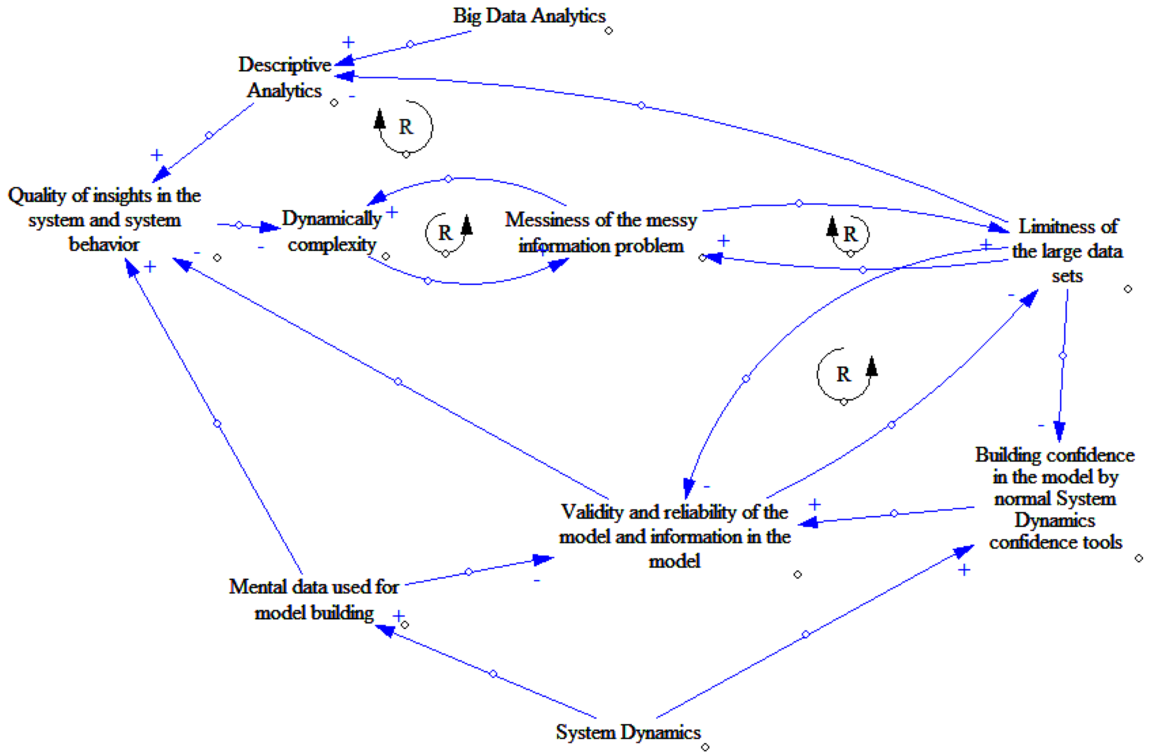
Big Data Analytics
Descriptive Analytics
R
Quality of insights in the system and system behavior
Dynamically complexity
R
Messiness of the messy information problem
R
Limitness of the large data sets
R
Building confidence in the model by normal System Dynamics confidence tools
Validity and reliability of the model and information in the model
Mental data used for model building
System Dynamics

Figure 4: Using system dynamics and big data analytics individually on messy problems

Figure 4 shows clearly that both separately methods are not capable of handling messy information problems themselves. They both are stuck in reinforcing feedback loops. However, a combination of both methods will reduce the problems of the large limited data sets and dynamic complexity. This relationship leads to the last addition of the conceptual model shown in figure 5. The extension of this conceptual and preliminary model is given in chapter four. Chapter four shows the results of the literature review that which will answers the question about how a combination of system dynamics and big data analytics establish a solution in solving messy information problems.

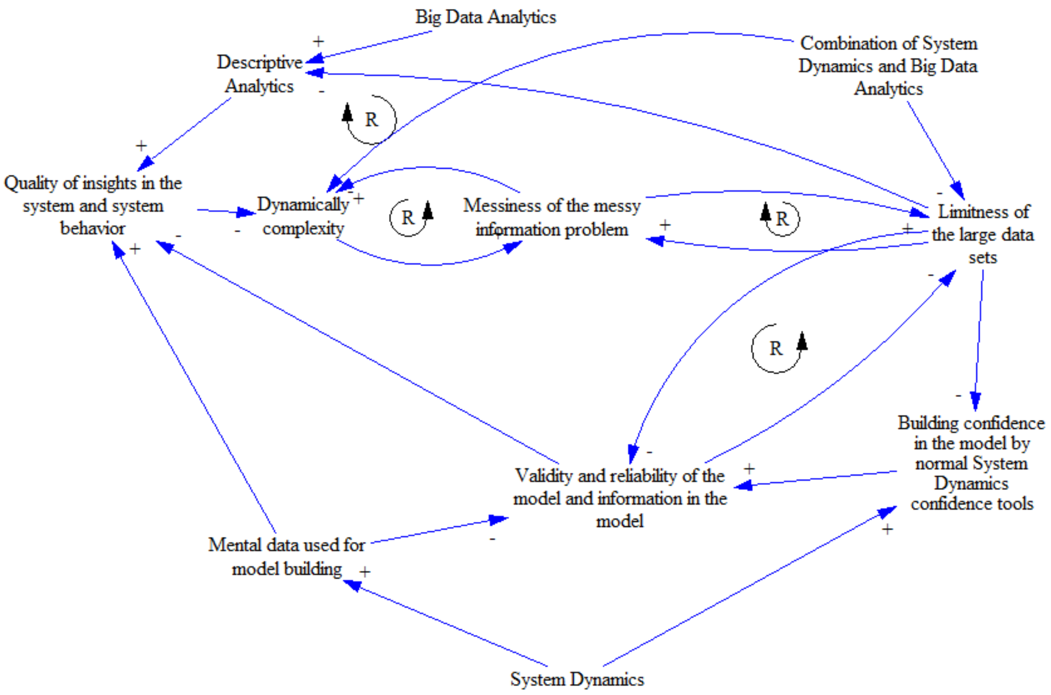
Big Data Analytics
Combination of System Dynamics and Big Data Analytics
Descriptive Analytics
Quality of insights in the system and system behavior
Dynamically complexity
Messiness of the messy information problem
Limitness of the large data sets
Building confidence in the model by normal System Dynamics confidence tools
Validity and reliability of the model and information in the model
Mental data used for model building
System Dynamics
R
R
R
R
+
-
+
-
+
-
+
-
+
-
+
-
+
-
+
-
-
+
-
+
+

Figure 5: Final preliminary model

3. METHOD

In this chapter the research method is described. In the first paragraph the arguments for the research design are discussed. In the second paragraph the literature review and the process of building a causal loop model are discussed. In the third paragraph the confidence tests that were used are explained.

3.1 Choice of research design

Several research designs were considered to discover which research design is best to apply to answer the main question. Three different options were suitable namely: starting a new project, analyzing existing projects or receiving insights based on parts of other projects. The considered research designs based on these options were: an experiment, a case study or a mixed design (e.g. a literature review and building and validating a causal loop diagram).

An experiment was not an option because of several reasons. First, the present study is an explorative study, and is not an evaluative study (Yin, 2014), which makes that an experiment is not suitable. Second, because of the high costs for application and the risk of no success makes that companies did not want to cooperate in an experiment. -(Sivarajah et al, 2016). At last limited time was available for the present study therefore it was impossible to carry out an experiment.

A case study was not an option as well. First, research projects who use a combination of system dynamics and big data analytics are rare and most of the time not public. As far as we are aware, only one research could be found, which is the research of Fiddaman (2017). This research had the limitation that the results and research process were mainly not public. Several reasons why combination is rare and not public could be given. First, a large percentage of the system dynamic projects are not published or public (Featherston & Doolan, 2012). Second, big data analytics most of the time contains privacy related data or threatening data, which is a reason to make big data projects not public (Sivarajah et al, 2016). The last reason is that big data analytics is a relatively new type of method. The slow adoption of big data analytics by organizations is caused by the high costs for implementation (Sivarajah et al, 2016). To conclude, no relevant studies are available for public, therefore a case study is not an option.

Several literatures are available about solving problems with one of the methods. Thus, a combination of the insights within these papers could help to solve the messy information problems. With these combined insights scholars can further develop knowledge in the field of solving messy information problems. In addition, the insights can help practitioners to create cases for further research and to refine the final model of the present study.

To establish a knowledge base of solving messy information problems by using a combination of system dynamics and big data analytics, a literature review is the best option. According to Randolph (2009), a literature review is a valid method for receiving methodological insights, discovering important variables relevant to the topic, identifying relationships between ideas and practices and understanding the structure of the subject. In the present research insights are demanded about what possibilities there are on combining system dynamics and big data analytics, which is in line with the goals of Randolph (2009). To gain new knowledge a model building process has been applied. A model is a good way for knowledge building according to Schwaninger and Groesser (2008). According to Sterman (2000) building causal loop diagrams is a good way to get insights in problem understanding, for example messy (information) problems. Because causal loop diagrams ask for different variables, causal relations and a structure, a literature review is suitable to support this model building.

3.2 Literature review and causal loop diagram building

To conduct the present study, a multistep process has been followed. The results of the process can be found in chapter four and five. The first step was a literature review. The first step within the literature review was selecting articles from the leading academic journals on big data analytics and system dynamics. All articles till the end of June 2018 are included in the present study. The articles that were used were only the articles that were marked as main articles from System Dynamics Review from the System Dynamics Society. The articles of Big data Review and Journal of Information and Management were selected from the database of the sciencedirect.nl website. For the articles of the Journal of Information and Management, the search concepts “Big Data” and “System Dynamics” were used. The discovered articles were used in the present research.

The second step was selecting the relevant articles based on their title. Articles were divided in categories. The first category was “Not a fit with the subject”, this category is for the articles that did not fit into the other four categories. The second category was “Case study/ Specific model”, an exception that was made for case studies about research methods, because they could give methodological insights. The third category was “Possible relevant”, this category existed of articles about tools and methods in system dynamics or big data analytics. The fourth category was “column or a personal story”, this category existed of articles that were subjective or not based on research. The last category was “Too specific research method”, which contained articles that focus on very specific points on tools.

In case category three was not detailed enough, category two and five could be used to fill in missing parts. In this study this option was not applied. The first categorization was done on title level. Afterwards the abstract of the remaining articles in category three were used for selection. At the end, the number of articles in column three of table 1 (e.g. summary & conclusion) were left. After reading the whole articles, column four of table 1 displays the number of articles that were used for the literature review and the causal loop model building. Table 1 represents the selected numbers of articles in each stage.

Table 1: Table with number of selected articles in each phase of the selection

	Start	Title	Summary & Conclusion	Final
System Dynamics Review	388	94	34	28
Journal of Big Data Review	128	39	22	22
Journal of Information and Management	64	20	3	3

The second step in the literature review was analyzing the articles based on relationships. For this phase each article was read and important parts in the text were marked. This selection was established by using at least one of the following criteria. First, the text part is related to one of the five sub concepts of the present study. Second, the text gives an explanation of the tool or method that was used. At last, the text part gives advantages, disadvantages or limitations of the tool or method. Based on the marked text parts relationships and variables were extracted. The marked text parts of each article were clustered for each article based on the found relationship. This step was applied in the program Excel.

The third step consists of building a causal loop diagram. If you as reader are not familiar with causal loop diagram, an explanation of causal loop diagrams is added into appendix B. The relationships and variables found in the second step were connected in this phase. Sometimes a translation was needed, therefore the specific variables were translated into a system dynamics and big data version. For example, “parameter names” (big data analytics) and “variables” (system dynamics). This was needed for combination of insights from system dynamics and big data analytics literature. The next step consisted of building small structures out of the insights for each sub concept of messy information problems, like ambiguity. The structures were built, according to the standards of Sterman (2000). Afterwards all small structures were combined with the preliminary model explained in chapter two. The starting point was the model in figure 2 (chapter 2), however it needed some changes before all the structures could be added. The main reason for these changes was that the model was to abstract and did not contain the sub-concepts of limited information and dynamic complexity and described big data analytics and system dynamics to abstract. At last the adapted model was extended and further adapted based on the discovered insights of the literature review. This resulted in the model explained in chapter four.

3.3 Confidence in the causal loop diagram

There are several criteria for using models for knowledge building (Schwaninger & Groesser, 2008). First, it needs to have the ability to support or falsify a theory, thus it should be testable. Second professionals need understanding of how the models works and how to use the models. This is achieved using the following criteria namely clarity, precision, validity, reliability and simplicity of the model. In addition, it needs to cover the field of interest of the professional. Moreover, the model should show clear parts for further development of knowledge (Schwaninger & Groesser, 2008).

In the present study the proposition to solve messy information problems with a combination of system dynamics and big data analytics which conforms the first criteria. The last criteria can be confirmed by validation of the model. Validation of a causal loop model is a controversial subject, because different perspectives are described in literature. Two main perspectives are the positivists and the social constructivists (Barlas, 1996). According to Barlas (1996) the positivists perspective is about statistical testing of the model and examine if the model is true or false. The output of the model should match with the “real” output. This perspective is different compared to the perspective of the social constructivists who are not only interested in the match of the output behavior, but also in the explanation of the behavior. The goal of the validation of a social constructivist model is to validate the internal structure. This internal system structure of the problem presentation can be illustrated in different ways, because it is a combination of different statements. These statements can be set up in different ways, representing the same problem. In conclusion there is no single model, however there are multiple possible models. This perspective has the assumption that each model contains the modelers world view. Models are not correct or incorrect, however they lie on a continuum of usefulness (Barlas, 1996). The present study is mainly about explanation of system behavior to discover the possibilities. Therefore, the model exists of multiple causal relationships. These relationships are built by the modeler itself based on scientific papers, which result in a social constructivism perspective on validation. In this perspective, validation tests are called confidence tests, because of this lack of true or false. The higher the confidence the better usability (Barlas, 1996).

According to Sterman (2000) and Forrester and Seinge (1980) different kinds of validation can be used. For causal loop diagram not all confidence tests are workable, because many tests are for models that contain quantitative data. In the present study a structure assessment test will be done. This test asks whether the model is consistent with knowledge of the real system relevant to the purpose. Barlas (p. 189, 1996) proposes two direct structure tests. The first test is a theoretical structure-confirmation test. The model in chapter four is already based on a literature review, therefore the added value of this test is low. The literature review used in the model in chapter four offers already a certain confidence in the model, therefore the added value of this test is low. The second test is an empirical structure-confirmation test. To apply this test, two type of sources are used. First the available information of the study of Fiddaman (2017) is used to confirm or disconfirm the found relationships. The other empirical source are interviews with experts. In total three respondents take part in the present study. Two respondents are academic professionals and in addition teachers in the field of System Dynamics from the Radboud University in Nijmegen. The other respondent is a lector from the HAN University of Applied sciences who has a team that is working with Big Data Analytics. Unfortunately, a fourth respondent dropped out, because of sickness and vacation. The people were selected based on their academic career and their professional skills. They have expertise on a conceptual level on system dynamics or big data analytics. In addition, they are professionals that have expierence in system dynamics or big data analytics. To conclude, they can judge the model from a theoretical perspective and practical perspective.

The process of the interviews consisted of two steps. The interview started by asking if they accept that the interviews were audio recorded. All the participants agreed. Next, our definition of messy information problems was to set the context of the interview. To examine if the model in this study is understandable and useable for professionals the model was shown to the participants. At the same time a story was told about the model elements. They were instructed that they can interrupt the story if they could not follow the story or they did not agree with certain things. The story was about a new project of solving a messy information problem, where the respondents participated in as an expert on system dynamics or big data. The project leader of the project made a model (the model in chapter four) to illustrate the possibilities of combining big data analytics and system dynamics to solve messy information problem. The project leader described the found relationships and possibilities of combinations in the story. The second step of the interviews were statements. The relationships that formed the loops in the model were transformed into statements and stated to the respondents. The respondents needed to confirm or falsify the statements and give an explanation about it. The analysis for the confidence test of the different found loops for combinations were done by labelling the interviews and by using the research description of Fiddaman (2017). The labelled interviews parts are added in a table for each loop, which can be found in chapter five. For each table a conclusion was written. Based on all conclusions an overall conclusion is given for each found combination of system dynamics and big data analytics.

4. MODEL CONSTRUCTION

In this chapter we extend the preliminary model of chapter two by the insights of the literature study. In The first paragraph the results of the concept of dynamic complexity and the extensions of the model related to these results are discussed. In the second paragraph the concept of limited information and the extensions of the model related to these results are discussed. In the last paragraph the loops in the final model are discussed.

4.1 Dynamic complexity

In chapter two we showed that the concept of dynamic complexity existed of two sub-concepts: feedback loops and time/delays. However, during our analysis we experienced that these sub-concepts were hard to work with. Both sub-concepts contained two other sub-concepts namely model building and model behavior understanding. Therefore, the decision has been made, that the sub-concepts used in the present study are model construction and model behavior understanding, because the model structure reveals system behavior (Allen, 1988).

4.1.1 Building the dynamic complexity representation

To build the dynamic complexity of a problem into a model, the real complexity of the situation should be simplified. In chapter two we introduced descriptive analytics, which can support in describing the dynamic complexity, as a results descriptive analytics can help to build the model. In this paragraph the different big data analytics tools that offer possibilities for identifying the structure elements which form the patterns of models are discussed. In the present study the discovered descriptive tools are data mining, machine learning and deep learning, which are discussed below.

4.1.1.1 Deep learning as a tool for structure building

A tool that combines variables finding and relationship building in a technological way is deep learning (Prusa & Khoshgoftaar, 2015). In regular machine learning approaches for text mining and learning approaches no formal solution is available for all problems, this will be discussed later. This means that a researcher needs to determine and implement the best possible solution. Deep learning has not this disability and can extract high level features out of low level data (Najafabadi et al., 2015). Najafabadi and colleagues (p2, 2015) explain deep learning in their paper as “automated extraction of complex data representation (features) at high levels of abstraction. These algorithms develop a layered hierarchical architecture of learning and representing data, where high-level (more abstract) features are defined in terms of lower level (fewer abstract features)”. In practice deep learning is a more advanced form of machine learning. Deep learning can be differentiated into different neural networks. One popular network is the convolutional neural network, which is effective in feature extraction and classification. Although learning these networks is a slow and computational expensive task (Prusa, Khoshgoftaar, 2017).

Within deep learning two fundamental building blocks are important: autoencoders and restricted boltzmann machines (RBM). Autoencoders are networks existed of three layers: input, hidden and output. The RBM contains only a visible and a hidden layer. Both building blocks perform best on non-local and global relationships and patterns in data (Najafabadi et al., 2015). Deep learning tries to reduce the error between input and simulated behavior. For messy information problems, both building blocks can be useful, depending on the problem situation. The first step in deep learning is describing the key variables of the system and collecting the data that describes the behavior of the key output variables (Abdelbari & Shafi, 2017). This process can be established by experts. In practice, deep learning is for example applied in semantic indexing, because it offers a more efficient way of presenting data and makes it useful as a source for knowledge discovery (Abdelbari & Shafi, 2017).

In case of messy information problems deep learning is interesting. We added the insights (figure 6) described about deep learning into the final model (figure 26), because of several reasons. First, deep learning decreases the interaction of humans in the process of feature extraction. Deep learning allows them to remove human bias in feature engineering and preservation of more information as the original data can be used for training. Abstraction in this case decreases the impact of faulty data (Prusa & Khoshgoftaar, 2017), thus the model is less sensitive to local changes. Furthermore, deep learning can handle complex, non-linear patterns, which are hard or impossible to handle for more traditional machine learning, (text) data mining or feature engineering algorithms (Najafabadi et al., 2015). At last, according to Prusa and Khoshgoftaar (2017) an advantage of deep learning compared to other techniques such as data mining is that it not necessarily requires specialized domain knowledge. According to Najafabadi, and colleagues (2015) a limitation of deep learning is that deep learning lacks appropriate objectives in learning good representations and therefore further research is needed. A critical view of the representations needs to be added in the model. A critical view can for example exist of applying theory that explains the representation or specific domain knowledge of an expert. However, this will be discussed later in this chapter.

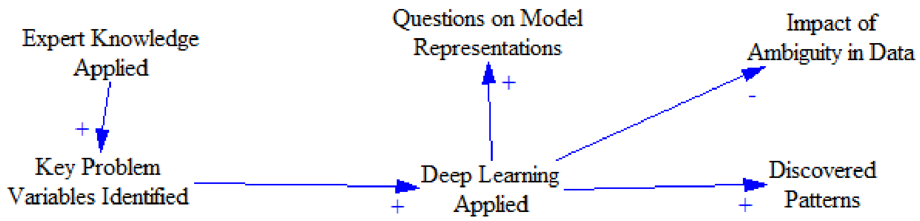
Expert Knowledge Applied
+
Key Problem Variables Identified
Questions on Model Representations
+
+
Deep Learning Applied
Impact of Ambiguity in Data
-
Discovered Patterns
+

Figure 6: Insights in Deep Learning

4.1.1.2 A deeper understanding of classification as tool for variable identification

Variables contain units of measurement and can be analyzed. A way of identifying variables is to create main topics and run algorithms to divide main topics into sub-topics. This process is defined in big data literature as classification (Mujamdar, Naraseeyappa & Ankalaki, 2017). A more specific way of classification is language detection algorithms (LDA), also known as text mining. Text mining is a widely used tool for classification. LDA offer a way in selecting several topics for identification. These algorithms differ from machine learning, because machine learning optimizes the number of clusters by cross validation and heuristics. LDA offers with this feature the possibility to control the granularity of the analysis (Pröllochs & Feurriegel, 2018). At the end the clusters are named based on the topics in the cluster. To gain a better understanding of how datamining works, an example of the paper of Herland, Khoshgoftaar & Wald (2014) is given. In their study they used a multistep approach on discussion forum posts, that is called SHIP. The first step existed of some basic text processing, to structure the data entries. The second step existed of entity extraction, only medical relevant posts were selected. The third step consisted of expression distillation whereby posts were divided over the predefined number of clusters. In this case five classifiers (e.g. personal experience, advice, information, support and outcome) were used. This part of the analysis has been done by using the J48 decision tree algorithm using the WEKA tool. The fourth step is aggregation, whereby the level is changed from post level to discussion level. In this step the data is aggregated to a level, where topics can be extracted. The outcomes can be helpful in the future to help patients by giving them the information they need based on pre-determined medical condition and to connect them to similar patients.

Within classification the clusters or the whole analysis mostly contains noise, because of the linguistic content and its characteristic imprecision (Pröllochs & Feurriegel, 2018). This problem arises in the field of financial markets, where managers face an incentive to frame their disclosures in a certain way, which causes biases. Based on the used sources, certain caution is needed, especially if only one source is used. Combining sources could be an outcome to this problem (Pröllochs & Feurriegel, 2018). Another solution for this problem is to identify high quality sources, based on their available contextual data. For example, Twitter data contains more context than search query engine data (Herland, Khoshgoftaar & Wald, 2014). Another high-quality source is purposive texts. Kim and Andersen (2012) describe purposive texts as: "First, purposive text data arise from a discussion involving key decision makers or stakeholders in the system under study. The participants in the discussion have a sophisticated knowledge of the system, and their expert knowledge becomes the basis of the causal maps being elicited. Second, purposive text data capture the participants' focused discussion on the system and the problem at hand. As a result, the data frequently depict causally and dynamically rich discussions. Third, the discussion captured in the data should reflect a frank and unfeigned conversation of the decision-making group.". Classification tools have one advantage, they reduce the impact of the modelers own assumptions about the system, however they will never exclude these biases completely (Kim and Andersen, 2012).

Another problem of using a text data mining tool is that sometimes random clusters can be made, because of unstructured, complex duplicative textual databases that contain many homonyms and synonyms (Al-Hassan, Alshameri & Sibley, 2013). This problem is a problem of ambiguity in texts. In such case replacing synonyms and erasing not important homonyms from the text (like parts of company names) can help to get better clustering. This process requires a certain understanding of the text and context, such as expert do (Al-Hassan, Alshameri & Sibley, 2013).

As we discussed above classification can be used for identification of variables but has some limitations. The data and the quality of the data that are used for classification have an impact on the quality of the results. For a messy information problem, a certain understanding of quality of the data is essential to judge if this tool is applicable (Sivarajah, Kamal, Irani & Weerakkody, 2016). A lot of information can be found in documents of organizations including purposive texts (Kim & Andersen, 2012). We decided to add these relationships (figure 7) about classification into the final model (figure 26). The insights into classification lead to the structure in figure 7.

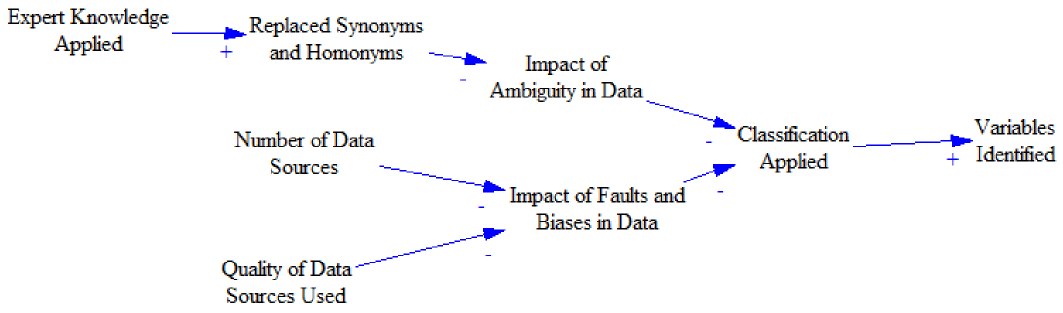
Expert Knowledge Applied
+
Replaced Synonyms and Homonyms
-
Impact of Ambiguity in Data
-
Classification Applied
+
Variables Identified
Number of Data Sources
-
Impact of Faults and Biases in Data
-
Quality of Data Sources Used
-

Figure 7: Insights into classification

4.1.1.3 A deeper understanding of clustering, as tool for parametrization

A special way of classification is clustering, which is also called unsupervised classification. Clustering is a quantitative process (Mujamdar, Naraseeyappa & Ankalaki, 2017), however it can be used for variable identification. Clustering techniques can be divided in two main types. The first type is probability-based methods. These methods assume that clusters come from a mixture of distributions. In fact, this is more parameters estimation. This type has the limitation that the tool becomes less applicable to massive data sets and streams (Aletti & Micheletti, 2017). The second type is distance-based approaches. These approaches depend on the distance. This type tries to minimize the mean squared distance between the data and their closest centers (Aletti & Micheletti, 2017). This type of classification can be used for variable identification. In the present study, we use the first type as clustering and the second type as classification, because it can be used for variable identification. Within these two types, many subtypes of clustering and classification algorithms have been created such as partitioning clustering (i.e. classification with a predefined number of clusters), hierarchical clustering (i.e. building a tree), density-based methods (i.e. clustering with a threshold) (Mujamdar, Naraseeyappa & Ankalaki, 2017).

Clustering has one important disadvantage, it becomes difficult when the data is high dimensional, for example images (Kaur & Datta ,2015). According to Kaur and Datta (2015) high dimensional data suffer from the curse of dimensionality. The first implication of this curse is that if the dimensionality of data grows, the relative contrasts between similar and dissimilar points decrease. The second implication is data tend to group together differently under the different sets of dimensions. However, there are some solutions for this problem (Kaur & Datta ,2015). First, divide the data into subspaces and then cluster them, so called subspace clustering. However, subspace clustering is a time expensive process. Second, the Apriori algorithm, that is based on hierarchal clustering, is also a promising approach to find all possible higher dimensional subspace clusters from the lower dimensional clusters using a bottom up process (Kaur, Datta, 2015). A third solution of dealing with high dimensional data is removing irrelevant clusters. Afterwards the top down algorithms of Proclus and Findit could be appropriate (Kaur & Datta, 2015). However, these solutions and this problem are very specific and image recognition is most of the time not a relevant process in messy information problems, because most images in these problems are graphs and can also be converted to numbers. That is why we left out these solutions and the problem. We know that clustering is to specify parameters, after classification identified the variables for the clustering process. These insights lead to the structure of figure 8 and were added into the final model (figure 26).

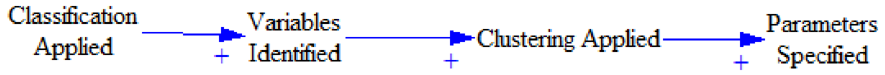

Figure 8: Insights into clustering

4.1.1.4 A deeper understanding of association rule extraction

Another data mining technique is association rule mining, that searches for relationships between variables. Kumar and Toshniwal (2015) discuss this technique in road accident data. In the paper association rule mining is used to identify variables that have their effect on the occurrence of an accident. Before they use association rule mining, they used K-mode clustering. “Old” techniques like regression analysis are still popular but are limited compared to association rule mining. These techniques have limited capacity to discover new and unanticipated patterns and relationships that are hidden in conventional data bases (Kumar & Toshniwal, 2015). However, K-mode clustering is a way of classification, for identifying variables. Therefore, first classification is needed to extract association rules afterwards, so called relationships or connections between the variables. Because relationships are essential building blocks in building patterns (Sterman, 2000). However, we think certain caution is needed, because sometimes relationships are found that are totally irrelevant. For example, older children are better in math compared to younger children. This is a fact; however, it does not mean that a person is more intelligent. These insights about association rule extraction lead to the structure of figure 9, which was added to the final model (figure 26).

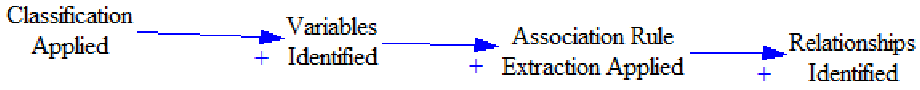

Figure 9: insights into association rule extraction

4.1.1.5 A deeper understanding of sentiment analysis for polarity extraction

A tool that can establish polarities based on relationships between variables is sentiment analysis. Sentiment analysis is also known as opinion mining and studies people's sentiment towards certain entities. Variables in this case are entities. Most sentiment analysis has been applied by using data mining (Sohangir, Wang, Pomeranets, Khoshgoftaar, 2018). According to Sohangir, Wang, Pomeranets and Khoshgoftaar (2018) the hierarchical learning in deep learning convolutional neural networks makes it perfect for sentiment analysis, because of the transformation of input over more layers. An important aspect of sentiment analysis is identifying the features that contain the sentiment, before classification can be executed (Fang & Zhan, 2015). El Alaoui and colleagues (2018) constructed a dynamic dictionary of words polarity based on selected set of hashtags related to a given topic.

To conclude, different methods can be used for sentiment analysis. Sentiment analysis in the present study is useful for identifying polarities of relationships. In case of a causal loop diagram, this is a polarity of the relationship. In stock and flow diagrams this polarity helps in building the formula. This was the last step in building patterns. That is why it is important to add this feature into the final model (figure 26). These insights lead to the following structure in figure 10.

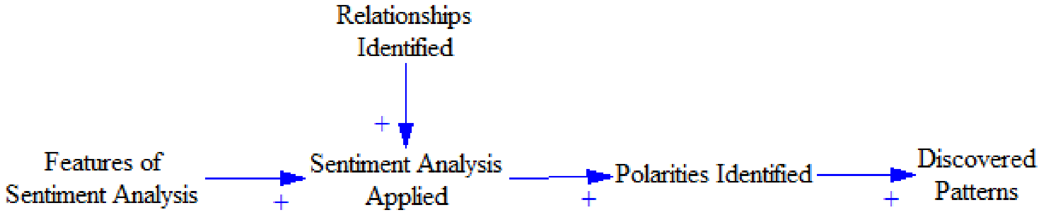

Figure 10: insights into sentiment analysis

4.1.1.6 Machine learning

Besides the datamining tools (e.g. classification, clustering, association rule mining; Lamari & Chah Slaoui, 2017), another tool within Big Data Analytics is machine learning. Within machine learning two types of learning can be recognized. The first type is incremental learning, which means that the learner updates the model of the environment when new significant experiences from stream data become available. The second type is dynamic ensemble learning, which means that data are divided in small data chunks. On each data chunk the classifier is trained independently. Finally, heuristic rules are developed to organize these classifiers into one super classifier. A classifier could be compared with variables that are connected. These two types of learning are relevant when concepts drift arises. Concept drift is the change of the impact of variables over time. Although incremental learning adopts not suddenly to concept drift, it is faster and more noise resistant. Ensemble learning adapts more easily, because it sets the size of data chunk and assigns different weighting values to different base classifiers (Zang, Zhang, Zhou & Guo, 2014). Executing machine learning frameworks like MapReduce provide an effective solution against the weakness of scalability of the incremental learning. The only limitation is that it does not offer iterations (Liu, Wang, Matwin & Japkowicz, 2015).

In conclusion, machine learning can identify variables and relationships as small patterns. In addition, it can identify concept drift. Thus, additional to the datamining techniques, machine learning is useful. Especially if there is disagreement between problem holders which variables are most important. This problem is also called ambiguity. People do not know they are talking about different subjects, because of this change over time. Each person misses a part of information. In this case the information about the change is missing. Because of the messy information problems, this identification of concept drift can decrease or clarify the problem. That is why these insights (figure 11) are added to the final model (figure 26).

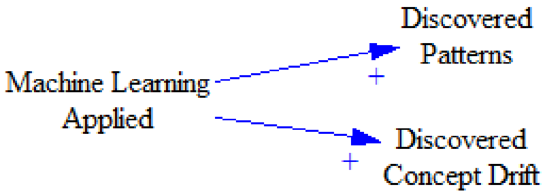

Figure 11: Insights into machine learning

4.1.2 Dynamic complexity understanding

In the first part of the paragraph we explained the building possibilities for the dynamic complexity. The next part is about the understanding of the dynamic complexity of the model by analyzing it. Understanding of the model arises by analyzing the structure and model's behavior which is caused by the structure (Sterman, 2000).

4.1.2.1 Time as part of the systems behavior

Important aspects of behavior are time, delays and behavior over time. According to Conboy, Dennehy and O'Connor (2018), time and delays do not receive the amount of attention in business analytics literature that is needed. However, speed and time are associated with better management and control of complexities, which lead to better performance and business value. However, this is not always the case. Therefore, time should be taken as a loose approach, because it arises with a certain uncertainty and time, speed and acceleration are not always good (Conboy, Dennehy & O'Connor, 2018). Especially in system dynamics problems it is important to consider many endogenous sources of system behavior, that can affect the system. When keeping this in mind better problem explanations can be given and possible policies and decision can be made (Richardson, 2011). Because time is an important aspect of behavior. Limited information of the problem causes uncertainty about the exact time usage in the model, very strict time use is necessary to reduce the consequences of policies that solve the problem. Therefore, we added this loose time approach into the final model (figure 26) to get the best understanding of dynamic complexity and solutions of the problem. Furthermore, these insights related to this loose time approach can be found in figure 12.

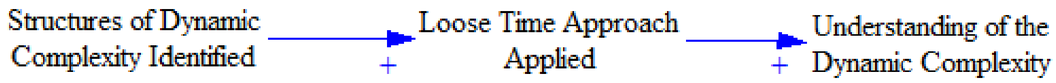
Structures of Dynamic Complexity Identified
+
Loose Time Approach Applied
+
Understanding of the Dynamic Complexity

Figure 12: Insights into the loose time approach

4.1.2.2 Reducing complexity in system structures

Relatively small models are easier to understand compared to larger models. However, larger models become hard to handle because of the limited capabilities of humans (Schoenenberg, Schmid, Ansah & Schwaninger, 2017). In the literature a lot of methods are described to gain a better understanding of models. One of these methods are variety filters. These filters reduce the complexity of a model to promote accurate interpretation of the model. These filters encompass interpretive model portioning, structural model portioning and algorithmic detection of archetypal structures (ADAS; Schoenenberg, Schmid, Ansah & Schwaninger, 2017). The goal of these filters is to reduce model complexity to receive similar but simpler causal structures. Although these filters increase the understanding of a model, the filters also cause information loss. Besides the filters, simplification can be used for understanding the model. Simplification is used to distil only the essential structures that produce the fundamental behavior. Simplification has the advantage that people understand the model easier. However, the disadvantage is that some people will find it too simplistic. If simplification is used depends on the use of the model, namely a larger the model offers in general more trust by clients. However, simplified versions offer the possibility after validation that they are transferable to other domains (Eberlein, 1989; Keren Sayal & Barlas, 2006). Model understanding is an important feature of solving messy problems. Therefore, the complexity will be reduced (Sterman, 2000). This better understanding is been established by making the problem understandable for people by fitting the complexity to the capacity of the brain (Schoenenberg, Schmid, Ansah & Schwaninger, 2017). That is why reducing complexity is added to the final model (figure 26), as shown in figure 13.

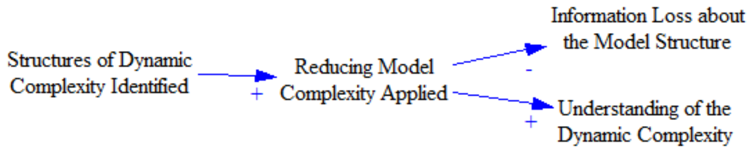
Structures of Dynamic Complexity Identified
+
Reducing Model Complexity Applied
Information Loss about the Model Structure
-
+
Understanding of the Dynamic Complexity

Figure 13: Insights into complexity

4.1.2.3 Sensitivity Analysis for identification of the dominant parameters

Another way of understanding a model is the quantitative perspective, for example identifying parameters that are important and to investigate how they affect the system behavior. In System dynamics literature these examples are also called sensitivity analysis (Walrave, 2016). Sensitivity analysis has many options for semi-automated methods. A semi-automated sensitivity analysis method has the advantages to go beyond ad hoc experimentation and conduct systematic analysis of intervention thresholds. In addition, it can handle main modes of behavior: (1) zero/constant behavior; (2) linear growth/decline; (3) exponential growth/decline; (4) goal seeking growth/decline; (5) S-shaped growth/decline; (6) growth and decline or decline and growth; and (7) oscillation with/without growth/decline (Walrave, 2016).

For solving messy information problems with policies, it is essential to find dominant parameters and the effects of parameters on the system, to build a policy that fits a wide and realistic range of influences (Sterman, 2000). Therefore, sensitivity analysis is an essential step in solving messy information problems. However, to do sensitivity analysis parameters should be specified to have a point to start from. These insights lead to the structure in figure 14, which is added to the final model (figure 26).

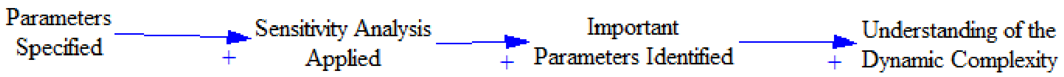
Parameters Specified
+
Sensitivity Analysis Applied
+
Important Parameters Identified
+
Understanding of the Dynamic Complexity

Figure 14: Insights into sensitivity Analysis

4.1.2.4 Loop and structure recognition for understanding

Loop identification and structure identification are important parts of understanding system behavior. Hayward and Boswell (2014) call this loop impact. They propose an algorithm that identifies which loop or combination explains stock behavior, this algorithm is based on the Pathway Participation metric method.
Another tool is a behavioral approach of feedback loop dominance in explaining how structure drives behavior (Ford,1999). To let this application succeed, the tool should analyze automated, they should have a clear view of loop dominance and a view on how a loop impacts the system behavior. In addition, analysists should be careful about shadow loops that could arise. They seem like dominant loops; however, they are not (Ford, 1999).

Another tool in the system dynamics field is statistical screening of the model to identify the high-level structures and parameters (Ford & Flynn, 2005). Simple correlation coefficient analysis could be used to show the relative importance of model inputs at different times, which is the same as concept drift as explained earlier (Ford & Flynn, 2005). The feedback mechanisms that drive system behavior once identified can be used to design and tests policies (Taylor, Ford & Ford, 2010). In chapter two we discussed that feedback loops are important for understanding dynamic complexity, therefore identification of loop impact is an important step, because it explains the main behavior. The same applies for statistical screening. At the end all tools try to identify important high impact parameters or structures in relation to the systems behavior, which is essential for understanding the dynamic complexity. That is why we added these insights (figure 15) into the final model (figure 26). However, it depends on the type of model and problem situation which exact tool can be applied.

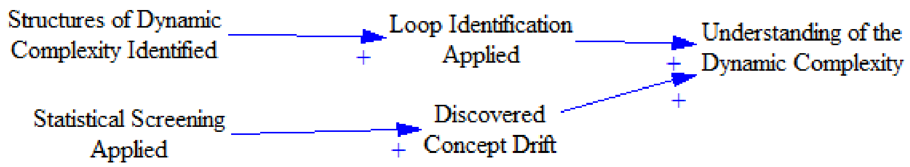
Structures of Dynamic Complexity Identified
+
Loop Identification Applied
+
Understanding of the Dynamic Complexity
+
Statistical Screening Applied
+
Discovered Concept Drift

Figure 15: Insights into loop and structure recognition

4.2 Limited information

The concept of limited information ambiguous, incomplete and biased and faulty data, as we discussed in chapter two. In this literature review we discussed several techniques to reduce ambiguity, incompleteness or biases and faults in data, to reduce their impact on the problem analysis. Within the next sub-paragraphs, the results of the literature review on the concepts ambiguity, incompleteness and biases and faults are discussed.

4.2.1 Ambiguity

Ambiguity arises when parts of the context are missing (Zhan & Dahal, 2017). Each source of ambiguity asks for a different solution to decrease the impact. In each sub-paragraph a source of ambiguity and a solution for this ambiguity is discussed.

4.2.1.1 Solving variables ambiguity by descriptions and machine learning

Constructing variables in general is a source of ambiguity, because people are creating the variables. Therefore, the definitions of these variables will be different between people. Therefore, an accurate description of the variable decreases the impact of ambiguity and misunderstanding (Felipe Luna-Reyes & Lines Anderson, 2003). According to Jacobsen and Bronson (1987) a variable has a good description if it meets the criteria of reliable, realistic and face validity. In addition, a variable need to be measurable and should have a clear unit, that leaves no space for ambiguity. However, it is difficult if there are no recognizable units of measurement. In such case a certain construction of other measurable units should be used (Jacobsen and Bronson (1987). An example variable is work pleasure, because work pleasure could be translated in the time an average time a person works at a company. However, each representation has its limitations, therefore a good consideration about the three criteria should lead to a workable outcome.

Another example of ambiguity in a variable is a business process that changes over years. Depending on the time, the person looks different to the process. A way to explore the process of change on topics is identifying concept-evolution (i.e. unknow classes in data), feature evolution (i.e. progression of new features and regression of old features) or concept drift (i.e. slow changes in the concept of stream). This identification process is possible with machine learning; however, this is challenging because of the infinite length stream data (Chandak, 2016). Moreover, identifying the changes in the process gives additional information, which can be used to clarify the context of data. According to Hurtado, Agarwal and Zhu (2016) a combination of association rule mining and forecasting can help to identify possible future trends on topics.

Variables are essential in analysis of dynamic complexity, because they are cornerstones of the model that visualizes the dynamic complexity (Sterman, 2000). Therefore, a high-quality description is needed to build and analyze the model in a correct way and to decrease the impact of ambiguity. A variable description and the impact can change over time, therefore identifying concept drift is important to understand the behavior over time. In conclusion, both a high-quality description of variable and identifying concept drift are important in solving messy information problems and are added as the structure in figure 16 to the final model (figure 26).

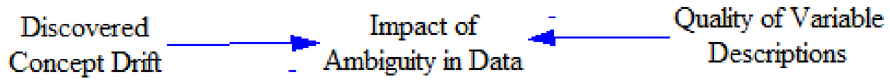

Figure 16: Insights into ambiguity

4.2.1.2 Association rule extract and ambiguity

One kind of ambiguity arises in texts, because many words have synonyms or homonyms. A latent semantics approach in association rule mining can discover topics in documents that even have no words in common (Hurtado, Agarwal & Zhu, 2016).

Especially in messy problems and messy information problems, documents are a big source of data (Kim & Andersen, 2012). In addition, messy information problems are most of the time problems that involve multiple stakeholders as discussed in chapter one and two. Each group of stakeholders has their own way of communication, which causes these synonyms and homonyms. Therefore, overcoming this source of ambiguity is essential to oversee the whole problem and to use the whole spectrum of document data sources. The insights described above are visualized in figure 17 and are added into the final model (figure 26).

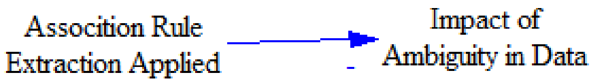
Assocition Rule Extraction Applied
-
Impact of Ambiguity in Data

Figure 17: Relationship of association rule extraction

4.2.1.3 Enrichment by information to overcome ambiguity in data

Most of the time context is missing in short texts (e.g. social media), which makes it difficult for methods like deep learning to extract patterns out of these texts. Semantic enrichment can be used to overcome this problem. Zhan and Dahal (p1, 2017) explain this process as: “This is accomplished by taking the nouns, and verbs used in the short texts and generating the concepts and co-occurring words with the help of those terms. The nouns are used to generate concepts within the given short text, whereas the verbs are used to prune the ambiguous context (if any) present in the text. The enriched text then goes through a deep neural network to produce a prediction label for that short text representing it’s category”. Thus, finding concepts of words from the short texts and identifying the words that appear frequently with those words are defined as semantic enrichment.

Besides ambiguity on a low level, ambiguity also rises on a problem definition level. Ambiguity in problem definition can be defined as what are the key drivers in the problem (Sterman, 2000), also called key variables (Abdelbari & Shafi, 2017). In this case ambiguity can be used to further strengthen the dynamic hypothesis. Therefore, Ambiguity needs to be transformed into questions and these questions needs to be answered. Questions should be tested, for example by using a theory or expert knowledge. Answers could lead to new questions and at some point, the dynamic hypothesis is rich enough for modelling. This enrichment leads to new variables in the model and to new possibilities for policy options. As a result, the modelling process becomes more purposeful (Mashayekhi & Ghili, 2012).

Enrichment of data is an important way to overcome ambiguity, because it adds the missing context. Nowadays social media is a huge and important source of information (Lee, 2017). However, less cases of usage of this data are found in the literature of system dynamics, these cases can be found in the literature of big data (Lee, 2017). Therefore, adding the possibilities with this data helps the practitioners of system dynamics to overcome ambiguity. Overcome ambiguity on problem level is an essential step in the modelling process in system dynamics and solving a messy problem (Sterman, 2000). In addition, it is one of the essential usages of mental data (Sterman, 2000). The insights described above are illustrated in figure 18 and are added into the final model (figure 26).

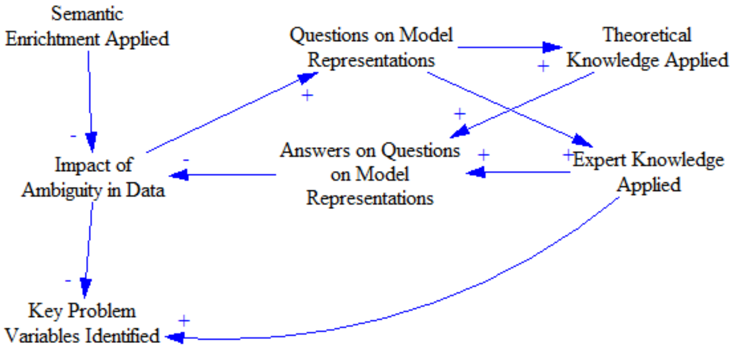
Semantic Enrichtment Applied
Questions on Model Representations
Theoretical Knowledge Applied
Impact of Ambiguity in Data
Answers on Questions on Model Representations
Expert Knowledge Applied
Key Problem Variables Identified
-
-
-
+
+
+
+
+
+

Figure 18: Insights into enrichment

4.2.2 Incompleteness

Incompleteness is an important cause of ambiguity, faults or biases in data. Incompleteness of data can be found on different levels. The first level is missing data in datasets (e.g. like time series data). The second level is missing data in hierarchy (i.e. like only a theme of a picture instead of elements of it). Both levels are discussed below.

4.2.2.1 Solving incomplete parameter data by using parameter estimation tools

Data can be missing on both levels during solving problems. Data could be missing for specifying parameters. The data could be partly missing or missing completely. In case of partly missing data, interpolators could be used or other statistical tools, such as mean substitution, could be used (Tilaye Wubetie, 2017). However, these last types of tools most of the time lead to biases in parameter estimation. Other algorithms show promising results, such as the expectation/maximization algorithm and multiple imputation algorithm. These algorithms are based on iterative solutions in which the estimated parameters are the imputed values. As a result, the estimated parameters will change. (Tilaye Wubeti, 2017). In case a complete variable is missing, these values could be guessed or estimated. Applying estimation can be done by using clustering. After clustering an estimation of the values can be made based on the main variable. Thus, measuring components can be constructed by investigating natural correlations (Tilaye Wubeti, 2017).

In messy problems, which include messy information problems, missing parameter data is an essential part of the problem (Sterman, 2000). Methods that can help to fill in these missing parts are therefore essential before the model can be analyzed for solving the problem. However, keep in mind that each tool that tries to fill in missing data, increases the impact of the faults and biases, because - the estimated values are not the exact real value. Every situation asks for a different tool, we clustered these tools as one super tool in figure 19. This figure is, added into the final model (figure 26).

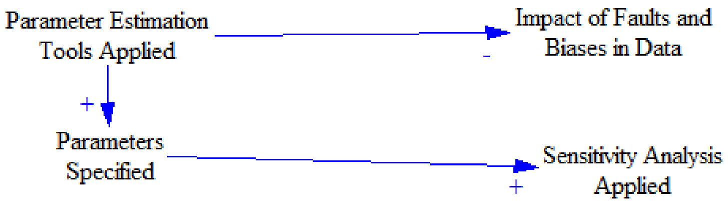
Parameter Estimation Tools Applied
Impact of Faults and Biases in Data
-
+
Parameters Specified
Sensitivity Analysis Applied
+

Figure 19: Insights into parameter estimation tools

4.2.2.2 Solving incomplete parameter data in the model

For solving missing data in hierarchy, indirect interference can offer a solution. According to Hosseinichimeh, Rahmandad, Jalali and Wittenborn (p.158, 2016) indirect interference is: “to match properties of empirical and simulated data in order to estimate the unknown parameters of the model of interest". Indirect interference has several advantages according to Hosseinichimeh et al. (2016). First, there are little limitations to the types of models to which it can be applied. The only limitation is that indirect interference can only be applied to models that are simulated for different values of its parameters. Second, it is relatively inexpensive to compute when the auxiliary model uses a maximum likelihood estimator. In addition, the auxiliary model parameters have small variances and they could be matched reliable with few simulations. Third, the indirect interference inherits the beneficial properties of the estimation method used for the auxiliary model. Fourth, indirect interference can be used for both estimating and validating a model. According to Hosseinichimeh et al. (2016), an important step in indirect interference is to identify archetypes (i.e. structures), which can be done by using the semi-automated ADAS method, that we mentioned in the paragraph 4.1.2.2. Although ADAS has some limitations. First, it cannot detect system archetypes, however it can only detect modeling components that fulfill the structural requirements as system archetypes. True archetypes are more than simple two-loop constellations. That is why the process still asks for a modeler judgement to interpret the results. Second, the output of the algorithms is sometimes hard to interpret especially in large models, because they only use structural requirements. (Hossenichimeh, et al., 2016).

An approach that has similarities with the indirect interference is the POPS algorithm of Yücel and Barlas (2011). The algorithm estimates parameters that meet the qualitative features of the desired system behavior. The algorithm identifies the pattern and estimates open parameters to meet the pattern. According to Yücel and Barlas (2011) the algorithm is applicable in model calibration, model testing and policy analysis. POPS is one of the few parameter search tools in the system dynamics field that trust on optimization-based heuristics for an effective search. However, POPS is the first tool that is pattern oriented, compared to others that are point based oriented. An advantage of this is that even when reference data is missing, POPS can run the model. Another advantage is that POPS can be used to find the changes in parameters for changing the system behavior. A disadvantage is that POPS is not able to receive an exact numeric fit to omitted variables in the model. This can be solved by using a multi-objective optimization algorithm, by using this algorithm the numeric aspects of objectives can be integrated in the model. (Yücel & Barlas, 2011).

In paragraph 4.2.2.1 we discussed estimation tools based on quantitative data of variables. In this paragraph we discuss tools that are based on the behavior of the model itself. Especially in messy problems, where not all data about variables is available, a lot of information is available about the behavior of the system (Sterman, 2000). This behavior is most of the time the motive for solving the problem, therefore it should not be left out in the model structure building process. As parameter estimation is an important aspect of the building process, we decided to illustrate these insights in figure 20. Figure 20 is added into the final model (figure 26).

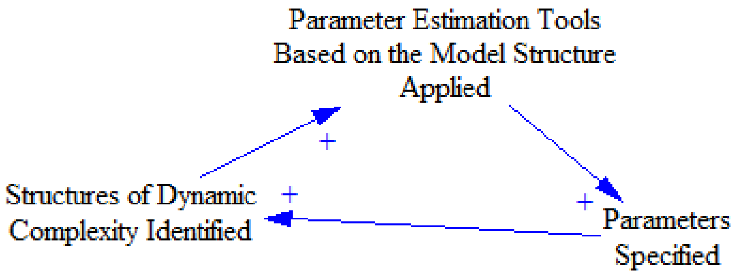
Parameter Estimation Tools Based on the Model Structure Applied
+
Structures of Dynamic Complexity Identified
+
+
Parameters Specified

Figure 20: insights into model-based parameter estimation

4.2.2.3 Solving incompleteness using mental models

Another way of solving incompleteness in data is the use of mental models. Databases are most of the time not complete. A lot of information about systems is stored in mental models (Ford & Sterman, 1998). According to Doyle and Ford (p. 17-21, 1998) mental models of systems are: “A mental model of a dynamic system is a relatively enduring and accessible but limited internal conceptual representation of an external system whose structure maintains the perceived structure of that system”. This knowledge is subjective, because it is personal and context specific (Doyle & Ford, 1988). The information can be found by people who work with the system. We call these people experts. To describe, examine and use this knowledge, methods are used for elicitation, articulation and description of knowledge (Ford & Flynn, 2005). Examples of these methods are interviews and coding. However, these methods and the mental data contain or cause bias, because the information is filtered (Forrester, 1987; Hall, Aitchison & Kocay, 1994; Ford & Flynn, 2005). Expert knowledge can be helpful to guess parameters. Although this is not an extremely trustworthy method, it sometimes is the only solution (Ford & Flynn, 2005). This causes uncertain parameters, but sensitivity analysis (Clemson, Tang, Pyne & Unal, 1995) or statistical screening could display if the uncertainty in the parameters is acceptable. Sometimes the parameters are not part of the essential structure in the problem, then it is not a problem to estimate the parameters (Ford & Flynn, 2005).

As we discussed above, mental data is an important source of information in system dynamics (Sterman, 2000). That is why mental data should be included into the final model. In paragraph 4.2.1.3 and 4.2.2.2 we already discussed two ways of applying expert knowledge. The part we missed was the part of using expert knowledge to build structure or falsify them. Missing numerical and written data cause missing structure parts. Mental data can prevent these missing structure parts, because it is a richer form of information (Forrester, 1987). That is why we illustrated this last part in figure 21 and we added this figure into the final model structure (figure 26).

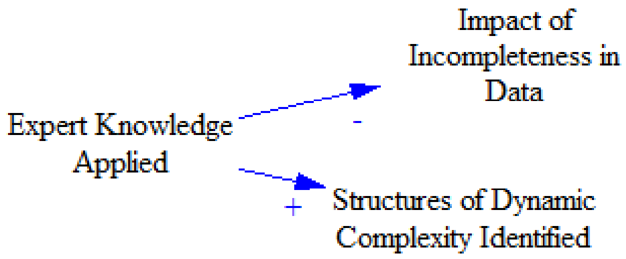
Impact of Incompleteness in Data
Expert Knowledge Applied
-
+
Structures of Dynamic Complexity Identified

Figure 21: insights into mental models

4.2.3 Faults and biases

As a result, faults and biases a certain context is missing or is wrong. These missing context causes ambiguity and makes it hard to solve problems. Biases are caused by humans who are interacting in the process. Faults are faults in the data and can exist for different reasons. First biases are discussed in this paragraph and then faults in data are discussed.

4.2.3.1 Theory to examine the outcome from faulty data

Biases caused by humans exist because of their limited processing capacity and their previous experience and knowledge (Goodwin, 2014). In a process of knowledge discovery humans always start with the knowledge they have in their minds. This sometimes result in missed possible explanations. In big data analytics the starting point is more process driven compared to traditional research methods that are more theory driven. Traditional ways of inductively proposing theories and deductive confirming theories are all theory driven approaches. For theory we used the definition of Elragal and Klischewski (p2, 2017): “Theory as a statement of relationships between units measured or constructed in the empirical world”. However, process driven approaches, for example big data analytics, have certain advantages compared to theory driven methods. Big data analytics is less influenced by biases and less subjective, but still asks for different theoretical guidance in each process step (Elragal & Klischewki, 2017). Big data analytics seeks to gain insights out of data and disrupts how research is conducted, by processing big data sets to generated relational insights and predictions. However, the shift from theory driven to process driven poses new challenges such as parameter selection, big data validity, reliability reflection and an overall theoretical framework supporting method selection and result interpretation (Elragal & Klischewki, 2017). Because big data analytics are almost impossible to check, this increase the chance of questions about the validity of the results. This is the V (i.e. veracity) that appears. The results should be validated by answering the validity (Sivarajah et al, 2016).

During the model building process, theoretical knowledge and expert knowledge can be used by disconfirming for example patterns, relationships and variables to improve the model. In addition, it judged the model to build confidence. More different patterns decrease the reliability of the model, because more explanations of the behavior are possible. However, expert knowledge elicited by interviews can cause biases because of the forgetting information effect and faults in the coding and analysis process (Lines Anderson et al., 2012).

A lot of tools in Big Data are not transparent because they often offer multiple explanations for the behaviors. These multiple explanations cause validity issues in the results. Removing the use of theoretical knowledge or mental data should directly lead to questions about the validity aspect of the model in the whole messy information problem solving process. To summarize, removing theorical knowledge or expert knowledge results directly in a decrease of validity of the model for practioners and for scientific importance. Theoretical knowledge is different from expert knowledge, because theoretical knowledge is scientifically proven. Therefore, we split expert knowledge and theorical knowledge as can be seen in figure 22. Figure 22 is a representation of the description above and is added in the final model (figure 26).

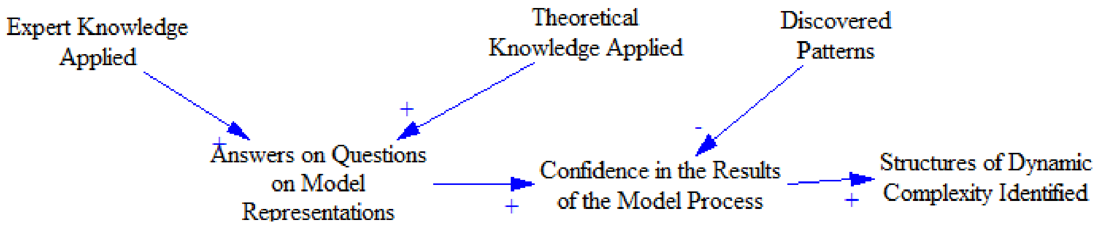
Expert Knowledge Applied
Theoretical Knowledge Applied
Discovered Patterns
+
+
-
Answers on Questions on Model Representations
Confidence in the Results of the Model Process
Structures of Dynamic Complexity Identified
+
+

Figure 22: Theoretical knowledge

4.2.3.2 Identifying faults to make solutions for it

Data bases most of the time contain contradictory data that undermines the soundness of the information. Old methods such as pie charts are difficult to analyze in big databases with contradictory data. Therefore Chika Nwagwu, Okereke and Nwobodo (2017) established an application named “ConTra” to identify contradictory information. In their tests ConTra scores very high on accuracy in big data sets. When contradictory data is identified a solution needs to be made, such as removing the contradictory data.

However, removing contradictory data leads automatically to incompleteness of the data, which can result in missing results. Besides removing other solutions could be used, such as further investigating what is responsible for the contradiction. To investigate if fluctuations are important there are general big data statistical tools to explore patterns and correlations. These general big data statistical tools overpower individual fluctuations (Chika Nwagwu, Okereke & Nwobodo, 2017). In addition, big data analytics is specialized in handling data that contains faults. Moreover, it is hard to decide when data contains faults and therefore removing data can result in missing essential information. We decided to not illustrate this insight in the model because we assume that big data is better in handling faults compared to practitioners. Practitioners can assume in large data sets which data is faulty and the impact of removing this data.

4.2.3.3 Sensitivity analyze the impact of possible faults

Sometimes it is not possible to correct the faults, because the wrong values could not be identified or replaced with better values. In such case sensitivity analysis can offer a solution, by identifying how sensitive a parameter is and by investigating which effect certain wrong values have within certain ranges. In such case the exact parameter is not important, but the pattern characteristics of the output behaviors are important. The sensitivity analysis reduces the impact of faults and biases in data, because it considers multiple modes of behavior of the system (Hekimoglu & Barlas, 2016).

Until now we only discussed parameter values and not graphical function, also called look-up variables. Look-up variables depend on variable variables. These variable variables most of the time contain a high degree of uncertainty (Eker, Slinger, Daalen van & Yücel, 2014). Therefore, a good analysis of the impact of these functions on the behavior is important, to anticipate better on possible effects when faults are in the data. Eker, Slinger, Daalen van and Yücel (2014) propose a method that is based on multiplication of the model function by a distortion function, which is a specific form of variable parametrization. In this way the function is transformed into the problem of parametric sensitivity, which has been discussed earlier in this paper.

Another type of sensitivity analysis is based on reaching certain goals and searching for the uncertainty that is possible within these goals. Sheng, Lee and Hay Lee (2012) propose a robust goal-seeking design that incorporates eigenvalue analysis with a mathematical programming approach. This approach has the advantage that it does not require specific probability distributions, who are assumed for the uncertain variables. In addition, it does not require that the user must specify subjective weights as in typical utility maximization approaches. The user specifies the set of design goals that needs to be satisfied (Sheng, Lee & Hay Lee, 2012).

In conclusion, sensitivity analysis was already introduced in paragraph 4.1.2.3. However, we did not discuss the effects of the limited information concept. In this paragraph we discussed how sensitivity analyses can be applied to overcome (partly) the concept of fault and biases in different settings. Letting out why sensitivity analysis is useful in the final model, decreases the usability of the model. Sensitivity analysis does not illustrate how to understand the effects of faults and biases in relation to the dynamic complexity understanding- We discussed multiple tools in this paragraph. Because each problem requires different tools, we took together as one variable. Figure 23 illustrates the visualized insights and is added into the final model (figure 26).

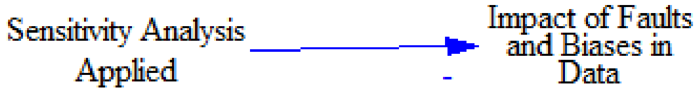
Sensitivity Analysis Applied
-
Impact of Faults and Biases in Data

Figure 23: insights into sensitivity analysis and faults and biases

4.2.3.4 Confidence and validity tests to explore the fault and bias impact

An important step for identifying the right dynamic complexity structures is to investigate important faults or biases in the model and data, by validation of the model. According to Barlas (1996) validation is defined as not only reproducing a system behavior, but also explaining how the behavior is generated and explaining how the behavior can be changed. Validation is hard, because there are no established formal tests that can be used in deciding if the internal structure is close enough compared to the real world. Therefore, most system dynamic practioners not use the term validation but use the term confidence. A lot of confidence tests are designed for validation of dynamic complex models. These tests can be divided in structure tests, and behavioral tests (Barlas, 1996). Structure tests determine the validity of the model by comparing the system to knowledge of the real system. Behavioral tests determine the validity of the model by applying certain behavior tests on model general behavior patterns. After these tests are applied, a test can be done to measure how accurate the model can reproduce the major behavioral (Barlas, 1996). Some experts in system dynamics are convinced that statistical significance tests are not that relevant because of the assumptions of the tests such as not autocorrelated, not cross-correlated or normally distributed (Barlas, 1996). These assumptions are almost never met by using a system dynamics model, because they are autocorrelated and cross correlated by their nature (Barlas, 1996). Another technical difficulty is that statistical significance tests will be ambiguous or misleading, if data are corrupted with measurement errors. In addition, there are no single output variables in system dynamics, which is called the multiple hypothesis problem (Barlas, 1996).

Validation or confidence building is an important step in solving a messy information problem. This step determines if the found structure is usable for designing solutions or policies. The reliability of a model will increase when the faults and biases in a system decrease. Figure 24 illustrates this step and is added into the final model (figure 26).

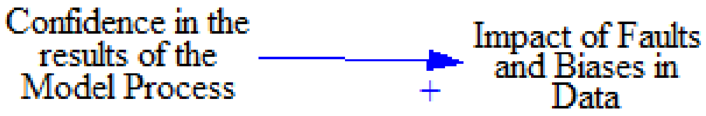
Confidence in the results of the Model Process
+
Impact of Faults and Biases in Data

Figure 24: insights in confidence and validity

4.3 The model construction

This paragraph is about the model construction. It starts with reorganizing the preliminary model from chapter 2, based on the insights found during literature review. Furthermore, the construction itself is discussed. The paragraph ends with the identification of the found loops.

4.3.1 Preliminary model

In chapter two a preliminary model was constructed. However, during the literature review we investigated that we must make some changes in the model before the insights we found could be added. The first change we made is dividing the limited information into the defined concepts of incompleteness, ambiguity and faults and biases. In this case we divided the variables 'into impact of incompleteness in data', 'impact of ambiguity in data' and 'impact of faults and biases in data' to replace the variable of 'limitless of the large data sets'. As discussed in chapter two incompleteness causes ambiguity, faults and biases, which are the relationships between the variables. In addition, faults and biases in data cause ambiguity, as discussed in chapter two as well. The reason for this change is to show the impact on the different concepts of limited data and to gain better insights into solving messy problems.

Another important change is that dynamic complexity is split into two sub concepts. In chapter two we discussed dynamic complexity into feedback loops and time and delay. However, during the literature review and model construction, we found that the concepts did not work. The concepts in the literature review were more about building the dynamic complexity and understanding the dynamic complexity. Building the dynamic complexity could be interpreted as the feedback loop. Furthermore, understanding of dynamic complexity could be interpreted as time and delay, in this way we can understand the behavior over time of the model. That is why we changed the sub-concepts into ‘structures of dynamic complexity identified’ and ‘understanding of the dynamic complexity’ for dynamic complexity. The last change we made is that we defined the exact system dynamics tools and big data tools and their effects. These changes lead to the change from the preliminary model illustrated in figure 5 (chapter two) to the end model illustrated in figure 25.

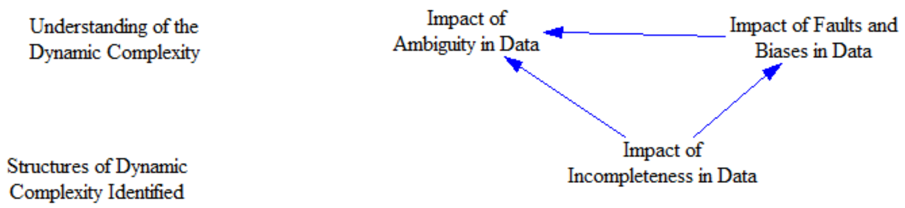
Understanding of the Dynamic Complexity
Impact of Ambiguity in Data
Impact of Faults and Biases in Data
Structures of Dynamic Complexity Identified
Impact of Incompleteness in Data

Figure 25: Starting point for model construction

4.3.2 Construction of the model.

The final model is constructed after stripping and changing the preliminary model (figure 5) into a starting point for model construction (figure 25). The model construction is done by combining all the separates structures from figure 6 to figure 24 into one structure. The combined structure, also called the final model is shown in figure 26.

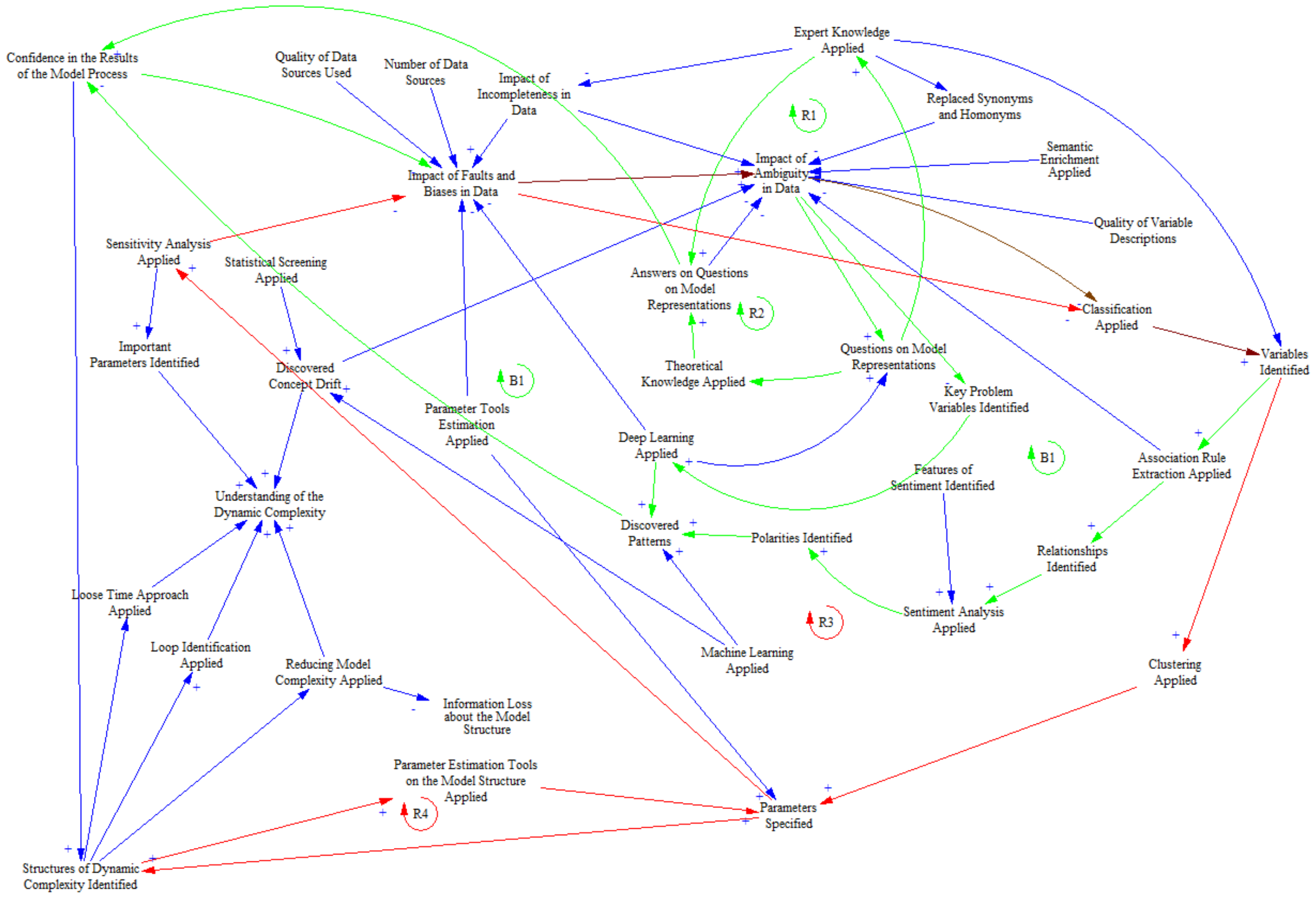
Confidence in the Results of the Model Process
Quality of Data Sources Used
Number of Data Sources
Impact of Incompleteness in Data
Expert Knowledge Applied
Replaced Synonyms and Homonyms
R1
Semantic Enrichment Applied
Impact of Ambiguity in Data
Impact of Faults and Biases in Data
Quality of Variable Descriptions
Sensitivity Analysis Applied
Statistical Screening Applied
Answers on Questions on Model Representations
R2
Classification Applied
Important Parameters Identified
Questions on Model Representations
Variables Identified
Theoretical Knowledge Applied
Discovered Concept Drift
B1
Key Problem Variables Identified
Parameter Tools Estimation Applied
Deep Learning Applied
Features of Sentiment Identified
B1
Association Rule Extraction Applied
Understanding of the Dynamic Complexity
Discovered Patterns
Polarities Identified
Relationships Identified
Loose Time Approach Applied
R3
Sentiment Analysis Applied
Loop Identification Applied
Machine Learning Applied
Reducing Model Complexity Applied
Clustering Applied
Information Loss about the Model Structure
Parameter Estimation Tools on the Model Structure Applied
R4
Parameters Specified
Structures of Dynamic Complexity Identified

Figure 26: Causal Loop Diagram

4.3.3 Loop identification

The next step is to identify the loops in the model, to investigate which combinations of system dynamic and big data analytics offer possibilities for solving messy information problems. The loops must- meet three criteria for selection:

1. The loops contain at least one sub-concept of messy information problems (e.g. ambiguity, incompleteness, faults and biases, dynamic complexity structure or dynamic complexity understanding).
2. The loop or combination of loops should contain at least one tool of system dynamics and one tool of big data analytics.
3. In case of a loop combination, the loop combination is a coherent process, that strengthens both tools.

Based on these criteria six loops and two structures could be identified. As can be seen in figure 26, we colored the first structure green and we colored the second structure red. When both a red and a green line is drawn on the same place, a brown line exists (figure 26). We first discuss discus the green structure and afterwards we discuss the red structure.

The green loops exist of four loops: reinforcing loop 1 (R1), reinforcing loop 2 (R2), balancing loop 1 (B1) and balancing loop 2 (B2; figure 26). First, R1 and R2 are two loops that both try to decrease ambiguity by answering questions that arise because of the ambiguity in the data. The only difference between both loops is that R1 answers these questions using expert knowledge and R2 answers these questions using existing theories. Thus, in the basis these loops are the same processes.

Second two balancing loops were found. Both loops try to find patterns, by using big data analytics tools, that cause confidence issues in the model. In B1 the tool that is used is data mining techniques. In B2 the tool that is used is deep learning.

When he structures are combined, they form a system archetype, because always a structure can be formed by a balancing and reinforcing loop and there is at least one relationship which connect the loops. This archetype is called fixes that fails (Meadows & Donella, 2008).

The red loops exist of two loops: reinforcing loop 3 (R3) and reinforcing loop 4 (R4). The R3 exists of the process of specifying parameters with data mining techniques, by first assessing what the effects are of faults and biases in the data by using sensitivity analysis. The R4 exists of the process of identifying parameters based on identified system structure with loops.

The loops combined form an archetype called the structure success to the successful, because there are two connected reinforcing loops (Meadows & Donella, 2008). The behavior of this archetype depends on which loop becomes stronger. Furthermore, if the resources, in this case parameters, are not available anymore these reinforcing loops will stop.

4.3.4 Understanding the whole system

After loop identification, the last step is to investigate if parts of the system structure cause any positive or negative side effects. First, the parameter estimation loop (R3) merges at some point with the green loops. Both loops can reinforce each other in a positive or negative way, depending on which red loop is stronger. Second, deep learning can strengthen R1 and R2 by delivering extra questions about model representations. This supports to fixes that fails characteristics of the archetype. Third, the model shows that ambiguity is the start for solving the limited information concept by using expert knowledge. However, this insight is already known in system dynamics, where mental models are the basis for building dynamic complex models. At last, the reinforcing loops (R1 or R2 and R3) form the system archetype, success to the successful (Meadows & Donella, 2008). These loops illustrate the battle between numerical and written knowledge compared to mental data. However, this is not really a coherent process, because it not strengthens the process of big data or systems dynamics. It only reduces the impact of faults and biases in data and the impact of ambiguity in data, compared to the red structure for example. The red structure uses big data analytics and system dynamics to estimate parameters, which helps the system dynamics process.

5. BUILDING CONFIDENCE IN THE MODEL

After building the causal loop diagram, confidence building is needed. For this confidence building we used expert interviews and the small description of the process of the research of Fiddaman (2017). The confidence building is discussed separately for each loop. We start with explaining both the green loops and the red loops. Furthermore, the interviews were conducted in Dutch, thus they were translated in the English language by the researcher. Moreover, it can be assumed that if no side note is made in the quotes below, the respondents confirmed the relationships during the story and the statements.

5.1 Reinforcing loop one: ambiguity and expert knowledge

The first reinforcing loop tries to decrease ambiguity by answering questions that arise – out of the ambiguity in the data, using expert knowledge. Table 2 shows the outcomes of the interviews and the research of Fiddaman (2017) on this loop. By reading the quotes it becomes clear that this loop is strong in answering questions that are necessary to get confidence into the model and the used data. However, two respondents said certain caution is needed to use this loop in practice, because you can lose valuable insights. In addition, the respondent said the situation is crucial, because if people keep things vague or out of sight, this could lead to wrong confidence. In conclusion, we established more confidence into the loop. However, we have two limitations. First, the analysist needs to know the characteristics of data quality. Second, if the situation leads to consequences for the experts, they could hold information back which leads to false confidence outcomes.

Table 2: Quotes interviews and text about loop reinforcing loop 1

Source	Quotes
System Dynamics Expert 1	- You can say it is a big problem (ambiguity), but it is also basically how people and data are interconnected. Sometimes people keep things vague with a purpose or is it only clear to them to a certain level. That is why they can't explain it deeper. That is why ambiguity is a problem but is there for a reason. You can take it away completely. The question is if you solve the problems by removing it." - On the other side, it has side effects, that can help to make the problem less messy. If you would force people to be clear in what they say, then you can cause exposure of the real difference in their opinions. This results in not signing up to the solution. Sometimes it is better if they just start solving it, while they think they understand it completely. At the end this becomes less messier.

	- If two people are involved in a messy problem, both persons have biases and they don't see the whole problem their self. That is why they think they don't agree, if they could see the whole picture their picture would be more the same as they think.
System Dynamics Expert 2	- With group model building it is the idea by adopting a certain process people are forced to make things explicit. Because they need to formulate it in clear causa and effect relationships, which reveals sometimes they mean something different. Thus, it exposes false consensus. - There is always something ambiguous in the method of system dynamics. If you find something surprising, which you are not sure if it is right, people want to correct it. It is a fault, so they change the model. Everything fits again and can be explained logically. However, that fault could be the insight what people were looking for from the beginning. Thus, it is an outcome. - I think expert knowledge is very powerful. We are walking sponges of information and if someone has expierence it is valuable. The mental models are very important, but you need to elicit it. I am convinced you just cannot tap information out of people.
Big Data Expert	- I you try to make both data (mental, written and numerical) the same, it could work out. However, it could also be you throw away valuable insights. - From the perspective off combination, so called triangulation. It is expected, that you combine multiple sources that give the same answer. Thus, combining is not bad. However, you need to prove that it could be applied in this situation.
Research Fiddaman (2017)	- "Building structural models is pretty quick but calibrating them and testing alternative formulations is a slow process."

5.2 Reinforcing loop two: ambiguity and theoretical knowledge

The second reinforcing loop tries to decrease ambiguity by answering questions that arise out of the ambiguity in the data, using theoretical knowledge. In table 3 we present the quotes of the interviews and the research of Fiddaman (2017) for this loop. The quotes show that this loop is possible and can add a better understanding of system structures. However, the same limitations should be made as described in the last subparagraph. First, the analysist needs to know the theory, to apply the theory in the correct way. Second, if the situation leads to consequences for the analysists, they could apply the theory in a way it fits their needs.

Table 3: Quotes interviews and text about reinforcing loop 2

Source	Quotes

System Dynamics Expert 1	- Understanding is other than solving. (of biases and faults) - It is the same as building a scientific theory. Dutch persons are longer because they drink milk, that is nice, but does not say that much. I would understand it if I knew the mechanism what causes this increase body length. Is because of stronger bones, I don't know. Is it a molecule that other people are missing? If you don't understand the process, they don't understand the problem. So, if big data tools aren't transparent, then a certain understanding misses. - Every test on the model, establishes confidence.
System Dynamics Expert 2	- There is always something ambiguous in the method of system dynamics. If you find something surprising, which you are not sure if it is right, people want to correct it. It is a fault, so they change the model. Everything fits again and can be explained logically. However, that fault could be the insight what people were looking for from the beginning. Thus, it is an outcome.
Big Data Expert	- I you try to make both data (mental, written and numerical) the same, it could work out. However, it could also be you throw away valuable insights. - From the perspective off combination, so called triangulation. It is expected, that you combine multiple sources that give the same answer. Thus, combining is not bad. However, you need to prove that it could be applied in this situation. - Many people see a correlation as a causality, but this isn't always. That is something people do, because they want to see it. It some amateur mechanism, people want to see something and find arguments to it. This succeeds even most of the time. You always can find prove even if the quality is bad.
Research Fiddaman (2017)	No relevant information was available.

5.3 Balancing loop one: confidence and data mining

Balancing loop one tries to find patterns that cause confidence issues by using data mining tools. In table 4 we present the quotes of the interviews and the research of Fiddaman (2017) for this loop. Based on the quotes, the issue arises about the question when a certain structure is a mistake or show the right insights. Furthermore, the relationship from patterns to confidence misses a condition in the model. If the people who work with the messy information problems trust the algorithm and data, then there are no confidence issues. However, in this case we assumed that the data and/or algorithm are not trusted enough. Based on the quotes, we can conclude that the loop can be judged as valid by the participants assuming that the data and/or algorithms are not perfect. This assumption is applicable because in the basis messy information problems have limited data.

Table 4: Quotes interviews and text about balancing loop one

Source	Quotes
System Dynamics Expert 1	- It is a first step. It only says something about the input and output, but the process in between is not transparent. However, more information about the structure is always valuable, but an extra step needs to be made. - Every test on the model, adds some confidence. - Yes, compared to a situation you don't have additional information, it is always valuable. However, compared to a clear validated model, it preferred the validated model.
System Dynamics Expert 2	- If I am in the social constructivism paradigm and I want the best possible representation of the mental models of respondents, then I am for sure there are biases in it. In such case big data is not valuable. If the model should reflect true reality, then there could be a role for big data. So, it depends on the project, if it is applicable. The other thing is, you can have a false idea of consensus, so thinking you mean the same, but in reality, you mean something else. Big Data can analyze this by in these documents you find these words combined with these, so I think you use the same but mean something different. You need to make an assumption. Did we do enough, to say the model is valid? Is the deviation an outcome instead of a fault? Where do you draw the line? - You need to draw the line somewhere. Here I trust the model and here I don't trust it. It is the same observation, that could be a fault as well as an insight. - One trap I see in this model is that we are convinced of model ownership, because commitment should increase the policy outcomes and implementation. The more methods are a black box, the less ownership.
Big Data Expert	- It is field in development, data science. It is still a mess, that is why the biggest achievements are celebrated within those tools that have immense quantities of data and a good infrastructure. So, the Google's of this worlds. Their strength is not the algorithms, but the data. - If you have trust in the algorithms, so you think they can be trusted. Then you also will have trust in the model. However, if you don't trust the algorithms, you don't trust the model as well.
Research Fiddaman (2017)	- "Ultimately, the Big Data approach didn't pan out. I think the cause was largely data limitations."

5.4 Balancing loop two: confidence and deep learning

The second balancing loop tries to find patterns with deep learning that causes confidence issues about the model. In table 5 we present the outcomes of the interviews and the research of Fiddaman (2017) for this loop. Based on the quotes we can conclude that deep learning offers a way for finding patterns and causes confidence issues in the model representations. In addition, it increases the impact of ambiguity, faults and biases. Two important things were mentioned during the interviews. First, deep learning should only be applied in cases where no other methods are better suitable for finding patterns, because of the high confidence issues. Second, the relationship from patterns to confidence in the model misses a condition in the model. As discussed earlier if the people who work with messy information problems trust the algorithm and the data, then there are no confidence issues. However, in this case we assumed that the data and/or algorithm are not trusted enough. Based on the quotes, we can conclude that the loop is judged as valid by the respondents assuming that the data and/or the algorithms are not perfect. This assumption is applicable because in the basis messy information problems contain limited information. Deep learning is not a direct reliable method.

Table 5: Quotes interviews and text about balancing loop two

Source	Quotes
System Dynamics Expert 1	- I think abstraction always helps to prevent faults. - It is a first step. It only says something about the input and output, but the process in between is not transparent. However, more information about the structure is always valuable, but an extra step needs to be made. - Yes, compared to a situation you don’t have additional information, it is always valuable. However, compared to a clear validated model, it preferred the validated model.
System Dynamics Expert 2	- I you can something about input and output, this assumes a lot. You are already far in specifying the boundaries of the problem. We use group model building to extract the problem behind the problem. Thus, applying problem structuring. But the input and output variables (key variables) are the outcome of a half year project. - If I am in the social constructivism paradigm and I want the best possible representation of the mental models of respondents, then I am for sure there are biases in it. In such case big data is not valuable. If the model should reflect true reality, then there could be a role for big data. So, it depends on the project, if it is applicable. The other thing is, you can have a false idea of consensus, so thinking you mean the same, but in reality, you mean something else. Big data can analyze this by in these documents you find these words combined with these, so I think you use the same but mean something different. - You need to draw the line somewhere. Here I trust the model and here I don’t trust it. It is the same observation, that could be a fault as well as an insight. - One trap I see in this model is that we are convinced of model ownership, because commitment should increase the policy outcomes and implementation. The more methods are a black box, the less ownership.

	- Mental models' rule, that is why in the whole words there a differences in concepts. If you name a variable it can have multiple explanations.
Big Data Expert	- The only thing you establish with deep learning is that you draw conclusions on possibly incomplete data. If it is correct and how it happens, that remains a black box within deep learning. Deep learning is applicable in cases where the problem or structure is not clear and never will be. - If the data is messy, then it is interesting what conclusions deep learning can draw. Thus, it is a way of coping. - In one case the algorithm came close in understanding, but something crucial for understanding was left out. - How do you know the conclusions are the right ones? Thus, that is the problem of deep learning. The outcomes are hardly to validate. Unless you have a group of people who did it by hand. Thus, applying deep learning while you don't have any idea about the outcomes and hoping they are right, that is tricky. I don't say it is there are no answers to it, maybe you can check them afterwards. - It is field in development, data science. It is still a mess, that is why the biggest achievements are celebrated within those tools that have immense quantities of data and a good infrastructure. So, the Google's of this worlds. Their strength is not the algorithms, but the data. - If you have trust in the algorithms, so you think they can be trusted. Then you also will have trust in the model. However, if you don't trust the algorithms, you don't trust the model as well.
Research Fiddaman (2017)	Not relevant information was available.

5.5 Reinforcing loop three: data mining and parameters

The third reinforcing loop tries to identify parameters with data mining techniques. In table 6 we present the quotes of the interviews and the research of Fiddaman (2017) for this loop. The respondents gave this loop less attention in the interviews to this loop, because probably they have more expertise on the topics of the other loops. However, the loop is confirmed by both the big data expert and system dynamics expert. Deciding on how incomplete the data is, goes in the same way as in a sensitivity analysis. It means that estimating can only be done with the right data, which is the main process in this loop with data mining. One important condition is that the characteristics of the data decide whether this method could be a success.

Table 6: Quotes interviews and text about reinforcing loop three

Source	Quotes

System Dynamics Expert 1	- Yes, that make it easier. (using data mining for parameters) - Yes, if you have a certain part of the model parameters based on quantitative information, then you have confidence in this part. What causes better guesses or estimation on other parts.
System Dynamics Expert 2	- By mapping parts of the system, it helps to develop even more parameters, that need to be estimated afterwards. - By mapping the system, you see more connection, you can use for estimation.
Big Data Expert	- I think this is to specific for the situation, it depends on the data and how the incompleteness in the data is.
Research Fiddaman (2017)	No relevant information was available.

5.6 Reinforcing loop four: system structure and parameters

The second reinforcing loop tries to find parameter values based on the found structure. In table 7 we present the quotes of the interviews and the research of Fiddaman (2017) for this loop. Based on the quotes both system dynamics experts confirm the loop structure. However, one expert proposes a missing link in the model namely a relationship from this loop to the confidence in the model structure variable.

Table 7: Quotes interviews and text about reinforcing loop four

Source	
System Dynamics Expert 1	- The context of the parameter is in these situations a little bit clearer. This makes it easier to estimate how big and in which range a parameter is. I think this is better, than just a guess without context.
System Dynamics Expert 2	- With this process you miss a validation probability, to check if everything is consistent. I you use the model to fill in the blind spot, you cannot use the model to validate. - The parameter is with that an artefact of the model. Not something you can validate with the model. Because then you validate the model with itself. - If you only miss one parameter, then you can guess the last one. If you have two open parameters, then you can have unlimited possibilities to fill in. It

Big Data Expert	No relevant information came out of the interview.
Research Fiddaman (2017)	No relevant information was available.

6. CONCLUSION AND DISCUSSION

In this chapter we answer the main question and discuss the results and the process. In the first paragraph we give an answer on the main question. In the second paragraph we discuss the implications for the existing literature, that we used in the theoretical framework and in the literature review. In the third paragraph a reflection on the methodological process and the limitations of the method are given. In the fourth paragraph we discuss the managerial insights. The last paragraph includes recommendations for future research.

6.1 Answer on the main question

The present study is a first step in understanding and solving messy information problems. This understanding emerged by decomposing messy information problems into five main concepts. These five concepts formed the starting point of the constructed model in the present study. Based on a literature review and the validation of the model, the present study tries to give an answer on the following question: Which possibilities exist for solving messy information problems when using a combination of system dynamics and big data analytics? Two possibilities were found in the model. These two possibilities consisted of two structures of archetypes.

The first archetype, named fixes that fails, consists of a conceptual level of two almost the same balancing loops and two almost the same reinforcing loops. The two balancing loops in this structure showed the process of two important big data processes, namely data mining and deep learning. Both processes construct patterns that could be a structure for the dynamic complexity structure of the messy information problem. However, these patterns caused confidence issues when they are used directly, which increases the impact of faults, biases and ambiguity in data. Ambiguity has a negative effect on key variables for deep learning and data input for classification. The two reinforcing loops stand for the application of both expert and theoretical knowledge that solve questions that arise out of the ambiguity of data. These solved questions increase confidence in the found patterns. This increased confidence decreases the impact of biases, faults and ambiguity in the data. However, during the validation process several criteria were discovered for applying these loops. First, if the outcomes of the solution for the messy information problem could affect a project member in a negative way, they can apply expert or theoretical knowledge in a wrong way. Because the structure is fixes that fails, this situation can even cause more serious failures. Second, using both theoretical and expert knowledge should not be directly applied without studying the deviation, because they can contain valuable insights. Thus, a certain understanding of the data is essential. Third, deep learning should only be used in situations where no better methods are available. For example, data mining can be used to build patterns, because deep learning is more a black box. At last, the assumption is met that the data and algorithm are not completely reliable.

The second archetype is success to the successful. This archetype consists of two reinforcing loops. The first loop tries to examine the faults and biases in data by using sensitivity analysis, which discovers the effect of the ranges of the effects of parameters. This effect decreases the impact of faults and biases and ambiguity in data, which offers possibilities estimating parameters by data mining. The second loop tries to estimate parameters based on existing knowledge of the dynamic complex structure. Both structures were confirmed during the validation. However, they have several side effects. Using the model structure knowledge for parameters estimation causes a loss of possible confidence tests during the validation phase of the model. Furthermore, the loop with data for estimation only can be successful if the data has reached a certain quality and someone is able to access this. Thus, according to the archetype the quality of the data and quality of the structure knowledge will decide which loop is stronger.

To conclude, both combinations of system dynamics and big data analytics offer possibilities to solve messy information problems. However, it is limited by the consequences the project delivers to participants, by the people who work with expert or theoretical knowledge. Another important limitation is the data itself, not all data will suitable for the combinations. A good examination can help to prevent useless analytics. In addition, a good examination after the analytics can investigate the valuable insights, instead of rejecting them to early.

6.2 Scientific implications

As far as we are aware, there are hardly no articles available about assessing data that is used in system dynamics. Examining the reliability of the data is a process at the end of a system dynamics model building (Barlas, 1996). The results of the validation of this model indicate that examining the quality of the data beforehand could help to use other methods into the model building process, for example data mining or deep learning. However, Sivarajah et al (2016) explained it is difficult to access the data by the V of Value. Value stands for value that is hidden in the data. Sensitivity analysis in combination with the structure of the insights of system dynamics structure could help to access the value on a parameter level.

According to Sivarajah et al (2016), veracity is an important feature of big data. Veracity stands for the validity of the results of the big data tools. Najafabadi, et al. (2015) found that good frameworks for accessing representations of deep learning are missing. Elragal and Klischewki (2017) discussed that a theory should play a different role in the knowledge discovery process of big data analytics. This study indicates a possible way of integrating theoretical and expert knowledge into the big data analytics process to overcome one of the biggest issues in big data, namely the veracity of the results (Sivarajah, et al., 2016). This process is comparable with the process of triangulation that is found in case study research (Yin, 2014). Triangulation can be defined as the process of converging of data collected from different sources, to determine the consistency of a finding (Yin, 2014). Each messy information problem is in the basis a case study, because you examine one problem. Therefore, these insights can also be used in messy information problems. In a messy information problem theoretical knowledge and expert knowledge are converged with written and numerical knowledge from the large databases. Abdelbari and Shafi (2017) tried to apply deep learning in several cases to extract a look-a-like of causal loop diagram. However, their results were disappointing. From the three applications, only one application gave a result. The present study added some criteria to the process of deep learning to hopefully find reliable results. Further investigation of the results of the study of Abdelbari and Shafi (2017) is needed to explain the failure of two of their projects.

6.3 Methodological reflection

The present study used a literature review and a causal loop diagram for building and analyzing which possibilities of combinations of system dynamics and big data analytics offer a solution for solving messy information problems. The chosen research design was the best research design for the present study, because other research designs were unsuitable due to no cooperation of a company, no available high-quality cases and the explorative nature of the main question. The combination of the results from the different validation sources in the present study show quite the same results compared to each other. Only some additional side effect was found in some of the sources. In addition, the findings from the model came from existing scientific literature, this confirms too the external validity.

The research design has some limitations for the internal validity. In future some changes can be made within the design to increase the internal validity. An important limitation is that only three respondents participated in the validation interviews.

Saturation of validation information ca not be done, based on the present study. More respondents can confirm whether the found results are applicable. In addition, the system dynamics experts were both experts on eliciting information from people. This resulted in a deeper interview on expert and theoretical knowledge, but a relatively shallow interview on the parameter estimation. In future research system dynamics experts with other expertise should be interviewed.

Another limitation is that all the experts of validation have a social constructivism science perspective. This could result in biased defined definitions and insights, because they accept certain assumptions unnoticed. Especially in big data analytics a more positivist science paradigm is stronger, because a computer is binary, and it is not able in handling ambiguity that well.

Furthermore, we used the most important journals in the field of big data analytics and system dynamics for our literature review. Maybe more insights could be found in other journals. In addition, other journals could be used as a theoretical structure confirming test (Barlas, 1996).

Despite these limitations, the present study is one of the first studies that examine the possible combinations of big data analytics and system dynamics to solve messy information problems. The results of this study should be interpreted with some caution. However, it shows that some combinations of big data analytics and system dynamics are useful to solve messy information problems.

6.4 Managerial insights

The present study delivers useful managerial insights. First, the present study illustrates in the theoretical framework the limitations of the methods separately in comparison to messy problems. As Ulrich (2003) discusses that both soft system and hard system methodologies separately used have each their own weaknesses and strengths, but at the end they are separately less powerful compared to a combination of hard and soft methodologies. The present study clearly indicates that a combination of system dynamics and big data analytics can establish better insights into messy information problems. Thus, a combination of hard and soft methodology experts in one team is preferred over a single methodology expert in a team. In addition, the present study shows a better integration of the use of the different levels of data (Forrester, 1992) discusses, because each methodology is focused on one or two levels of data. System dynamics itself is more focused on mental data and on numerical data, compared to big data which is more focused on written and numerical data.

Second, the present study shows clearly the weaknesses of certain tools. However, explanations and use of other tools to overcome these weaknesses are also given. For example, deep learning delivers patterns, but is not able to decide if the pattern illustrates something valuable. Theoretical or expert knowledge solve this weakness by adding additional information into the final pattern to decide whether the pattern is valuable.

Third, within system dynamics in the special application of group model building literature, a lot has been written about consensus and commitment. Within this special application of system dynamics consensus and commitment is most of the time about defining the problem or the system (Vennix, 1996). In an application, where consensus is not reached easily, concept drift identification can help to support this process by identifying when concept drift happened to add these insights into the process. These insights lead to a better understanding about the structure and the change over time.

6.5 Recommendations for future research

Based on the results some future research directions can be identified. The first recommendation is a research recommendation. First, the validation process of confidence building process was small in the present study. Future research should extend the current number of experts with different expertise and scientific paradigms to increase the confidence building. In addition, more tests should be added to further increase the confidence in the model, for example using the possibilities of the model in experiments or analyzing of case studies. Moreover, there is a recommendation for practioners. The recommendation is to start a project which will apply the insights of the present study into practice and will write down all circumstances and experiences. This project could give insights in further specification of the criteria or side effects of the applications

The present study did not focus on the exact criteria for data in relation to the algorithms. Future research can focus on these criteria of data in relation to the available algorithms. New insights can specify the exact combinations of problems and algorithms for solving messy information problems.

Validity in the present study has been studied based on the interview data of the validation of the model. Based on the conclusions of the validation, more insights are needed to find out which tool or combinations of tools exclude certain validity. This information is missing in the present model, by adding this information the model will be more applicable.

At last, the research of Fiddaman (2017) showed promising results by adding agent-based modelling into the combination of system dynamics and big data analytics. Most databases contain local knowledge. By using agent-based modelling, possibilities arise for adding local information into the problem.

Overall, limited previous research have investigated the combinations of system dynamics and big data analytics to solve messy information problems. Future research is needed to gain better insights in these combinations

REFERENCES

- Abdelbari, H., Shafi, K. (2017). A computational intelligence-based Method to ‘learn’ causal loop diagram-like structures from observed data. *System Dynamics Review*, 33, 3-33.
- Aletti, G., Micheletti, A. (2017). A clustering algorithm for multivariate data streams with correlated components. *Journal of Big Data*, 4:48.
- Al-Hassan, A.A., Alshameri, F., Sibley, E.H. (2013). A research case study: Difficulties and recommendations when using a textual data mining tool. *Information and Management Journal*, 40(7), 540-552.
- Allen, P.M. (1988). Dynamic models of evolving systems. *System Dynamics Review*, 4(1-2), 109-130.
- Barlas, Y. (1996). Formal Aspects of model validity and validation in system dynamics. *System Dynamics Review*, 12(3), 183-210.
- Barlas, Y., Carpenter, S. (1990). Philosophical roots of model validation: two paradigms. *System Dynamics Review*, 6(2), 148-166.
- Bleijenbergh, I.L. , Van Engen, M.L. , Rouwette, E.A.J.A. (2013). Validity and utility of participatory modeling: analyzing impediments to women’s careers at Dutch universities. *Under review at Organizational Research Methods*. [Paper is available via Blackboard, please do not spread.]
- Bryson, J. M. 2004. What to do when stakeholders matter. Stakeholder identification and analysis techniques. *Public Management Review*, 6(1), 21-53.
- Callado, A., Kelner, J., Sado, D., Kamienski, C.A., Fernandes, S. (2010). Better network traffic identification trough the independent combination of techniques. *Journal of Network and computer applications*, 33 (4), 433-446.
- Campbell, D. (2001). The long and winding (and frequently bumpy) road to successful client engagement: one team's journey. *System Dynamics Review*, 17(3), 195-215.
- Chandak, M.B. (2016). Role of big-data in classification and novel class detection in data streams. *Journal of Big Data*, 3:5.
- Chika Nwagwu, H., Okereke, G., Nwobodo, C. (2017). Mining and visualizing contradictory data. *Journal of Big Data*, 4:36.
- Clemson, B., Tang, Y., Pyne, J., Unal, R. (1995). Efficient methods for sensitivity analysis. *System Dynamics Review*, 11(1), 31-49.

- Conboy, K., Dennehy, D., O'connor, M. (2018). 'Big time': An examination of temporal complexity and business value in analytics. *Information and Management Journal*, In Press, Corrected Proof.
- Doyle, J.K., Ford, D.N. (1998). Mental models' concepts for system dynamics research. *System Dynamics Review*, 14, 3-29.
- Duggan, J. (2008). Equation-based policy optimization for agent-oriented system dynamics models. *System Dynamics Review*, 24, 97-118.
- Eberlein, R.L. (1989). Simplification and understanding of models. *System Dynamics Review*, 5(1), 51-68.
- Eker, S., Slinger, J., Daalen van, E., Yücel, G. (2014). Sensitivity analysis of graphical functions. *System Dynamics Review,* 30, 186-205.
- El Alaoui, I., Gahi, Y., Messoussi, R., Chaabi, Y., Todoskoff, A., Kobi, A. (2018). A novel adaptable approach for sentiment analysis on big social data. *Journal of Big Data,* 5:12
- Elragal, A., Klischewski, R. (2017). Theory-driven or process-driven prediction? Epistemological challenges of big data analytics. *Journal of Big Data*, 4:19.
- Enserink, B., Koppejan, J.F.M., Mayer, I.S. (2012). Public Policy Analysis – New Development. Chapter 2. *International series in operations Research & Management Science,* 179, 11-40.
- Fang, X., Zhan, J. (2015). Sentiment Analysis using product review data. *Journal of Big Data,* 2:5.
- Featherston, C.R. and Doolan, M. 2012. A Critical Review of the Criticisms of System Dynamics. *The 30th International Conference of the System Dynamics Society*, Retrieved from, (https://www.systemdynamics.org/conferences/2012/proceed/papers/P1228.pdf)
- Felipe Luna-Reyes, L., Lines Anderson, D. (2003). Collecting and analyzing qualitative data for system dynamics: methods and models. *System Dynamics Review*, 19, 271-296.
- Fiddaman T. (2017). *A tale of big data analytics and system dynamics*. Accessed on the 5th April 2018 from http://metasd.com/2017/08/a-tale-of-big-data-and-system-dynamics/
- Ford, A., Flynn, H. (2005). Statistical screening of system dynamics models. *System Dynamics Review*, 21, 273-303.
- Ford, D.N. (1999). A behavioral approach to feedback loop dominance analysis. *System Dynamics Review*, 15, 3-36.

- Ford, D.N., Sterman, J.D. (1998). Expert knowledge elicitation to improve formal and mental models. *System Dynamics Review*, 14, 309-340.
- Forrester, J. 1992. Policies, decisions, and information sources for modeling. *European Journal of Operational Research*, 59, 42-63.
- Forrester, J.W. (1987). Lessons from system dynamics modeling. *System Dynamics Review*, 3(2), 138-149.
- Forrester, J.W. and Senge, P.M. (1980): Tests for Building Confidence in System Dynamics Models. In System Dynamics. *TIMS Studies in the Management Sciences*, Vol. 14, 209–228.
- Gandomi, A., & Haider, M. 2015. Beyond the hype: Big data concepts, methods, and analytics. *International Journal of Information Management*, 35(2), 137–144
- Gigerenzer, G., Todd, P.M. and the research group (1999). *Simple Heuristics that make us smart.* Oxford: University Press Oxford.
- Goodwin, P., & G. Wright (2014). *Decision Analysis for Management Judgment. Fifth Edition.* Chichester: Wiley and Sons.
- Groesser, S.N., Schwaninger, M. (2012). Contributions to model validation: hierarchy, process, and cessation. *System Dynamics Review*, 28, 157-181.
- Hall, R.I., Aitchison, P.W., Kocay, W.L. (1994). Causal policy maps of managers: formal methods for elicitation and analysis. *System Dynamics Review*, 10(4), 337-360.
- Hayward, J., Boswell, G.P (2014). Model Behavior and the concept of loop impact: a practical method. *System Dynamics Review*, 30, 29-57.
- Hekimoglu, M., Barlas, Y. (2016). Sensitivity analysis for models with multiple behavior modes: a method based on behavior pattern measures. System Dynamics Review 32, 332-362.
- Herland, M., Khoshgoftaar, T.M., Wald, R. (2014). A review of data mining using big data in health informatics. *Journal of Big Data*, 1:2.
- Homer, J. (1996): Why We Iterate – Scientific Modeling in Theory and Practice, *System Dynamics Review,* 12(1), 1–19.
- Hossenichimeh, N., Rahmandad, H., Jalali, M.S., Wittenborn, A.K. (2016). Estimating the parameters of system dynamics models using indirect inference. *System Dynamics Review*, 32, 156-180.
- Hurtado, J.L., Agarwal, A., Zhu, X. (2016). Topic discovery and future trend forecasting for texts. *Journal of Big Data*, 3:7

- Jacobsen, C., Bronson, R. (1987). Defining sociological concepts as variables for system dynamics modeling. *System Dynamics Review*, 3(1), 1-7.
- Janssen, M., H. v.d. Voort, A. Wahyudi. Factors influencing big data decision-making. *Journal of Business Research* 70, 338-345.
- Joseph, R. C., & Johnson, N. A. 2013. Big data and transformational government. *IT Professional*, 15(6), 43–48.
- Kaplan, S. (2008). Framing Contests: Strategy Making Under Uncertainty. *Organization Science*, 19, 729–752.
- Kaur, A., Datta, A. (2015). A novel algorithm for fast and scalable subspace clustering of high-dimensional data. *Journal of Big Data*, 2:17.
- Kerem Saysel, A., Barlas, Y. (2006). Model simplification and validation with indirect structure validity tests. *System Dynamics Review*, 22, 241-262.
- Kim, H., Andersen, D.F. (2012). Building Confidence in causal maps generated from purposive text data: mapping transcripts of the federal reserve. *System Dynamics Review,* 28, 311-328.
- Kreps, D.M., Wilson, R. (1981). Reputation and imperfect information. *Journal of Economic theory,* 27, 253-279.
- Kumar, S., Toshniwal, D. (2015). A data mining framework to analyze road accident data. *Journal of Big Data*, 2:26
- Lamari, Y., Chah Slaoui, S. (2017). Clustering categorical data based on the relational analysis approach and MapReduce*. Journal of Big Data*, 4:28.
- Lee, I. 2017. Big Data: Dimensions, evolution, impacts and challenges. *Business Horizons*, 60, 293-303.
- Lines Anderson, D., Felipe Luna-Reyes, L., Diker, V.G., Black, L., Rich, E., Andersen, D.F. (2012). The disconformity interview as a strategy for the assessment of system dynamics models. *System Dynamics Review,* 28, 255-275.
- Liu, X., Wang, X., Matwin, S., Japkowicz, N. (2015). Meta-MapReduce for scalable data mining. *Journal of Big Data*, 2:14.
- March, J.G (1987). Ambiguity and Accounting: The illusive link between information and decision making. *Accounting, Organizations and Society*, 12 (2), 153-168.
- Martinez-Mayano, I.J. and Richardson, G.P. 2013: Best Practices in System Dynamics Modeling. *System Dynamics Review*, 29(2), 102–123.
- Mashayekhi, A.N. Ghili, S. (2012). System dynamics problem definition as an evolutionary process using the concept of ambiguity. System Dynamics Review 28, 182-198.

- Meadows, Donella H.: *Thinking in Systems--A Primer*, 2008. White River Junction: Chelsea Green Publishing.
- Mosekilde, E., Aracil, J., Allen, P.M. (1988). Instabilities and chaos in nonlinear dynamic systems. *System Dynamics Review*, 4 (1-2), 14-55.
- Mujamdar, J., Naraseeyappa, S., Ankalaki, S. (2017). Analysis of agriculture data using data mining techniques: application of big data. *Journal of Big Data*, 4:20.
- Najafabadi, M.M., Villanustre, F., Khoshgoftaar, T.M., Seliya, N., Wald, R., Muharemagic, E. (2015). Deep learning applications and challenges in big data analytics. *Journal of Big Data*, 2:1
- Osborne, M.J., Rubinstein, A. (1994). *A Course in Game Theory*. Chapter 2. Cambridge: The MIT Press.
- Pröllochs, N., Feuerriegel, S. (2018). Business Analytics for strategic management: Identifying and assessing corporate challenges via topic modelling*. Information and Management Journal*, In Press, Corrected Proof.
- Prusa, J.D., Khoshgoftaar, T.M. (2017). Improving deep neural network design with new text data representations. *Journal of Big Data,* 4:7.
- Rahn, R.J. (1985). Aggregation in system Dynamics. *System Dynamics Review,* 1(1), 111-122.
- Randolph J.J. (2009). A guide to writing the dissertation literature Review. *Practical Assesment, Research & Evaluation* ,14(13).
- Richardson, G.P. (2011). Reflections on the foundations of system dynamics. System Dynamics Review 27, 219-243.
- Schoenenberg, L, Schmid, A., Ansah, J., Schwaninger, M. (2017). The challenge of model complexity: improving interpretation of large causal model's trough variety filters. *System Dynamics Review,* 33, 112-137
- Schoenenberger, L., Schmid, A., Schwaninger, M. (2015). Towards the algorithmic detection of archetypal structures in system dynamics. *System Dynamics Review*, 31, 66-85.
- Schwaninger, M. & Groesser, S. N. 2008. Model-Based Theory-Building with System Dynamics. *Systems Research and Behavioral Science*, in press.
- Shafir, E., Simonson, I. and Tversky, A. (1993). Reason based choice. *Cognition*, 49, 11-36.
- Sheng Ng, T., Lee Sy, C., Hay Lee, L. (2012). Robust parameter design for system dynamics models: a formal approach based on goal seeking behavior. *System Dynamics Review*, 28, 230-254.

- Sivarajah, U., Kamal, M., Irani, Z., Weerakkody, V. 2016. Critical Analysis of Big Data challenges and analytical methods. *Journal of Business Research*, 70, 263-286.
- Social research association 2003. *Ethical Guidelines*. Accessed on the 13th of December 2017, from http://the-sra.org.uk/wp-content/uploads/ethics03.pdf
- Sohangir, S., Wang, D., Pomeranets, A., Khoshgoftaar, T.M. (2018). Big Data: Deep learning for financial sentiment analysis. *Journal of Big Data*, 5:3.
- Sterman, J. D. 2000. *Business dynamics. Systems thinking and modeling for a complex world*. Boston, MA: McGraw-Hill.
- Sterman, J.D. (1988). Deterministic chaos in models of human behavior: methodological issues and experimental results. *System Dynamics Review*, 4(1-2), 148-172.
- Taylor, T. R. B., Ford, D. N., & Ford, A. 2010. Improving model understanding using statistical screening. *System Dynamics Review*, 26, 73-87
- Taylor, T.R.B., Ford, D.N., Ford, A. (2010). Improving model understanding using statistical screening. *System Dynamics Revi*ew, 26, 73-87.
- Tilaye Wubetie, H. (2017). Missing data management and statistical measurement of social-economic status: application of big data. *Journal of Big Data*, 4:47.
- Tversky, A., Kahneman, D. (1974). Judgement under uncertainty: Heuristics and biases. *Science* 185, 1124-1131.
- Ulrich, W. (2003). Methodology choice: Critical System Thinking as critically systemic discourse. *The Journal of the Operational research society*. Vol 54, (4), 325-342.
- Vennix, J. A. M. 1996. *Group model building. Facilitating team learning using system dynamics*. Chichester: Wiley & Sons
- 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.
- Walrave, B (2016). Determining intervention thresholds that change output behavior patterns. *System Dynamics Review,* 32, 261-278.
- Wang, H., Z. Xu, H. Fujita, S. Liu. Towards felicitous decision making: An overview on challenges and trends of Big Data. Information Sciences, 367(8), 747-765
- Yang, J. (2018). Game-theoretic modeling of players, ambiguities on external factors. *Journal of mathematical Economics*, 52, 31-56.

- Yang, J., Yecies, B. (2016). Mining Chinese social media UGC: a big-data framework for analyzing Douban movie reviews. *Journal of Big Data*, 3:3
- Yang, Y., Zhang, K., Wang, J., Vinh Nguyen, Q. (2015). Cabinet Tree: an orthogonal enclosure approach to visualizing and exploring big data. *Journal of Big Data*, 2:15.

- Yin, R.K (2014). *Case Study Research. Design and Methods*. London: Sage Publications.
- Yücel, G., Barlas, Y. (2011). Automated parameter specification in dynamic feedback models based on behavior pattern features. *System Dynamics Review* 27, 195-215.
- Zang, W., Zhang, P, Zhou, C., Guo, L. (2014). Comparative study between incremental and ensemble learning on data streams: Case study. *Journal of Big Data*, 1:5
- Zhan, J., Dahal, B. (2017). Using deep learning for short text understanding. *Journal of Big Data*, 4:34.

Appendix A: Explanation of causal loops

System Dynamics is an unfamiliar method for most people. One type of model in system dynamics is a causal loop diagram. In this appendix we will focus on causal loop diagrams. Vennix (p 32., 1996) explains causal loop diagrams as: “A causal diagram takes the form of variables which are linked by arrows. The variable at the tail of the arrow is considered to have a causal effect on the variables at the point. Arrows are further denoted by a ‘+’ or ‘-‘sign. A ‘+’ sign represents a so-called positive causal relationship, which indicates that both variables change in the same direction (i.e. both increase or decrease). On the other hand, a ‘-‘sign denotes a negative relationship, which indicates that both variables change in opposite directions.”.

Furthermore, these relationships can form loops, loops are variables that are connected to each other with arrows in the same direction. Figure 1 shows two loops. The loop on the left is a loop with only positive signs, also called a reinforcing loop. Logically, a reinforcing becomes stronger. On the right a balancing loop is displayed. A balancing loop is a goal seeking loop and is characterized by an uneven number of negative signs. Loops together can form certain system structures, so-called system archetypes. These archetypes show similar behavior (Meadows, 2008).

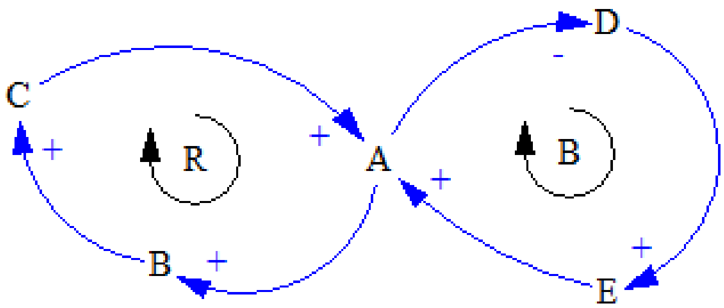

Figure 1: Balancing and reinforcing loop

- Meadows, Donella H.: *Thinking in Systems--A Primer*, 2008. White River Junction: Chelsea Green Publishing.
- Vennix, J. A. M. 1996. *Group model building. Facilitating team learning using system dynamics*. Chichester: Wiley & Sons