Bayesian Networks to Analysis Electronic Government using the K2 Algorithm

***De la Torre-Gea, Guillermo Alfonso[1], Horacio González-Pérez[1], Diego Soto-Hernández[1]**
[1]Universidad de la Sierra Sur, Oaxaca, México.
* gtorre@abanet.mx

Gabriela García Manzo[2]
[2] Instituto de Investigación y Desarrollo de Tecnologías Garman A.C. México.

***Abstract*— Public agencies have fostered the development of e-government services over the past decade, promoting more and better administrative services through digital channels. However, information technology management does not have the ability to directly assess the performance of the e-government model, since its scope of control includes several indicators of governance maturity. Current e-government frameworks are adequate to describe governance, but lack the ability to predict changes in governance maturity indicators that affect governance performance. There have been numerous studies that propose new models applied to specific situations. The objective of this study was to determine and deduce the dependencies between the variables assigned to e-government in Mexico during a survey conducted by INEGI in 2013, using the K2 algorithm. The behavior of all the variables studied using the algorithm K2 shows that, when using a transactional web portal, all variables are related; Structure of Procedures and services available on the web" is the variable with the highest number of dependencies. Having an e - government web portal defines the issue of the procedure and increases the tax paymen by 22%, vehicle registration payment 13% and civil registration procedures in 8%. The growth of transactional web portals should be considered as a priority theme within development public politics.**

Keywords—E-Government, Bayesian Networks, data mining, public politics.

I. INTRODUCTION

Governments and public agencies have fostered the development of e-government services over the past decade, promoting more and better administrative services through digital channels. The goal of e-government is to achieve internal efficiency in a government organization, supported by information technology as a process facilitator. However, information technology management does not have the ability to directly evaluate the performance of the e-government model, since its scope of control includes several maturity indicators of governance, such as the existence of different activities and processes, documents, metrics and roles.

On the other hand, current e-government frameworks are adequate to describe governance, but lack the ability to predict changes in governance maturity indicators that affect governance performance. That is why several authors have been given the task of developing prediction models applied to e-government. Simonsson et al [1] presented an application based on Bayesian networks for the prediction of e-government performance. The resulting Bayesian network could be used to support IT governance decision-making. Similarly, the impact of this process has not been fully evaluated. Fernández-i-Marín [2] analyzed through European countries the influence of e-government policies on their adoption, under different levels of Internet penetration, which allowed evaluating the promotion of e-government.

In this paper, a certain level of Internet penetration and policies focused on e-government have a substantial impact on the adoption of technology by citizens, highlighting the importance of investing in e-government; When their effects may be greater. The Bayesian inference used allowed the research to avoid common artificial assumptions in comparative political research, to design more flexible models and to present the results in an explicit way.

E-government security is considered to be one of the crucial factors in achieving an advanced stage of e-government. As the number of services increases, a higher level of security is required. According to Fadl Elssied et al. [3], fuzzy set theory is very useful for evaluating e-government security. De-Juanas et al. [4] proposed an instrument to verify the presence of quality indicators of scientific web portals related to the Social and Health Sciences.

Three studies were carried out from a non-experimental mixed-type design: document analysis and qualitative and quantitative assessments; The result was the development of an instrument, QuaSciWeb, which contains 6 categories of analysis that group 57 items: 1) identity and authorship; 2) user interface; 3) content; 4) data navigation and retrieval; 5) user experience; 6) visibility and 7) disclosure. The

instrument allows assessing the quality of portals and web pages of a scientific nature, identify the contents necessary for its improvement, and provide concrete guidelines that can be used by web designers and programmers.

As Internet services increase, government systems reinvent themselves as e-government services, emerging new techniques. Zhang et al. [5] presented a new method for e-procurement that allows searching the optimal scheme of acquisitions, through the evaluation and selection of services, applying a trapezoidal fuzzy number similarity algorithm to support element-based collaborative filtering and approach Bayesian, where services can be expressed as static values and represented as diffuse values.

The main problem to evaluate the performance of an e-government portal is the high number of variables that must be handled simultaneously. Dondeynaz et al. [6] proposed a set of models with 25 variables that suggest the use of a Bayesian network to perform the modeling of methods, due to the ease of adapting to complex probability distributions, and to integrate a qualitative approach for the data analysis, offering the advantage of integrating preliminary knowledge in probabilistic models. The statistical performance of the proposed models ranges from 20% to 5% of error rates and allows evaluating the relationships between human development, external support, governance aspects, economic activities, and access to water supply and sanitation.

When we analyze one set of variables at a time, this set is more than the sum of its parts. A Bayesian network model was developed by Hoshino et al. [7] which included the linkages between factors for the fishing communities of the Kei Islands in Indonesia, based on detailed local surveys. The results showed that cumulative impacts of multiple factors on major social, economic, and environmental outcomes may be much greater than the impact of a single source, implying that management or policy intervention could be more effective when multiple factors are addressed simultaneously. Hence the importance of using heuristic tools such as Bayesian networks.

There have been numerous studies that propose new models applied to specific situations. Some algorithms have not yet been tested as the K2 that allow obtaining Bayesian network structures from the observed data even in incomplete sets. The statistical methods of automatic learning have been applied in Bayesian statistics; however, mechanical learning may employ a variety of classification techniques to produce distinct models of Bayesian networks.

Bayesian networks are techniques that allow the analysis of many variables immediately [8], allowing inferences that other techniques do not allow. Therefore, the objective of this study was to determine and deduce the dependencies between the variables assigned to e-government in Mexico during a survey conducted by INEGI in 2013, using the K2 algorithm.

II. BAYESIAN THEORY

Bayesian network models are the representation of the knowledge used in the field of Artificial Intelligence for approximate reasoning [9]. According to Correa [10], the nodes correspond to concepts or variables and their links correspond to relations or functions. Functional relationships describe causal inferences expressed in terms of conditional probabilities in which variables are defined in a discrete or qualitative domain, as shown in equation 1.Prepare Your Paper Before Styling.

$$P(x1,\ldots,xn) = \prod_{i=1}^{n} P(xi \mid parents\,(xi)) \qquad (1)$$

Hruschka et al. [11], found that a Bayesian network could be used to identify previously undetermined relationships between variables. Bayesian networks describe and quantify these relationships even with an incomplete data set. The Bayesian network solution algorithm allows the calculation of the expected probability distribution of the output variables. The result of this calculation depends on the probability distribution of the input variables. Overall, a Bayesian network can be perceived as a joint probability distribution of a collection of discrete random variables (equation 2).

$$P(cj|xi) = P(x_i|c_j)\,P(cj) \,/ \sum_{k} P(x_i|c_k)P(ck) \qquad (2)$$

The most representative method of automatic learning in the artificial intelligence approach is the algorithm K2. This method is widely used even if it has the drawback of requiring the specification of an enumeration order in variables. The idea is to maximize the probability of the structure given the data in the search space of the acyclic graph directed respecting this order of enumeration [12].

III. MATERIALS AND METHODS

In order to perform this work, an Intel Core-i7 computer with 8 GB of RAM and Windows 8 operating systems was used. A data set was divided into two intervals using the Elvira v 162 software, to be used in Bayesian network model development, which describes the relationships between all the studied variables.

The Bayesian network analysis was carried out in three stages suggested by García-Manzo [13]:

A) Pre-treatment: The algorithm of imputation by zeros was used, which substitutes the values lost and unknown by the value equal to zero. No need for parameters. The discretization of the variables was of massive type, using the algorithm Equal frecuency with number of intervals equal to five.

B) Processing: It was done using the learning method K2 learning, with maximum likelihood and maximum number of parents equal to 5, without restrictions.

C) Post-processing: Formulation of the scheme of dependencies between variables: From the data set of the 8 variables, a dependency analysis was performed between all of them.

This analysis consisted in the quantification of the different types of dependence and gave as a result the identification of the existing causal relationships between the said variables. The structural learning of the Bayesian networks was carried out from the identification of the causal relationships between variables, which determined the structures.

After obtaining the parametric learning network, we calculated the conditional probabilities for the variables that show relation or dependence. The state variables are shown in Table 1 and correspond to the survey questions: National Census of Government, Public Security and State Penitentiary System (INEGI, 2014), subject Electronic government, procedures and services.

TABLE I. STATUS VARIABLES: E-GOVERNMENT, PROCEDURES AND SERVICES.

Name	Coding
Structure	ESTRUCTU
Topic of the process	TEMAS_WB
Catalog of procedures	CON_CATA
Access to catalog	ACC_CATA
Contents to the catalog	CONT_CAT
Type of Web services	***TIPO_WEB***
Process management	GEST_TRAM
Theme of the process	LIST_TEMA

Likewise, data from the same survey were taken on the type of connection in the infrastructure of the e-government portals for each Mexican entity, whose status variables are presented in Table 2.

TABLE II. STATUS VARIABLES: TYPE OF CONNECTION IN THE INFRASTRUCTURE OF E-GOVERNMENT PORTALS.

Name	Coding
State	ENTIDAD
Computer hardware available	ESTRUCTUR
Existence of a single Website for all Institutions	SIT_WEB
Type of computer network	CINT_TIP
Internet connection	CON_INT
Website in the Institution	FUN_SIWB
Function performed by the Institution	FUNCION

Both sets of data were analyzed independently, in the first case with the objective of knowing the variables related to the quality of the digital government service attended by the Institution, in terms of the functionality of the TRANSACCIONAL Web Site, this is possible to make payments For services. In the second set of data, the inference was made according to whether there is an internet connection that counted the institutions of Public Administration of the federative entities.

IV. RESULTS AND DISCUTION

According to the analysis performed using the ELVIRA software described above, we can visualize that, using the K2 algorithm, all the variables shown in Table 1 are related differently;

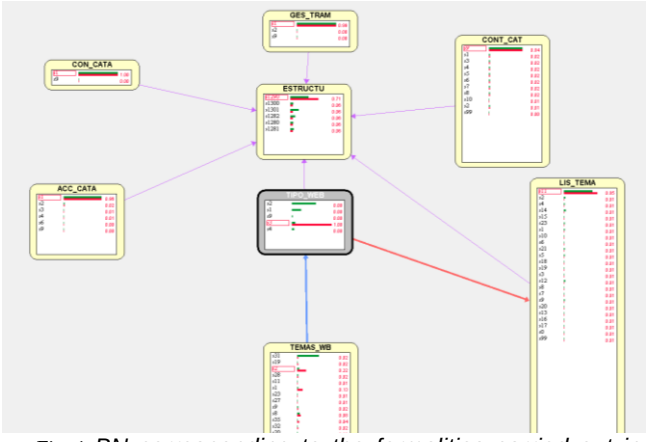

Fig. 1. *BN corresponding to the formalities carried out in the e-government portals for each Mexican entity.*

The intention of this work was to determine how all variables studied behave when using an e-government web portal that performs transactional operations, so the results are shown in Table 3.

TABLE III. STATUS VARIABLES WHEN THERE IS A TRANSACTIONAL WEB PORTAL.

Name	value	*P(x)\P(TIPO_WEB)*
Structure of procedures and services	Procedures by theme in web portal	0.71
Subject of processing	Tax payment	0.22
Procedures catalog	Existence of a catalog of services and formalities	1.0
Access to the catalog	Available on website	0.96
Contents of the catalog	The procedures are grouped by frequency of use	0.94
Management of processing	Management of any procedure in any of the Institutions	0.99

There is a significant difference in the use of a web portal of g - government when the type of web is informative that when it is transactional, this difference defines the subject of the process towards payment of taxes in a 22%, Payment of Vehicle ownership (13%) and Civil Registry procedures with (8%), which indicates that transactional web portals are a source of income for the government. On the other hand, this type of analysis should allow governments to obtain more accurate studies that result in a political intervention is more effective when multiple factors are addressed simultaneously.

The results of the Bayesian analysis corresponding to the type of connection in the infrastructure of the e-government portals shown in Figure 2, allow us to understand the importance of having a good network infrastructure by the Institutions, where access Internet is necessary; where access Internet is a critical variable; Since 88% is defined between the Local (57%), Metropolitan (21%) and World (20%) network type.

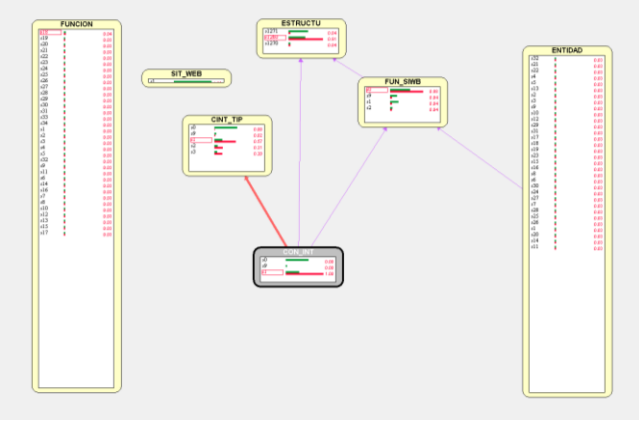

Fig. 2. *BN corresponding to the type of connection in the infrastructure of the electronic government portals for each Mexican entity.*

Two variables studied are independent of the type of infrastructure: Existence of a single Website for all Institutions and Function performed by the Institution, which we consider not necessary to include in this study, since they do not provide any analysis; Which gives us evidence of poor design of INEGI surveys. Therefore it is necessary to incorporate the use of multivariate analysis tools such as Bayesian Networks in a stage prior to the design of the questionnaires.

The variable "Website Existence in the Institution" allows us to know that the number of transactional web portals is not enough in the case of Mexico, so it should be considered as a priority theme within the development policies, both to increase The tax collection, as well as to increase the quality of the services offered by the State and Federal governments.

V. CONCLUSION

The behavior of all the variables studied using the algorithm K2 shows that, when using a transactional web portal, all variables are related in different ways; "Structure of Procedures and services available on the web" is the variable with the highest number of dependencies, while the "Process topic" is related to the "Web portal type". Having an e - government web portal defines the "Issue of the procedure" and increases the "Tax payment" by 22%, "Vehicle Registration Payment" 13% and "Civil Registration Procedures" in 8%, for Both transactional web portals are a source of revenue for the government. For this, having a good network infrastructure by the institutions is imperative. It is necessary to incorporate the use of multivariate analysis tools such as the Bayesian Networks in a stage prior to the design of the questionnaires. The growth of transactional web portals should be considered as a priority theme within development public politics, both to increase tax collection and to increase the quality of services

ACKNOWLEDGMENT

The preferred spelling of the word “acknowledgment” in America is without an “e” after the “g.” Avoid the stilted expression “one of us (R. B. G.) thanks .”. Instead, try “R. B. G. thanks.”. Put sponsor acknowledgments in the unnumbered footnote on the first page.

REFERENCES

[1] Mårten Simonsson, Robert Lagerström, and Pontus Johnson. A Bayesian network for IT governance performance prediction. In Proceedings of the 10th international conference on Electronic commerce (ICEC '08). ACM, New York, NY, USA (2008). Article 1 , 8 pages. DOI: http://dx.doi.org/10.1145/1409540.1409542

[2] Fernández-i-Marín, Xavier. The Impact of e-Government Promotion in Europe: In-ternet Dependence and Critical Mass. Policy & Internet, 3(4); 1944-2866 (2011). DOI: 10.2202/1944-2866.1093

[3] Nadir Omer Fadl Elssied, Othman Ibrahim, Adil Ali A.alaziz and Adil Yousif. Re-view Paper: Security in E-government Using Fuzzy Methods. International Journal of Advanced Science and Technology Vol. 37 (2011).

[4] Ángel De-Juanas, Rodrigo Pardo, Alfonso Diestro, Amelia Ferro, Javier Sampedro. Construcción de un instrumento de verificación de la calidad de portales y redes de investigación de carácter científico en Internet. Revista española de Documentación Científica, Vol 35, No 4 (2012).

DOI: http://dx.doi.org/10.3989/redc.2012.4.900.

[5] Shuai Zhang, Chengyu Xi, Yan Wang, Wenyu Zhang, and Yanhong Chen, "A New Method for E-Government Procurement Using Collaborative Filtering and Bayesian Approach," The Scientific World Journal, vol. 2013, Article ID 129123, 10 pages (2013). doi:10.1155/2013/129123

[6] C. Dondeynaz1,2, J. López Puga3, and C. Carmona Moreno. Bayesian networks modelling in support to cross-cutting analysis of water supply and sanitation in de-veloping countries. Hydrol. Earth Syst. Sci., 17, 3397-3419. (2013).

[7] Hoshino, E., I. van Putten, W. Girsang, B. P. Resosudarmo, and S. Yamazaki. A Bayesian belief network model for community-based coastal resource management in the Kei Islands, Indonesia. Ecology and Society 21(2) (2016).

http://dx.doi.org/10.5751/ES-08285-210216

[8] Espinoza-Huerta, T. D., Ortíz-Vázquez, I. C., García-Manzo, G., & De la torre-Gea, G. A. (2015). A Multivariable Computational Fluid Dynamics Validation Method Based in Bayesian Networks Applied in a Greenhouse. International Journal of Agriculture Innovations and Research, 4(1), 67-71.

[9] Gámez, J. A., Mateo J L, Puerta, J. M. Learning Bayesian networks by hill climbing: efficient methods based on progressive restriction of the neighborhood, Data Min. Knowl. Discov, 22, (2011) 106.

[10] Correa, M., Bielza, C., Paimes-Teixeira, J., Alique, J. R. Comparison of Bayesian networks and artificial neural networks for quality detection in a machining process, Expert Syst Appl, 36 (3) (2009) 7270.

[11] Hruschka E, Hruschka E, Ebecken N. F. F. Bayesian networks for imputation in classification Problems, J Intell Inform Syst, 29 (2007) 231.

[12] Guoliang L, Knowledge Discovery with Bayesian Networks, Ph. D. thesis, National University of Singapore, Singapore, 2009.

[13] García-Manzo, G., De la Vega-Flatow, J. N., Martínez-Alcaráz, S. L., Quijada-López, R. M., Rodríguez-Reyes, C. S., & De la Torre-Gea, G. A. (2016). Determination of relationship patterns in EEG and BVP signals using the K2 learning algorithm. Ponte Journal, 72(3): 22 – 32.