EGOV-CeDEM-ePart

Bayesian-belief Networks for Supporting Decision-making of the Opening Data by the Customs

Ahmad Luthfi*, Boriana Rukanova**, Marcel Molenhuis***, Marijn Janssen****, Yao-Hua Tan*****

*Delft University of Technology, the Netherlands/Universitas Islam Indonesia, Indonesia
a.luthfi@tudelft.nl/ahmad.luthfi@uii.ac.id
**Delft University of Technology, the Netherlands, b.d.rukanova@tudelft.nl
***Customs Administration of the Netherlands, the Netherlands, jm.molenhuis@belastingdienst.nl
****Delft University of Technology, the Netherlands, m.f.w.h.a.janssen@tudelft.nl
*****Delft University of Technology, the Netherlands, y.tan@tudelft.nl

Abstract: Open government data initiatives are part of the endeavor process of governments to show that they are accountable and transparent organizations. Opening more datasets to external data analytics providers or other government organizations holds the potential to help governments to improve their processes by promoting a better understanding and enhancing the decision-making. Nevertheless, the decision-making to disclose datasets is challenging. Decision-makers often refuse to open their datasets due to several potential risks. In situations like the Dutch Customs, a dataset can contain competitive sensitive data, and multiple parties have to agree to open it. Given this complex situation, in this paper, we test a Bayesian-belief Network method for supporting the decision to open data. Our work contributes to Customs in their efforts to disclose more datasets and helping decision-makers in the process of evaluating data and defining strategies of how to move from closed to open decisions.

Keywords: Bayesian-belief Networks, Decision-making, Open Data, Customs, Risks

Acknowledgement: This research was funded by Indonesia Endowment Fund for Education (LPDP), the Ministry of Finance of Republic of Indonesia. This study was also partially funded by the PROFILE Project (nr. 786748), which is funded by the European Union's Horizon 2020 research and innovation program.

1. Introduction

Government institutions play an essential role and have the power in opening of public data. Being both a data publisher and a policy-maker, the government has a particular locus to define strategies and tools for opening its data that improves the decision-making process (Luthfi & Janssen, 2019). Besides, opening more datasets can promote a better understanding, stimulate great ideas, enhance transparency, and other value proportions (Janssen, Charalabidis, & Zuiderwijk, 2012; Kucera &

Chlapek, 2014). However, during the decision-making process for opening datasets, the governments and external stakeholders can have different roles and motivations (Gonzales-Zapata & Heeks, 2015).

Regardless of the underlying motivation for opening data, analyzing and making decisions on the status of the dataset before releasing it to the appropriate stakeholders is often challenging and not trivial. The government should take into account several risks (Martin, Foulonneau, Turki, & Ihadjadene, 2013). The possible risks could include unlocking sensitive personal data, competitive information, and opening inaccurate data (Luthfi & Janssen, 2017). As a result, these potential risk factors can influence accountability and even degrade the reputation of the government institutions (Martin et al., 2013).

In this study, we introduce a supporting tool where a conceptual model was previously developed based on the healthcare case study (Luthfi & Janssen, 2017). At that time, the proposed decision-making model was still described in a high-level overview. The prior model employed sequential steps to analyze the selected dataset. For the analysis of the dataset, a non-actual dataset sample and simulated model using the Bayesian-beliefs network method was used. Besides, a quantitative approach was used to estimate the possible adverse-risks level while constructing the causal Bayesian networks. The empirical setting for this study is a pilot project (called the Dutch Living Lab) which is part of the PROFILE[1] EU-funded research project for developing data analytics solutions for Customs (Rukanova et al., 2019). The main research question that we set to explore in this paper is "to what extent the decision support tool for opening data is applicable to the Customs context?". Hence, the objective of this study is to explore the feasibility of the decision supporting tool using Bayesian-beliefs network that was developed for the context of opening data in the healthcare domain (Luthfi & Janssen, 2017) to the context of the Customs case study.

2. Theoretical Background

In this section we first review existing models that have been applied so far in the context of opening data. We then present in more detail the decision-making model using Bayesian Networks method, which is the method we further develop and enhance in this paper.

2.1. Decision-making Models for Opening Data

In the literature study, we found that there are various models for making decisions to open data. The five systematic models that contribute to the open data domain were identified, as follows: (1) Trade-off the risks values (Zuiderwijk & Janssen, 2015). This model provides structured steps for analyzing the benefits and risks of disclosing data. (2) Decision-support framework (Buda et al., 2015). This model provided a prototype that was based on the insight of open data ecosystems. (3) Multiple Criteria for decision-making (Luthfi, Janssen, & Crompvoets, 2018). This model used a fuzziness theory to analyze the uncertainty problems and provide decision alternatives. (4) Costs and benefits of opening data (Luthfi & Janssen, 2019). This model was developed based on the

[1] https://www.profile-project.eu/

Decision Tree Analysis method. This model is used to estimate the potential advantages and disadvantages of releasing data. (5) Interactive decision-making process (Luthfi & Janssen, 2017; Luthfi et al., 2018). This model proposed a Bayesian-belief Networks method to construct the causal relationships of the decision-making process to open data in the case of health patient records. This model contributes an interesting perspective of how to examine the risks and benefits of opening data by providing sequential iteration process. The model uses a suppression technique like k-anonymity to anonymize such sensitive attributes.

The prior research listed above has explored the feasibility of these models in the context of opening data. In this paper we focus specifically on further developing the last model that we listed, namely the one using Bayesian-belief Networks method for opening data (Luthfi & Janssen, 2017)[2].

2.2. Prior Study to Open Data Using Bayesian-belief Networks

In prior research (Luthfi & Janssen, 2017) a conceptual model was developed to analyze the possibility of adverse-risks in the open data domain that makes use of Bayesian-belief Networks theory. The main motivation of the prior research was to deliver new knowledge to the decision-makers and other related stakeholders on how to make decisions to open data by using a scientific and structural manner. This model proposed four main sequential steps to analyze the potential risks, namely retrieving and decomposing dataset, evaluating, assessing, and decision-making. This model examined the health patient records dataset as an example case, and developed a systematic simulation to test the conceptual model.

Besides, to estimate the level of possible risks, quantitative approach was employed. In the assessment step, during the iterative process of decision-making, the model normalizes the table by removing several sensitive attributes of the dataset based on the Bayesian network employment. While this model demonstrates an initial application of how Bayesian Networks theory can be applied in the context of open data (Luthfi & Janssen, 2017), this previous research it did not fully integrate all the Bayesian-belief Network rules.

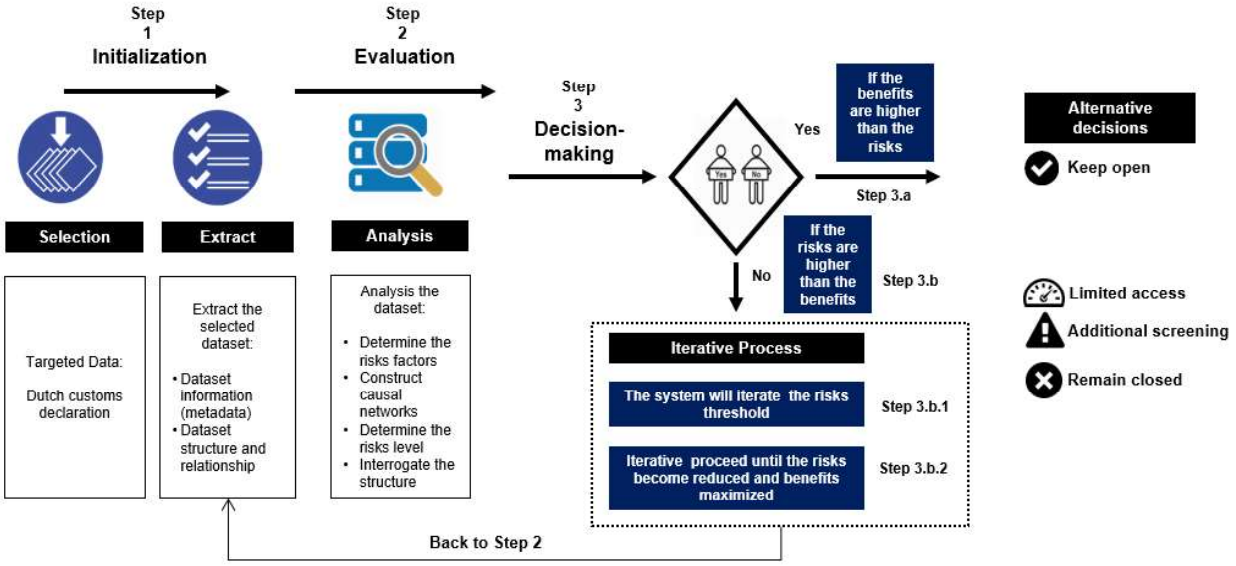
3. A Novel Conceptual Model

In this study, we use a systematic approach to apply the decision-making process to open data from the insight of Dutch Customs. As a starting point for developing our model we used the conceptual model that was initially developed in the previous study by using the health patient records dataset (Luthfi & Janssen, 2017). In this paper, we are adapting the prior decision-support model by modifying some steps to make it more effective but still take into account the comprehensive overview. In the previous model, there are four main steps, namely retrieving and decomposing dataset, evaluation, assessment, and decision-making. In this paper, we propose a new more effective process by merging the evaluation and assessment steps to become a single evaluation

[2] In this paper we focus only on further developing the use of Bayesian Networks theory in the context of opening data. Further research can also examine and further develop the use of the other models or combination thereof in the context of the opening data but this is out of the scope of the current paper.

process by implementing the Bayesian-belief Network rules. The objective to combine these steps is because the tool can employ the steps in the same process. Besides, the previous model was focused on the binary decision (open and closed), whereas in this new model, we can provide more dynamic decisions. The four decisions that are possible to take are as follows: (a) open the dataset; (b) maintain limited access to the dataset; (c) introduce additional screening; and (d) remain closed. The new proposed conceptual model for this study is presented in Figure 1.

Figure 2: A Novel Conceptual Model of Decision Support to Open Data (adapted from (Luthfi & Janssen, 2017))



In the first step (Initialization), we retrieve the datasets from the data provider. In this step, the decision support tool will extract the selected dataset into a machine- readable structure. In the second step (Evaluation), we analyze the dataset using Bayesian-belief networks method. There are four sub-steps, namely (a) determine the risk factors, (b) construct the causal relationship of the risk factors, (c) determine the risks level, and (d) interrogate the structure. Next, we evaluate the latest status of the dataset based on the single classification result in Step 2. In this step, the constructed Bayesian-belief Networks is interrogated to get the final state of the current level of risks ("high", "moderate", and "low"). Finally, the decision is made in the step 3. The expected result from this step is to provide a single classification of the dataset status, namely open, limited access, additional screening, and closed dataset. In the case that the data providers consider reanalyzing the dataset because of potential other risk effects, the decision support tool can iterate the process. The iteration process aims to update the dataset status (back to step 2) to keep certain parts of the dataset is able to be disclosed.

4. Case Study Analysis

4.1. Examine the Decision Support Tool

In order to observe and evaluate the decision support model, we apply the tool to the context of the Dutch Customs case study. In this paper, we employ the three main steps from the conceptual model shown in Figure 1.

Step 1. Initialization

In this step, the authentication process is required to indicate the groups and levels of the users namely: administrator, data analysist (experts), and decision-makers. For example, we give a privilege level from the data analyst or expert. Then, the tool selects the datasets from the data provider. The original dataset structure used in this case study is derived from the Dutch Living Lab, namely Vereeenvoudigde Aangifte e-Commerce. In this process, the decision support tool will extract the selected dataset into a readable and machine structure. The tool can select a data source from multiple database platforms like CSV, XML, JSON, etc. and ensure that the metadata of the dataset is well structured. Afterward, the tool constructs the dataset structure and its relationships.

Step 2. Evaluation

The first sub-step of this process is determining the risk factors. In this step, the tool asks the data analyst or expert to select a single or multiple risks category of the attribute. There are several risk factors provided in this tool, namely privacy infringement, data inaccuracy, data misinterpretation, data sensitivity, and data ownership. In this case, the expert selects the data sensitivity issue as we want to examine the selected dataset in terms of sensitive level. In the second sub-step, the tool constructs a causal network of the risk factors that are determined in the previous sub-step. This causal networks are developed based on the Bayesian-belief Networks formulation. The causal networks can play a role to represent a set of risks variables and their conditional dependencies via a directed acyclic graph, as shown in Figure 2. The third sub-step of this process is determining the risks level. The earlier studies conducted in the healthcare sector (Luthfi & Janssen, 2017) used a quantitative approach to determine the risks level of attributes. For that study the availability of the experts and data analysts in this field was sufficient enough to quantify and estimate the of risk level of the selected dataset. Nevertheless, in the Dutch Customs case, such accurate expertise for estimating the details the risks level is limited. Besides, in practice, doing quantitative approach will take an effort and is time consuming. Therefore, in this paper we adapt the approach to a qualitative approach to level of the risks, namely “high risks”, “moderate risks”, and “low risks”, as can be seen in Figure 4.

The last sub-step of this process is interrogating the dataset structure. In this sub-step, the tool develops the group of the Dutch Customs declaration dataset including the risks level. The goal of this interrogation is to visualize the explicit status of each attribute in terms of the data sensitivity issues. There are three color signals shown by the tool to indicate the risks level. The red attributes represent the high risk level, the yellow attributes indicate the moderate risk, and the green attributes reflect the low risk.

Figure 3: Bayesian Networks Causal Relationships

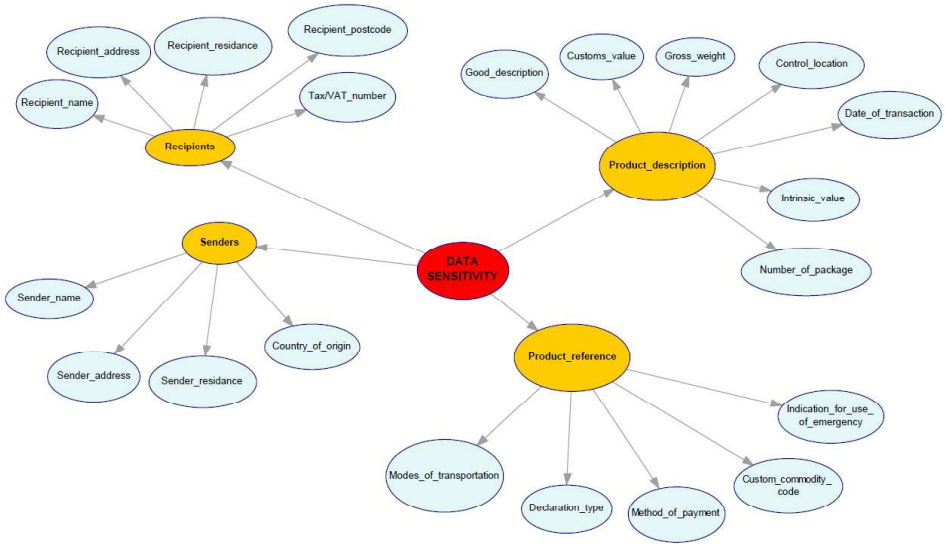


Figure 4: Determine Risk Level

Recipients

Code	Description	High	Moderate	Low
A5	Recipient's name	☑	☐	☐
A6	Recipient's address	☑	☐	☐
A7	Recipient's residence	☐	☑	☐
A8	Recipient's postcode	☐	☑	☐
A9	Tax/VAT number	☐	☑	☐

Senders

Code	Description	High	Moderate	Low
A10	Sender's name	☑	☐	☐
A11	Sender's address	☑	☐	☐
A12	Sender's residence	☐	☑	☐
A13	Country of offer/origin	☐	☑	☐

Product Description

Code	Description	High	Moderate	Low
A17	Goods description	☐	☑	☐
A18	Customs value	☐	☑	☐
A19	Gross weight	☐	☑	☐
A20	Control location	☐	☑	☐
A21	Date of transaction	☐	☑	☐
A24	Intrinsic value	☐	☑	☐
A26	Number of packages	☐	☑	☐

Reference

Code	Description	High	Moderate	Low
A1	Modes of transportation	☐	☐	☑
A3	Declaration type	☐	☐	☑
A15	Method of payment	☐	☐	☑
A16	Customs commodity code	☐	☐	☑
A14	Indication for use of emergency procedure	☐	☐	☑

Step 3. Decision-making

In this final step, the tool provides the information of the status of dataset attributes into four decision alternatives. Based on the analysis process (step 2), the tool recommends to use additional screening with respect to some sensitive attributes like recipient_name, recipient_address, sender_name, and sender_address. In order to help the data analyst to follow up the additional screening decision, the tool also take into account the action plan. In this case, we propose a salted cryptography algorithm to train the attributes. This method uses concatenating technique to randomly blur the plain text of the data value (Dubrawsky, 2009).

5. Discussion and Conclusions

Based on the entire process of the decision-making process of opening data in this study, we found some positive results. First, for the decision-support tool, it is applicable to use a qualitative approach in estimating the potential risks level of the dataset. The qualitative work has proven to make the time more efficient because the Customs expert does not require such high expertise to define and compare the risk by reflecting on the numbers (quantitative). Second, Bayesian-belief Networks is a suitable method that can analyze multiple datasets with different case studies by using systematic steps. The tool is not only able to construct the causal relationships of the risk factors, but also to interrogate the attributes by grouping the risks level. Third, by examining the real-life environment of the Dutch Customs declaration dataset, the decision-support tool has shown the more rigorous result and is recommended to be used by decision-makers.

Regarding contribution to theory, in this study we extend the method developed earlier (Luthfi & Janssen, 2017) by: (1) merging steps in the method to simplify the process and make it more efficient; (2) by incorporating a more sophisticated use of the Bayesian-belief Networks in the method compared to what was done earlier; (3) by extending the tool and demonstrating that it can support also qualitative analysis in addition to quantitative analysis. Our study also demonstrated a broader applicability of the method and the tool beyond the healthcare domain where it was originally developed also to the Customs domain. which is a very different domain of application. This increases our confidence that the tool can be applied across domains.

With respect to practice, this study contributes new insights to policy-makers and decision-makers, in particular in the Customs domain, to support their decision-making process in opening data. The decision-support tool is applicable to multiple datasets and different case studies. The use of a synthetic qualitative approach could be beneficial for government organizations who have a limited number of experts to assess the risks level. Besides, the tool can make a better understanding of the decision-makers regarding the possibility of changing policy from closed to open data sets. For future work, we recommend using multiple methods to evaluate the dataset, such as Multi-Criteria Decision Making and K-Nearest Neighbour algorithm, to get more rigorous results and findings of the proposed conceptual model to open data.

References

Buda, A., et al., Decision support framework for opening business data, in Department of Engineering Systems and Services. 2015, Delft University of Technology: Delft.

Dubrawsky, I., General Cryptographic Concepts, in Eleventh Hour Security+. 2009. p. 135-151.

Gonzales-Zapata, F. and R. Heeks, The Multiple Meanings of Open Government Data: Understanding Different Stakeholders and Their Perspective. Government Information Quarterly, 2015. 32: p. 441-452.

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.

Kucera, J. and D. Chlapek, Benefits and Risks of Open Data. Journal of System Integration, 2014. 1: p. 30-41.

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.

Luthfi, A. and M. Janssen, Open Data for Evidence-based Decision-making: Data-driven Government Resulting in Uncertainty and Polarization. International Journal on Advanced Science Engineering and Information Technology, 2019. 9(3): p. 1071-1078.

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.

Martin, S., et al., Risk Analysis to Overcome Barriers to Open Data. Electronic Journal of e-Government 2013. 11(1): p. 348-359.

Rukanova, B., et al. Value of Big Data Analytics for Customs Supervision in e-Commerce. in International Conference on Electronic Government. 2019. San Benedetto del Tronto, Italy: Springer.

Zuiderwijk, A. and M. Janssen, Towards decision support for disclosing data: Closed or open data? Information Polity, 2015. 20(2-3): p. 103-107.

About the Authors

Ahmad Luthfi

Ahmad Luthfi is a PhD researcher at Delft University Technology, the Netherlands. He holds his master's in the field of Computer Science at Gadjah Mada University, Indonesia. His research interests are in the area of Open Government Data, Decision-making Process, and Decision Support Systems.

Boriana Rukanova

Dr. Boriana Rukanova is a researcher at Delft University of Technology. Her research interest include digital infrastructure innovations in international supply chains, upscaling of innovations, and value of data analytics for government supervision.

Marcel Molenhuis

Marcel Molenhuis is a senior consultant data & analytics at Secretary Coordination Group Innovation (CGI) in Customs Administration of the Netherlands.

Marijn Janssen

Prof. dr. Marijn Janssen is full professor in ICT & Governance at the Delft University of Technology and head of the Information and Communication Technology section. His research is focused on ICT-architecting which multiple public and private organizations.

Yao-Hua Tan

Prof. dr. Yao-Hua Tan is full professor of Information and Communication Technology at the department of Technology, Policy and Management of the Delft University of Technology. His research fields are IT innovation for e-customs to make international trade more secure and safe; IT architectures for data sharing and compliance management for international supply chains; artificial intelligence and data analytics for customs risk targeting and improve logistic efficiency in international trade.