Characterization of the use of e-commerce and e-government in Oaxaca as an indicator of the Digital divide by means of Bayesian classifiers

***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*— The Digital Divide, which contributes to reinforcing the differences that exist between countries, has formed an isolation in the lifestyle in diverse communities that do not have access services to the Internet. Therefore it is necessary to implement information systems according to regional characteristics, which should be a budgetary policy priority. Despite the growing importance of this problem and its manifestations as an emerging inequality, there is a lack of research, both in theoretical and empirical terms. Having reliable data and systematically compiled with categories of analysis relevant to the case, would make it possible to approximate a characterization of the Digital Divide in Mexico; As well as to know their behavior and identify their evolution and trends. Bayesian networks are techniques that allow multivariate analysis, to obtain inferences. Therefore, the objective of this study was to determine and deduce the dependencies between the variables assigned to trade and e-government in Oaxaca and compared to Mexico City, using the data obtained by a survey carried out by INEGI in 2013, Use of the K2 algorithm.**

Keywords—E-Government, Bayesian Networks, data mining, Internet access, network infrastructure.

I. INTRODUCTION

The major changes that characterize the present society are the generalization of the use of technologies, communication networks, rapid technological and scientific development, and the globalization of information. The accelerated development of telecommunications and informatics in the second half of the twentieth century, as well as the Internet in the last decade, has influenced all fields of human endeavor.

The Digital divide is not exclusively linked to technological aspects because it is a combination of socio-economic factors and in particular of limitations and lack of telecommunications and information technology infrastructure (Serrano-Santoyo, 2003). Digital divide, which contributes to reinforcing the differences between countries, has shaped a technological difference resulting from other social, economic, political, generational, geographical, cultural and gender gaps (Caridad-Sebastian and Ayuso-García, 2011).

Differentiated access to knowledge on both the generator and consumer sides persists, so the need to implement information systems in line with regional characteristics should be a policy priority in science and technology (Garduño-Oropeza et al., 2008).

Mexico is a country of inequalities of diverse facets and expressions to which the Digital divide is being added. Despite the growing importance of this problem and its manifestations, there is a lack of research, both in theoretical and empirical terms.

More than half of the national population does not have access, use and appropriation of the Internet, this being the most characteristic service of the Information and Knowledge Society, since for 2014, of 112 million 336 thousand inhabitants of Méxco, only 51.2 million were users of the network according to the Mexican Internet Association (AMIPCI), while 60 million are disconnected.

Regarding the gender dimension, the aforementioned study assures that the distribution of internet users is distributed by 50% between men and women; In terms of age groups, there are important differences: in 2013, the largest number of users was in the group aged 13 to 18 (24%), while the lowest access group was the elderly Of 55 (4%) (AMIPCI, 2014: 7).

Having reliable data and systematically compiled with categories of analysis relevant to the case, would make it possible to approximate a characterization of the BreDig in Mexico, as well as to know its behavior and identify its evolution and trends. According to data from the Organization for Economic Co-operation and Development (OECD), our country is located on the 32nd of the 34 countries belonging to that organization in the broadband category, with transmission speeds much lower than those recorded on average in those nations (OECD, 2012: 14).

Other statistics released by the World Bank reflect a significant backwardness of the country in

connectivity to the Internet and the use of ICTs, which puts the country at the 76th position among 142 economies in the world (World Bank, 2013b), even below Countries with lesser or similar development: if Chile appears with an index of 61.4 and Argentina 56, Brazil with 50, Colombia with 49, Jamaica with 47 and Venezuela with 44, Mexico appears with 38, next to Peru, surpassing Ecuador (35), Honduras (18) and Haiti (10).

The National e-Mexico System proposed by the government since 2000 was designed from a modernizing and techno-deterministic perspective, giving ICT the ability to solve the structural problems of the country, which proposed to make a quantum leap "In the development and not only" closing the BreDig, but also the one of education, health, access to the markets and the existent with the government, especially the one that prevails between the governmental government and the locals "according to a Report made by the Secretariat of Communications and Transportation (SCT).

Raising the BreDig is a task of great proportions and indispensable to carry out, to stop the advance of a new social inequality and to take advantage of the opportunities for the development of the country that the ICT can contribute, in terms of generating knowledge and Support for productive activities.

In order to propose a formal solution on the use of some technology as a means of reducing BreDig, it is necessary to characterize the conditions in the use of internet in marginalized communities. A survey carried out by INEGI (2013) provided relevant information for this purpose, whose results can be obtained from its portal as "hard data" that can be used as inputs in an inference process, using multivariate analysis tools.

On the other hand, Bayesian networks are techniques that allow the analysis of many variables immediately (García-Manzo et al., 2016), 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 the use of internet in marginalized communities in the state of Oaxaca and to compare them with the Mexico City, through a multivariate analysis applied to the results obtained Of this survey, using the algorithm K2.

II. Bayesian Network Theory

According to Gamez [9], Bayesian network models are the representation of the knowledge used in the field of Artificial Intelligence for approximate reasoning. The nodes correspond to concepts or variables and their links correspond to relations or functions [10]. 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.

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

A Bayesian network could be used to identify previously undetermined relationships between variables [11]. 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 probability distribution of the input variables depends on the result of this calculation. Overall, a Bayesian network (equation 2) can be perceived as a joint probability distribution of a collection of discrete random variables.

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

The K2 algorithm is the most representative method of automatic learning in the artificial intelligence approach. This method is widely used even if it has the drawback of requiring the specification of an enumeration order in variables. According to Guoliang [12], the objective of a Bayesian Network model 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.

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 were 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.

Data processing was performed to obtain them in comma-separated text format (CVS), for later Bayesian analysis using ELVIRA software through the algorithm K2, which according to García-Manzo et al. [13] has provided good results in its implementation as an analysis tool. The data were separated into 25 columns referring to e-Commerce in Mexico, each of which corresponds to a question taken as an analysis variable, where the variable pag-gob corresponding to the payment of government services and formalities appears, as shown in Table 1.

TABLE I. List of study variables.

QUESTION	ANSWER	VARIABLE
7.10 In the last 12 months, have you shopped online? Exclude purchases for work and include those made for your business	1 Yes 2 No b not answer	Commercial
7.11 In the last twelve months, what was the largest amount Did purchase make online?	a) Amount b) not answer	Monto
7.12 In the last 12 months, how often have you shopped online?	1 Daily 2 week 3 Month 4 six month 5 a year 9 not answer	Frecuencia
7.13 In the last 12 months, what did you buy online? 01 books?	1 Yes 2 No 9 not answer	Libros

	1 Yes	
02 music and videos?	2 No	
	9 not answer	
	1 Yes	Music
03 computers?	2 No	
	9 not answer	Comp
	1 Yes	
04 software?	2 No	Softw
	9 not answer	
	1 Yes	
05food and drinks?	2 No	Alimentos
	9 not answer	
	1 Yes	
06 tickets?	2 No	Boletos
	9 not answer	
	1 Yes	
07 electronic devices?	2 No	Elect
	9 not answer	
	1 Yes	
08 home services?	2 No	Hogar
	9 not answer	
	1 Yes	
09 personal services?	2 No	
	9 not answer	Person
	1 national?	
7.14 Is the Internet site where you have made purchases …	2 foreigner?	Nal
	3 both?	
	9 Unknown	
	9 Not answer	
	1 Yes	
7.15 ¿Tuvo algún problema al realizar la compra?	2 No	
7.16 ¿El problema que tuvo fue…	9 Not answer	Problem
	1 Yes	
	2 No	
1 relacionado con la entrega?	9 Not answer	Entrega
2 con la seguridad del pago?	1 Yes	
	2 No	Pago
	9 Not answer	
3 con la seguridad de los datos personales?	1 Yes	Segur
	2 No	
	9 Not answer	
4 Other problems	1 Yes	
	2 No	
	9 not answer	
	Specific it	
	9 Not answer	
7.17 En los últimos doce meses, ¿ha realizado pagos por Internet?	1 Yes	Pago
	2 No	
7.18 En los últimos doce meses, ¿cuál es el monto del mayor pago realizado por Internet?	9 Not answer	
	amount	pag-monto
	9 Not answer	
	1 Daily	
7.19 ¿Con qué frecuencia ha realizado pagos por Internet?	2 weekly	pag-frec
	3 monthly	
	4 six month	
	5 yearly	
	9 Not answer	
7.20 ¿Los pagos que ha realizado por Internet son…	1 Yes	
1 servicios bancarios y financieros?	2 No	pag-banca
	9 Not answer	
2 servicios y trámites de gobierno?	**1** Yes	**pag-gob**
	2 No	
	9 Not answer	
3 servicios educativos?	1 Yes	
	2 No	pag-edu
	9 Not answer	
4 bienes y servicios para el hogar?	1 Yes	
	2 No	pag hogar
	9 Not answer	
5 bienes y servicios personales?	1 Yes	
	2 No	pag-person
	9 Not answer	
6 Otro pago	1 Yes	
	2 No	
	9 Not answer	
	Specific it	

The data were filtered in two federative entities that we consider to represent the BreDig in our country: Cd. De México and Oaxaca, to later obtain a BN for each entity.

The Bayesian network analysis was carried out in three stages suggested by Ortíz-Vázquez et al. [14]: a) Pre-treatment, b) Processing, and c) Post-processing. After obtaining the parametric learning network, we calculated the conditional probabilities for the variables that show relation or dependence, taking as an event the variable gob to make the inference when we have 100% probability, which gives us The behavior of all variables when the payment of services to the government has been made.

IV. RESULTS AND DISCUSSION

TABLE II. *A PRIORI* AND *A POSTERIORI* PROBABILITY VALUES OF THE STUDY VARIABLES FOR OAXACA AND CD. DE MÉXICO.

VARIABLE	OAXACA (2,778 Surveys)		MEXICO (2,303 Surveys)	
	a priori	*a posteriori*	*a priori*	*a posteriori*
comercio	4%	47%	9%	64%
monto	$500.00 95% ($52,782.00)	$500.00 56% ($52,782.00)	$10,000.00 89% ($1,844,703.00)	$10,000.00 37% ($5,568,654.00)
frecuencia	1 vez al año 96%	1 vez al año 67%	1 vez al año 93%	1 vez a la semana 44% 1 vez a los 6 meses 23% 1 vez al mes 30%
libros	1%	14%	3%	28%
musica	1%	13%	5%	64%
comp	1%	7%	1%	5%
softw	1%	22%	4%	29%
alimentos	0	4%	1%	21%
boletos	2%	23%	97%	83%
elect	97%	68%	2%	16%
hogar	97%	66%	2%	14%
person	98%	77%	6%	34%
nal	99%	87%	Nal y extranjero 95%	Nal y extranjero 68% Nal. 21%
problem	0%	5%	1%	6%
entrega	0%	2%	100%	97%
pago	0%	2%	0%	2%
segur	0%	0%	0%	0%
pago	4%	96%	9%	94%
pag-monto	$400.00 95% ($42,225.00)	$5,000.00 10% ($652,830.00)	$2,200.00 88% ($401,274.72)	$500.00-$99,000.00 88% ($1,597,250.00)
pag-frec	1 vez al mes 95%	1 vez al mes 39% 1 vez a 6 meses 28% 1 vez al año 21%	1 vez a la semana 91%	1 vez al mes 56% 1 vez a 6 meses 23% 1 vez a la semana 11%
pag-banca	3%	49%	97%	74%
pag-gob	1%	**100%**	4%	**100%**
pag-edu	1%	22%	1%	14%
pag hogar	1%	37%	3%	30%
pag-person	97%	46%	5%	41%

Based on the a priori and a posteriori probability values of the study variables for Oaxaca and Cd. De México, an estimate was made of the payment to the government for services through the electronic platform

performed per year, as shown in Table 3, by the following equations:

Total amount = (Amount) (% Amount) (surveys) (% trade) (3)

Total pay-amount = (pay-amount) (% pay-amount) (surveys) (% trade) (4)

Payment-total amount / year = (Payment-total amount) (fre-fee) (% fre-fee) (5)

TABLE III. ESTIMATED PAYMENT TO GOVERNMENT FOR SERVICES THROUGH THE ELECTRONIC PLATFORM

Oaxaca		México	
a priori	*a posteriori*	*a priori*	*a posteriori*
$481,365.00/año	$30,963,726.90/año	$18,988,319.75/año	$2,060,452,500.00/año

Figure 1 shows the result of the inference of the model obtained by the Bayesian classifier using the algorithm K2 when it is certain that the user has already made some payment to the government through the internet in the State of Oaxaca.

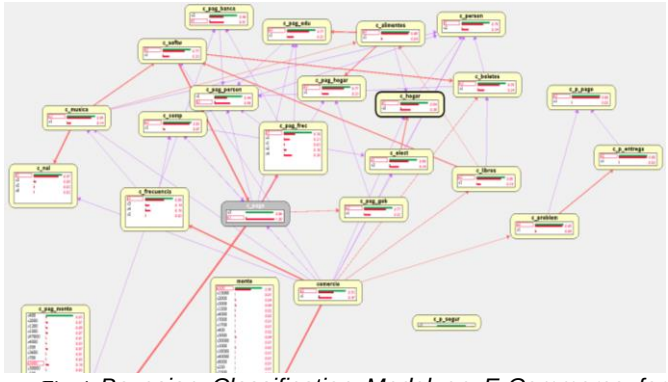

Fig. 1. *Bayesian Classification Model on E-Commerce for the State of Oaxaca, Mexico*

Comparing the results of the BN between DF and Oaxaca, higher percentages are observed in DF, which are attributable to the fact that the digital culture is most ingrained, which can be attributed to technological lag and geographic determinism. However, it is important to mention that the structure of BN also presents some variations between cities compared; which are mentioned as follows:

TICKETS: national.
AMOUNTS: 10 to 20 times greater in Mexico City.
PROBLEMS: in DF only 6% presented problems of mainly DELIVERY (98%).
PAYMENT TO GOVERNMENT: FOOD AND BOOKS
MUSIC - NATIONAL: Oaxaca buys while DF does not.
SOFTWARE: Oaxaca buys while DF does not.

These changes in the structure of the network of dependencies allow us to understand that these differences are not only in terms of quantities or percentages but differences in customs or activities of daily living.

V. CONCLUSION

BN allows us to analyze a large number of variables at a time and infer specific situations that lead us to perform a synthesis work on a larger scale than using the conventional tools of statistics. BN must be implemented as a tool for analyzing and predicting the country's development in e-government policies.

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] Arturo Serrano Santoyo, E.M.M., La brecha digital, in La brecha digital, Mitos y Realidades, UABC, Editor. 2003, UABC: México. p. 175.

[2] Caridad Sebastian, Mercedes, & Ayuso García, María Dolores. (2011). Situación de la BreDig de género y medidas de inclusión en España. Investigación bibliotecológica, 25(55), 227-252. Recuperado en 23 de marzo de 2017, de http://www.scielo.org.mx/scielo.php?script=sci_arttext&pid=S0187-358X2011000300009&lng=es&tlng=es.

[3] Garduño Oropeza, Gustavo; Aguado López, Eduardo; Rogel Salazar, Rosario; Zúñiga, María Fernanda; (2008). Redalyc: una alternativa a las asimetrías en la distribución del conocimiento científico. Ciencia, Docencia y Tecnología, Noviembre-Sin mes, 11-30.

[4] INEGI, 2013 CNGSPE, 2013.

[5] AMIPCI, 2014: 5

[6] OCDE, 2012:14

[7] SCT, 2001: 229

[8] 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.

[9] De la Torre-Gea, G., Horacio González-Pérez, Diego Soto-Hernández, García-Manzo, G. (2017). Análisis Bayesiano del Gobierno Electrónico en México mediante el Algoritmo K2.

[10] Correa, M., Bielza, C., Paimes-Teixeira, J., Alique, J. R. (2009).Comparison of Bayesian networks and artificial neural networks

[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] De la Torre-Gea, Guillermo; Ortiz-Vázquez, Irving Cesar, Rico-García, Enrique; Guevara-González, Ramon; Morales-Vázquez, Karla Adriana, Rodriguez-Mendez, Lilia Irene; Fernandez-Loyola, Rodrigo; Pérez-

Robles, Juan Francisco. (2016). Bayesian Analysis of a CFD Model to the Absorption of Ammonium Concentrations in Zeolite. Journal of Scientific and Engineering Research, 3(4); 258-266.

[14] Ortíz-Vázquez, Irving Cesar; Rodriguez-Mendez, Lilia Irene; Soto-Zarazúa, Genaro; Pérez-Robles, Juan Francisco; Rico-García, Enrique;. And De La Torre-Gea, Guillermo Alfonso. (2016). Analysis of large commercial greenhouses in warm climates using cfd and bayesian networks. Journal of Global Ecology and Environmen, 5(2); 2454-2644.