Decision Tree Analysis for Estimating the Costs and Benefits of Disclosing Data

Luthfi, Ahmad; Janssen, Marijn; Crompvoets, Joep

DOI
10.1007/978-3-030-29374-1_17

Publication date
2019

Document Version
Accepted author manuscript

Published in
Digital Transformation for a Sustainable Society in the 21st Century - 18th IFIP WG 6.11 Conference on e-Business, e-Services, and e-Society, I3E 2019, Proceedings

Citation (APA)
Luthfi, A., Janssen, M., & Crompvoets, J. (2019). Decision Tree Analysis for Estimating the Costs and Benefits of Disclosing Data. In I. O. Pappas, I. O. Pappas, J. Krogstie, L. Jaccheri, P. Mikalef, Y. K. Dwivedi, & M. Mäntymäki (Eds.), *Digital Transformation for a Sustainable Society in the 21st Century - 18th IFIP WG 6.11 Conference on e-Business, e-Services, and e-Society, I3E 2019, Proceedings* (pp. 205-217). (Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics); Vol. 11701 LNCS). Springer. https://doi.org/10.1007/978-3-030-29374-1_17

Decision Tree Analysis for Estimating the Costs and Benefits of Disclosing Data

Ahmad Luthfi[1,2 *[0000-0001-5416-1529]], Marijn Janssen[1[0000-0001-6211-8790]], Joep Crompvoets[3[0000-0003-1077-597X]]

[1]Delft University of Technology
Faculty of Technology, Policy and Management
Jaffalaan 5, 2628 BX Delft, The Netherlands
{a.luthfi, m.f.w.h.a.janssen}@tudelft.nl
[2] Universitas Islam Indonesia, Yogyakarta, Indonesia
ahmad.luthfi@uii.ac.id
[3] Katholieke Universiteit Leuven, Leuven, Belgium
joep.crompvoets@kuleuven.be

Abstract. The public expects government institutions to open their data to enable society to reap the benefits of these data. However, governments are often reluctant to disclose their data due to possible disadvantages. These disadvantages, at the same time, can be circumstances by processing the data before disclosing. Investments are needed to be able to pre-process a dataset. Hence, a trade-off between the benefits and cost of opening data needs to be made. Decisions to disclose are often made based on binary options like "open" or "closed" the data, whereas also parts of a dataset can be opened or only pre-processed data. The objective of this study is to develop a decision tree analysis in open data (DTOD) to estimate the costs and benefits of disclosing data using a DTA approach. Experts' judgment is used to quantify the pay-offs of possible consequences of the costs and benefits and to estimate the chance of occurrence. The result shows that for non-trivial decisions the DTOD helps, as it allows the creation of decision structures to show alternatives ways of opening data and the benefits and disadvantages of each alternative.

Keywords: Decision Tree Analysis, Estimation, Costs, Benefits, Open data, Open government, Investments

1 Introduction

During the past decade, government institutions in many countries have been started to disclose their data to the public. The society expects that governments become open and that their becomes easy to re-use [1, 2]. The opening of the data by the governments can provide various opportunities including increased transparency, accountability but also to improve decision-making and innovation [3, 4]. However, opening of data is more cumbersome and many datasets remain closed as they many contain personal or sensitive data. Decisions to disclose are often made based on bina-

ry options like "open" or "closed" the data, whereas also parts of a dataset can be opened or datasets can be pre-processed in such a way that they can be opened data. A Decision tree analysis (DTA) can help decision-makers in estimating the investments needed to process data before releasing.

There are several advantages in using DTA to the decision-making problems. First, DTA can make the proses understandable and a method that relatively easy to interpret [5, 6]. Second, DTA is able to take into account both continuous and categorical decision variables [5, 7]. Third, DTA provides insight into which variables are the most important to comprehend the outcome of the alternative decisions [8]. Fourth, a decision tree can perform a classification without requiring in-depth knowledge in computational algorithm [5, 9].

The objective of this paper is to develop a decision tree analysis for open data (DTAOD) to estimate the costs and benefits of disclosing data. This will help us to gain insight into the potential of using DTA for supporting the opening of data. A decision tree is a decision support tool that uses a tree-like model of decisions and their possible consequences of conditional control statements [7, 10]. DTA is chosen as it can serve a number of purposes when complex problems in the decision-making process of disclosing data are encountered. Many complex problems in decision-making might be represented in the payoff table form [9]. Nevertheless, for the complicated problem related to investment decisions, decision tree analysis is very useful to show the routes and alternatives of the possible outcomes [7].

The developed DTA consists of the following four steps [5, 8], as follows: First, define a clear decision problem to narrow down the scope of the objective. Factors relevant to alternative solutions should be determined. Second, structure the decision variables into a decision-tree model. Third, assign payoffs for each possible combination alternatives and states. In this step, payoffs estimation is required to represent a specific currency of amount based on the experts' judgment. Fourth, provide a recommendation of decisions for the decision-makers.

This study used three alternative decisions when deciding to disclose data, namely "open the whole dataset", "provide limited access to dataset", and "keep the dataset closed". Opening decision refers to the governments enable releasing their data without any restriction. Limited access means the level of openness data is restricted to a specific group of users, and closed decision refers to the government should not share the data. Each decision has costs and benefits that need to be quantified. In addition, payoffs need to estimate the probability of investments. For example, the open decision might require potential investment for data collection, processing and data visualization.

This research can support decision-makers and other related stakeholders like business enablers and researchers, to create a better understanding of the problem structure and variants of opening data. Furthermore, this study contributes to the limited literature about decision support for disclosing data and it is the first work using DTA. This paper is consists of five sections. In Section 1 the rationale behind this research is described, Section 2 contains the related work of decision-making approaches to open data domain. In Section 3, the DTA approach is presented, including research method, related theories, and proposed steps in constructing DTA. Section 4 provides

systematically the development of DTA. Finally, the paper will be concluded in Section 5.

2 Related Work

2.1 Overview of Methods for Deciding to Open data

In the literature, there are various methods in analyzing to open data. Four types of approaches for decision-making of opening data were identified. First, an iterative decision-making process in open data using Bayesian-belief networks approach. Second, proposed guidance to trade-off the chances of value and risk effects in opening data. Third, a framework to weight the risks and benefits based on the open data ecosystem elements. Fourth, a fuzzy multi-criteria decision making (FMCDM) method to analyze the potential risks and benefits of opening data. The several related methods in analyzing to disclose data can be seen in Table 1.

Table 1. The overview in the literature

	Method	Overview and limitations
1	Iterative model of decision support for opening data [11, 12]	The use of Bayesian-belief networks approach is to construct the relational model of decision support in opening data. The outcomes of this model can be used to prevent the risks and still gain benefits of opening data. Several suppression approaches are used like removing sensitive attributes and designing k-anonymity of a dataset. This method, furthermore, also introduces binary decisions, namely: open and closed the dataset.
2	Trade-offs model [11, 13]	This method provides guidance for weighing the potential values and risks of opening data. Interview sections are based on some certain groups of government employees like civil servants and archivists. There is no specific methods nor algorithm found to develop the trade-off model. The trade-offs model can only be used for decision-making in a sense of the simple problem. The decision alternatives are defined in binary expression, namely: open and closed.
3	A framework of decision support in open data [14, 15]	A developed prototype is based on the following concept of open data ecosystems. The proposed model is exclusively for business and private organizations. There is no evaluation and assessment model introduced in this framework.

4	A fuzzy multi-criteria decision making (FMCDM) [16, 17]	As a practical methodology for dealing with fuzziness and uncertainty in Multiple Criteria Decision-making (MCDM), Fuzzy AHP (FAHP) has been applied to a wide range of applications. Fuzzy AHP has been implemented to a broader domain of studies. Fuzzy analytical hierarchy process (FAHP) is utilized by collecting input from experts' knowledge and expertise. The scores for each criterion are summed up to rank the importance of the decision alternatives. Four main criteria are used like data sensitivity and data ownership representing risks criteria, and data availability and data trustworthy as benefits criteria.

However, none of these related existing approaches uses a method to analyze and estimate the possible costs-benefits of opening data for a specific problem. DTA can play a role in providing different steps and expectations of the decision-making process.

2.2 Theory of decision tree analysis

DTA is introduced for the first time in the nineteen sixties and primarily used in the data mining domain. The main role of using this method is to establish classification systems based on multiple covariates in developing a prediction of alternative variables [5, 9]. A decision tree is a graphical representation of specific decisions with possible consequences of a series of related decision alternatives. This theory allows an individual or organizations to trade-off possible actions against another action based on the probabilities of risks, benefits, and costs of a decision-making process [5, 18]. DTA uses a tree-like model of decisions to drive informal discussion a method or algorithm that able to predict the alternative decisions into the numerical form [10]. In the case of opening data, DTA is used to identify and calculate the value of possible decision alternatives by taking into account the potential cost-adverse effects. In general, a decision tree consists of three main elements, namely [5, 7-9]:

1. **Decision node**. A decision is represented by a square that refers to the decision to implement or not to implement a project of work including the amount of money invested. In this study, a decision node represents the alternative in deciding whether the data should be opened, preprocessing before data can be opened, or keep the data closed.
2. **Chance node**. An uncertainty situation or event is indicated by a circle. The uncertain situation may happen before or after the event or both before and after decisions. In this paper, the chance node represents the potential costs and benefits of dataset status. For example, costs and benefits of opening data may lead to uncertainty situation like the possible costs of data collection and data visualization.

3. **Decision outcomes**. The outcome of a decision process is showed with a triangle. When all the uncertainties and change node have been completed and there are no more decisions to be changed and made, at this stage the decision-makers are able to know the estimation payoff they will need to take into account. In this study, the outcome indicates that the amount of money (Euros) and investments needed when the decision-makers define a decision of opening the data.

The existing literature provides insight into the advantages of using DTA the decision-making process. First, DTA can generate understandable the estimation process and is easy to interpret [5, 6]. Second, DTA is able to take into account both continuous and categorical decision variables [5, 7]. Third, DTA provides a clear indication of which variable is becoming the most important in predicting the outcome of the alternative decisions [8]. Fourth, a decision tree can perform a classification without requiring in-depth knowledge in computational [5, 9].

The use of DTA in this study can manage a number of variables of the costs and benefits in opening data. In this situation, DTA can support the decision-makers in deciding how to select the most applicable decision. Besides, this method is able to subdivide heavily skewed variable into a specific amount of ranges. Fig.1 shows the example of decision tree notation with alternatives of choices in the case of open data decision.

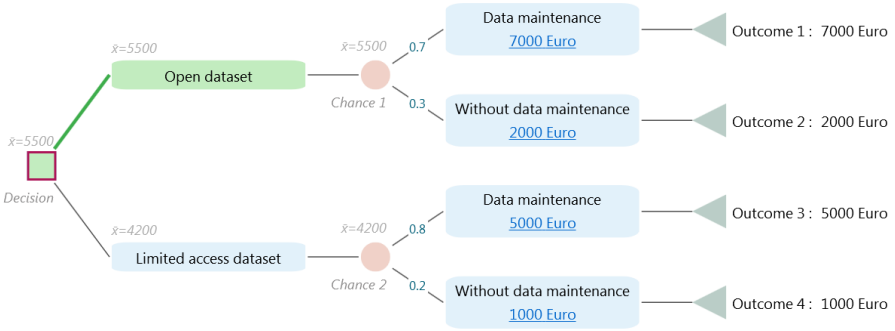


Fig. 1. An example of DTA

The objective of this decision tree illustrated in Fig.1 is that the decision-makers are trying to find the expected monetary value (EMV) of probability decisions, namely open dataset and limited access to the dataset. The EMV is the probability-weighted average of the outcomes [5, 7]. The use of EMV in DTA can be defined in two main benefits. First, EMV helps decision-makers to understand the possible investments of alternative actions. Second, DTA supports selecting the most appropriate alternatives by weighing the costs of two alternative decisions.

In order to get the probability of an outcome in opening data case shows in Fig.1, the probabilities along the branches of the tree need to be multiplied. Beforehand, we first should define that there are two alternative decisions in this case, namely: open the dataset or provide limited access to the dataset. Heavily skewed variable need to

be subdivided into a specific amount of ranges. In this example, the ranges of the possible costs are between 0 to 1000 Euros. To obtain the expected monetary value from the example in Fig.1**,** the probability-weighted average of the four outcomes is calculated by summing the data maintenance activity with the probability of each outcome. This, give the outcome 0.7 x 7000 + 0.3 x 2000 = 5500 Euro. In a similar vein, the costs of the limited access alternative can be calculated 0.8 x 5000 + 0.2 x 0.2 x 1000 = 4200 Euro. In this example, the DTA shows that the investment needed to open a dataset is higher than the limited access to alternative decisions.

3 Research Approach

In this study, we use experts' judgment to assign payoffs possible consequences of the costs and benefits in opening data including the changes. The expert judgment is used because of their capability to interpret and integrate the existing complex problems in a domain of knowledge [19, 20]. To do so, we interviewed four experts from three postgraduate researchers and one professional with open government data and costs-benefits investment experiences consideration. The objective of this approach is to provide comprehensive insight in predicting the costs and benefits in the open data field. In order to minimize the potential bias of the payoffs process, a systematic interview protocol for employing expert judgment was developed. There are some considerations in selecting the experts for this study. First, we select the experts based on their knowledge in the open data field. Second, best practices in estimating the costs and benefits investment in open data domain should take into account. The data collection using experts judgment followed three main steps [20, 21]:

1. **Define the background and preparation**. This step involves the identification and justification of the costs and benefits for which experts' assessment is required. To prevent possible bias during the quantification process, the interview protocol should in line with the experts' knowledge and their best practice in the field of open data. In addition, we determine some requirements of criteria for the experts like tangible evidence of expertise and their reputation and availability to participate in the entire process of interview sections.
2. **Structuring and decomposing tasks**. In order to calibrate the elicitation problems, the interview process should easy to understand and able to transfer the feedback from the experts into the pay-off table. In some cases, the experts might have their own opinion in the sense that they can contribute to the understanding and personal beliefs.
3. **Elicitation process**. In this step, an iterative process involves the measurement of the expert's quantification and distribution should take into account. The experts should assess the adequacy of the probability distribution, and provide a repeating process to adjust the results of the elicitation process.

The selected experts use their understanding and reasoning processes as they refer to their experiences to make judgments [21, 22]. However, understanding the current issues and having logical reasons behind predicting costs and benefits in open data

domain is not trivial. The costs and benefits estimation requires sufficient knowledge and complex experiences in a specific field [23]. There are some barriers and limitations of the expert judgments elicitation. First, during the elicitation process, the experts might possibly quantify the answers inconsistently because of the unclear set of questions from the interviewer. To cover this issue, we design a list of questions protocol as structured as possible and easy to comprehend by the experts. The use of specific terminologies in the field of open data, for instance, should be clearly defined. Second, the use of experts' judgment is potentially time-consuming and experts are often overconfidence that can lead to uncertainty estimation [19, 24]. To tackle this issue, we use aggregate quantitative review by subdividing heavily skewed variable into a specific amount of ranges.

3.1 Steps in Developing the DTA

In order to effectively manage and construct a decision tree based analysis, and to represent a schematic and structured way, in this paper we use four main steps in developing DTA [5-7], as follows: First, define a clear problem to narrow down the scope of the DTA. Relevant factors resulting in alternative solutions should be determined as well. This step could involve both internal and external stakeholders to seek the possible options for a better decision-making process.

Second, define the structure the decision variables and alternatives. The structure of the problems and influence diagram require to be interpreted into formal hierarchical modeling. In this step, organizations need to construct decision problems into tree-like diagrams and identify several possible paths of action and alternatives.

Third, assign payoffs and possible consequences. In this step, the EMV formula is required to help to quantify and compare the costs and benefits. EMV is a quantitative approach to rely on the specific numbers and quantities to perform the estimation and calculations instead of using high-level approximation methods like agree, somewhat agree, and disagree options. For this, experts' judgment is used to estimate the pay-off of possible consequences of the costs and benefits and to estimate the chance of occurrence.

Fourth, provide alternative decisions and recommendations. After successfully assigning payoffs the possible consequences and considering adjustments for both costs and benefits, decision-makers can select the most appropriate decision that meets the success criteria and fit with their budget. These steps will be followed when developed the DTAOD.

4 Developing the DTAOD: step-by-step

4.1 Step 1: Define the Problems

The problem of opening data consists of three main aspects. First, decision-makers have a lack of knowledge and understanding in estimating the costs and benefits of open data domain and its consequences. Second, decision-makers might consider how to decide on the opening of data. Too much data might remain closed due to a lack of

knowledge of alternatives. Yet, the decision-making process in the open data field remains a complex circumstance and fraught with ambiguities. Third, decision-makers have no means to estimate the potential costs and benefits of opening data. Therefore, decisions are often made based on binary options like “open” and ‘closed” the data.

The three main aspects mentioned above tend to make the decision-makers are often reluctant to open their data. In this study, we develop a structured approach to estimate the possible costs and benefits of opening data. Considering the debatable situation of decision alternatives, the DTAOD method provides insight and applicable decisions in open data domain like allowing the data opened, giving some suppression methods, and keeping the data closed. By defining these problems, we expect that this study can deliver new insights into the inquiries and contribute to the governments and decision-makers in estimating the possible costs and benefits of disclosing data.

4.2 Step 2: Structure the Decision Alternatives

The decision-making process in opening data can be time-consuming and might require many resources. To understand better the consequences of each possible outcome, decision-makers require simplifying the complex and strategic challenges. Therefore, the DTA presented in this paper can construct a model and structure the decision alternatives whether the data should be released or closed. In this step, the list of decision alternatives and possible paths should be modeled into structure manner. The possible decision alternatives are a finite number of possible future events and represent uncertainties through the probabilities and possible distributions.

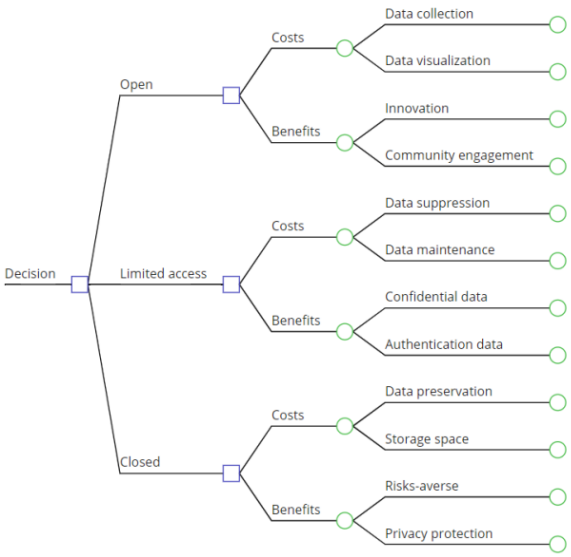


Fig. 2. Decision alternatives and possible paths

Fig.2 illustrates the decision alternatives and various types of possible paths in deciding the complex problems of opening data. We define three main decision nodes, namely “open”, “limited access”, and “closed”. The first decision refers to the governments allow releasing their data to the public with less or without restrictions. Second, the limited access indicates that the level of openness is restricted to a specific group of user. Third, closed decision refers to the government should keep the data exclusively. For each decision nodes, they have costs and benefits factors that require quantifying and payoffs to estimate the probability of investment. For example, the open decision might potential costs-adverse from the activity of data collection and data visualization. This means the government should consider by opening the data they will invest the amount of money to cover these activities. At the same time, disclosing the data without restriction can contribute some benefits to society like creating innovation and improving community engagement. The similar situation with two other decision alternatives like limited access and closed. Each decision has its costs and benefits that require the calculation process. The detailed explanation related to costs and benefits classification and the payoff process can be seen in Section 4.3.

4.3 Step 3: Assign Payoffs and Possible Consequences

In this step, the assign numerical values to the probabilities including the action-taking place, and the investment value expected as the outcome will be carried out. The objective of the assign payoff mechanism is to weight the trade-off objectives, and the potential risks-adverse [5]. In this paper, the assign payoffs represent the outcome for each combination in a table namely table of payoffs and possible consequences. This table uses costs terminology that represents the negative impact of a decision like value for the expense and potential lost revenue [5, 8]. While benefits-averse, indicate the positive influence to a decision like a net revenue stream, potential income, and other profit elements [8, 9].

Implementing the entire process of assign payoffs and its possible consequences is a complex situation. Although the decision tree approach lends a structure and a methodology to evaluate choices, the governments might invest many resources and consume time earnestly. Besides, the way to determine numerical values for the potential costs and benefits factor remains intuitively. The mistakes in estimating numerical values, mistaken assumptions, and the difficulties in quantifying the expected monetary values are able to reverse the estimation process significantly.

To create simpler the translation mechanism from the qualitative to the quantitative estimation, first, we use the minimum and maximum amount of money that may affect the costs and benefits factors in decision alternatives. Second, we develop an interview protocol by using clear and common terminologies in open data domain including the influenced sub-factors of the costs and benefits. The result of the assign payoffs and the possible consequences from the selected experts as presented in Table 2.

Table 2. Assign payoffs and possible consequences of the costs and benefits in opening data

Alternative Decisions	Expert judgment (probability in percentage)					Expert judgment (invesment in Euro)						
	1	**2**	**3**	**4**	**Mean**	**1**	**2**	**3**	**4**	**Mean**	**Total**	**Outcome**
1. Open												
- Costs factors												
a. Data collection	65	67	58	62	63	15.500	16.200	16.500	16.600	16.200	30.238	46.438
b. Data visualization	35	33	42	38	37	14.250	15.500	12.100	14.300	14.038		44.276
- Benefits factors												
c. New knowledge	58	62	54	63	59	12.300	14.450	14.000	13.000	13.438	26.796	40.234
d. Community engagement	42	38	46	37	41	15.235	11.600	13.800	12.800	13.539		40.335
2. Limited Access												
- Costs factors												
e. Data supression	66	58	54	55	58	16.000	16.500	17.000	14.500	16.000	32.725	48.275
f. Data maintenance	34	42	46	45	42	16.000	17.000	16.800	17.100	16.725		49.450
- Benefits factors												
g. Confidential data	55	65	44	45	52	18.000	17.600	17.700	18.200	17.875	35.000	52.875
h. Authentication data	45	35	56	55	48	18.500	17.500	16.850	16.500	17.338		52.338
3. Closed												
- Costs factors												
i. Data preservation	72	68	62	70	68	13.000	14.500	13.500	14.200	13.200	27.588	40.788
j. Storage space	28	32	38	30	32	16.000	15.850	12.200	13.500	14.388		41.976
- Benefits factors												
k. Risks-averse	52	56	57	60	56	9.300	10.500	12.000	10.000	10.450	22.513	32.963
l. Privacy protection	48	44	43	40	44	11.000	13.000	11.750	12.500	12.063		34.576

Table 2 presents the result of the assign payoffs between three alternative decisions, namely: “open”, “limited access”, and “closed”. This table includes the expert judgment in estimating the probabilities of the costs and benefits, and the numerical values given to predict the investment of money in the euro currency. When the entire process of assign payoffs has completed, we can calculate the average numerical values of the costs and benefits percentages possibilities. For example, data collection factor might probability invests 63% of the revenue stream instead of a data visualization program (37%). This means, that the most significant money investment from this opening decision is data collection.

Data collection refers to a mechanism of gathering the dataset on the variables of interest from the holders or owners by using specific manners and techniques [25]. Data visualization, furthermore, refers to the action in presenting the dataset into an interactive and user-friendly interface and the ability to effectively capture the essence of the data [26]. Regarding the issue of the potential investment of money between data collection and data visualization, it is noticeable that deriving data from data providers can potentially cost expense higher than the visualizing the data. In addition, according to experts, data collection requires more than 16K Euros on average of investments, which is higher than data visualization about (14K). Therefore, the total costs for opening data decision from data collection and data visualization equal to approximate 30K Euros. Figure 3 is the complete decision tree showing all alternatives.

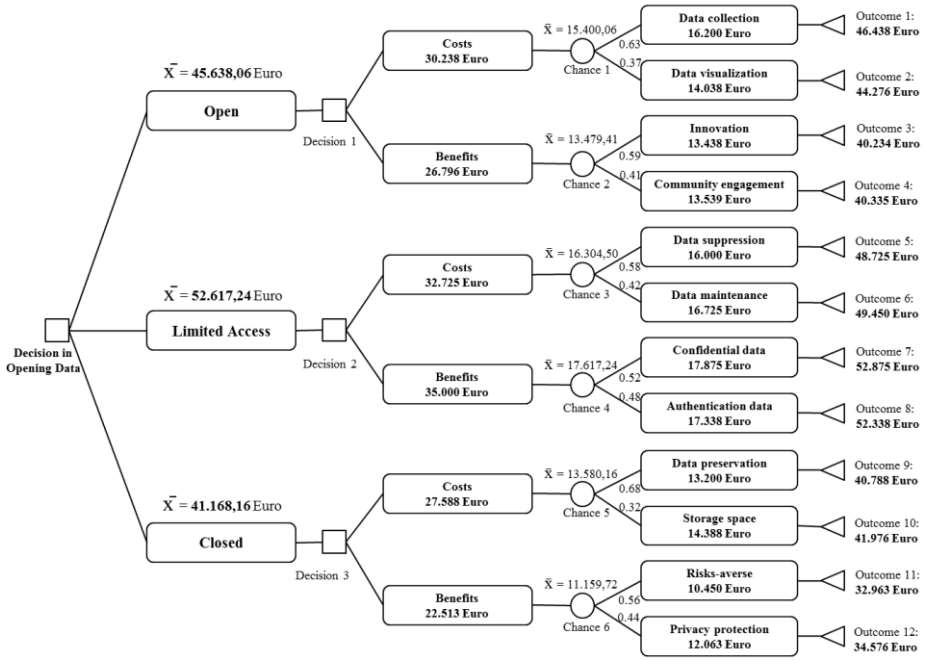


Fig. 3. Decision tree analysis to estimate the costs and benefits in open data domain

The process shown in the decision tree Figure 3 results in the payoff result depicted in Table 2. From the constructed data, we are able to compare the costs and benefits of the three decision nodes. The number values stated on each sub-element indicate the prediction of money expenses. For example, to obtain the expected monetary value from an open decision, we have to do some structured ways. First, we need to know about the average costs of data collection and data visualization by calculating the probability and estimation of the amount. Here, we calculate (0,63 x 16.200 Euro) + (0,37 x 14.038 Euro) = 15.400,06 Euro. Second, we need to estimate the costs of the open data decision by adding up the value of data collection and data visualization whereby (16.200 Euro + 14.038 Euro) = 30.238 Euro. Third, we require estimating the outcome for each sub-costs factor. To do so, the amount of data collection and data visualization should be added to the potential total costs whereby (16.200 Euro + 30.238 Euro) = 46.438 Euro (outcome 1). Whereas, the outcome 2 is obtained from (14.038 Euro + 30.238 Euro) = 44.276 Euro. Finally, after we do the same way to the benefits of factors, we require estimating the total investment of the open decision. Before we calculate the process, it is important to compare the highest potential investment between the costs and benefits factors. The reason is to determine the highest priority of the potential investment between costs and benefits consideration. In this case, the highest probability is the costs factors (30.238 Euro) instead of its benefits (26.796 Euros). Therefore, the total average of expected monetary value (EMV) for “open” decision is equal to the EMV of the costs adding up to the total value of the costs whereby 15.400,06 Euro + 30.238 Euro = 45.638,06 Euro.

4.4 Step 4: Provide decision and recommendations

Based on the constructed decision tree analysis (in Fig. 3), the final step in developing decision tree analysis is making a decision and providing some recommendations presented in decision action plans. To provide the most suitable decision between the three alternatives (open, limited access, and closed) to the decision-makers, we take into consideration the weighting process of the costs and benefits affect in open data. Next, from the EMV results, the DTA can recommend a decision as to the highest priority that might influence the investment of institutional revenue streams. We classify the findings of the study into two parts, namely:

1. Possible paths and with total payoffs

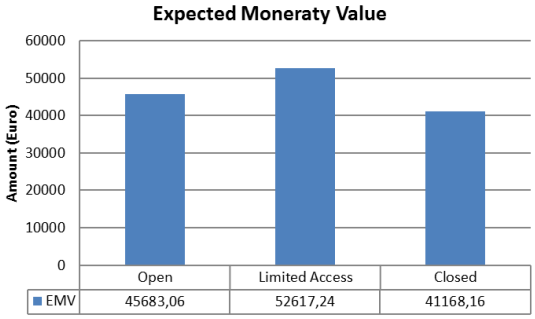
The first finding from the decision tree analysis is the possibility of the nodes and paths and its chances, as can be seen in Table 3. Every decision alternatives provide the estimation of payoffs in the euro currency. Based on these results, it can be concluded that the highest investment for the costs factor in open data domain is data maintenance where the cost almost 50K euros. Data maintenance, in this case, is the sub-nodes of the limited access decision. Meanwhile, it is noticeable that the highest potential benefit by implementing the decision is confidentiality of the data where about 52K Euros that would be a new benefit for the government institutions. In this case, the limited access decision one the hand can potentially have high costs and on the other hand, can result in high new revenues.

Table 3. Possible nodes, paths, and estimation payoffs

Terminal	Total Payoff
Decision → Open → Decision 1 → Costs → Chance 1 → Data collection	46.438 Euro
Decision → Open → Decision 1 → Costs → Chance 1 → Data visualization	44.276 Euro
Decision → Open → Decision 1 → Benefits → Chance 2 → New knowledge	40.234 Euro
Decision → Open → Decision 1 → Benefits → Chance 2 → Community engagement	40.335 Euro
Decision → Limited access → Decision 2 → Costs → Chance 3 → Data suppression	48.725 Euro
Decision → Limited access → Decision 2 → Costs → Chance 3 → Data maintenance	49.450 Euro
Decision → Limited access → Decision 2 → Benefits → Chance 4 → Confidential data	52.875 Euro
Decision → Limited access → Decision 2 → Benefits → Chance 4 → Authentication data	52.338 Euro
Decision → Closed → Decision 3 → Costs → Chance 5 → Data preservation	40.788 Euro
Decision → Closed → Decision 3 → Costs → Chance 5 → Storage space	41.976 Euro
Decision → Closed → Decision 3 → Benefits → Chance 6 → Risks-averse	32.963 Euro
Decision → Closed → Decision 3 → Benefits → Chance 6 → Privacy protection	34.576 Euro

2. Expected monetary value (EMV)

The expected monetary value (EMV) resulted from the decision tree analysis shows that the limited access decision could gain the highest monetary value of about 52K Euro. It is following the open decision in approximately 45K Euro, and the decision to keep closed the data can contribute around 41K Euro. The EMV of each decision is derived from the probability-weighted average of the expected outcome. Fig. 4 presents the detailed of EMV result and ranges of the possible investment. This EMV result can recommend the decision-makers in estimating and quantifying the amount of money required includes the investment strategies. The main advantage of the DTAOD is that the EMV can help the decision-makers in selecting a suitable choice, which involves higher benefits with less investment.



Decision	EMV	Range
Open	45.638,06 Euro	44.276 – 46.438 Euro
Limited Access	52.617,24 Euro	52.338 – 52.875 Euro
Closed	41.168,16 Euro	40.788 – 41.976 Euro

Fig. 4. The expected monetary value and investment ranges

5 Conclusion

Many government organizations are reluctant to disclose their data, because they have limited insight into the potential costs and possible adverse effects. Processing data or opening datasets partly can overcome this problem. However, this requires investments. In this study, we presented the DTAOD method to estimate the potential investments and merits of opening a dataset. This method was found to be useable by decision-makers to decide to disclose data. There are several advantages found in using DTAOD in this study. First, the decision tree can provide a better understanding of the possible outcomes of a decision alternative. Second, the proposed decision tree provides insight into selecting an informed decision. However, this is highly dependent on the alternatives that are formulated and included in the decision tree. Third, the decision tree is able to allocate the values in estimating the costs and benefits in open

data domain based on expert judgments. This provides insight into the activities needed for opening data and the associated costs and benefits.

At the same time, using DTAOD might not be easy. First, during the assign payoff process, a small change in the quantification of numerical values can lead to a large chance in the entire structure of the decision tree. Second, the calculations are based on information from experts, but these might not be correct or biased towards openness or closeness. This result shows that the high and low of expected monetary values (EMV) of a decision will influence the decision made.

This study contributes to a better understanding of the problem structure and comes up with new insight in estimating the costs and benefits of releasing data for the policy-makers. In the future research, we recommend using a different method like paired comparison, multi-voting, and net present value (NPV) methods to quantify the assign payoffs as this study using a single expert judgment.

References

1. Lourenço, R.P., *An analysis of open government portals: A perspective of transparency for accountability.* Government Information Quarterly, 2015. **32**(3): p. 323-332.
2. Ubaldi, B., *Open Government Data: Towards Empirical Analysis of Open Government Data Initiatives.* OECD Working Papers on Public Governance, 2013. **22**: p. 60.
3. Zuiderwijk, A. and M. Janssen, *Open Data Policies, Their Implementation and Impact: A Framework for Comparison.* Government Information Quarterly, 2013. **31**(1).
4. Janssen, M., Y. Charalabidis, and A. Zuiderwijk, *Benefits, Adoption Barriers and Myths of Open Data and Open Government.* Information System Management, 2012. **29**(4): p. 258-268.
5. Delgado-Gómez, D., J. C.Laria, and D. Ruiz-Hernández, *Computerized adaptive test and decision trees: A unifying approach.* Expert Systems with Applications, 2019. **117**: p. 358-366.
6. Yeoa, B. and D. Grant, *Predicting service industry performance using decision tree analysis* International Journal of Information Management, 2018. **38**(1): p. 288-300.
7. Yuanyuan, P., B.A. Derek, and L. Bob, *Rockburst prediction in kimberlite using decision tree with incomplete data.* Journal of Sustainable Mining, 2018. **17**: p. 158-165.
8. Adina Tofan, C., *Decision Tree Method Applied in Cost-based Decisions in an Enterprise.* Procedia Economics and Finance, 2015. **32**: p. 1088-1092.
9. Song, Y.-y. and Y. Lu, *Decision tree methods: applications for classification and prediction.* Shanghai Archives of Psychiatry, 2015. **27**(2): p. 130-135.
10. Zhou, G. and L. Wang, *Co-location decision tree for enhancing decision-making of pavement maintenance and rehabilitation.* Transportation Research Part C, 2012. **21**: p. 287-305.

11. Luthfi, A. and M. Janssen, *A Conceptual Model of Decision-making Support for Opening Data*, in *7th International Conference, E-Democracy 2017*. 2017, Springer CCIS 792: Athens, Greece. p. 95-105.
12. Luthfi, A., M. Janssen, and J. Crompvoets. *A Causal Explanatory Model of Bayesian-belief Networks for Analysing the Risks of Opening Data*. in *8th International Symposium, BMSD 2018*. 2018. Vienna, Austria: Springer International Publishing AG.
13. Zuiderwijk, A. and M. Janssen, *Towards decision support for disclosing data: Closed or open data?* Information Polity, 2015. **20**(2-3): p. 103-107.
14. Buda, A., et al., *Decision support framework for opening business data*, in *Department of Engineering Systems and Services*. 2015, Delft University of Technology: Delft.
15. Luthfi, A., M. Janssen, and J. Crompvoets. *Framework for Analyzing How Governments Open Their Data: Institution, Technology, and Process Aspects Influencing Decision-Making*. in *EGOV-CeDEM-ePart 2018*. 2018. Donau-Universität Krems, Austria: Edition Donau-Universität Krems.
16. Luthfi, A., et al. *A Fuzzy Multi-criteria Decision Making Approach for Analyzing the Risks and Benefits of Opening Data*. in *17th IFIP WG 6.11 Conference on e-Business, e-Services, and e-Society, I3E 2018*. 2018. Gulf University for Science and Technology (GUST), Kuwait: Springer LNCS 11195.
17. Kubler, S., et al., *A state-of the-art survey & testbed of fuzzy AHP (FAHP) applications.* Expert Systems with Applications, 2016. **65**: p. 398-422.
18. Yannoukakou, A. and I. Araka, *Access to Government Information: Right to Information and Open Government Data Synergy.* 3rd International Conference on Integrated Information (IC-ININFO), 2014. **147**: p. 332-340.
19. Beaudrie, C., M. Kandlikar, and G. Ramachandran, *Using Expert Judgment for Risk Assessment* Assessing Nanoparticle Risks to Human Health. 2016.
20. Veen, D., et al., *Proposal for a Five-Step Method to Elicit Expert Judgment.* Frontiers in Psychology, 2017. **8**: p. 1-11.
21. Mach, K., et al., *Unleashing expert judgment in assessment.* Global Environmental Change, 2017. **44**: p. 1-14.
22. Walker, K.D., et al., *Use of expert judgment in exposure assessment: Part 2. Calibration of expert judgments about personal exposures to benzene.* Journal Of Exposure Analysis And Environmental Epidemiology, 2003. **13**: p. 1.
23. Rush, C. and R. Roy, *Expert judgement in cost estimating: Modelling the reasoning p.* Concurent Engineering, 2001. **9**: p. 271-284.
24. Knol, A., et al., *The use of expert elicitation in environmental health impact assessment: a seven step procedure.* Environmental Health, 2010. **9**(19): p. 1-16.
25. Kim, S. and Y.D. Chung, *An anonymization protocol for continuous and dynamic privacy-preserving data collection.* Future Generation Computer Systems, 2019. **93**: p. 1065-1073.
26. Xyntarakis, M. and C. Antoniou, *Data Science and Data Visualization*, in *Mobility Patterns, Big Data and Transport Analytics*. 2019. p. 107-144.