Expert Systems with Applications 40 (2013) 6728–6734

Contents lists available at SciVerse ScienceDirect

Expert Systems with Applications

journal homepage: www.elsevier.com/locate/eswa

Policy making for broadband adoption and usage in Chile through machine learning

Gonzalo A. Ruz *, Samuel Varas, Marcelo Villena

Facultad de Ingeniería y Ciencias, Universidad Adolfo Ibáñez, Av. Diagonal Las Torres 2640, Peñalolén, Santiago, Chile

ARTICLE INFO

Keywords:
Broadband penetration
Policy making
Clustering analysis
Bayesian networks

ABSTRACT

For developing countries, such as Chile, we study the influential factors for adoption and usage of broadband services. In particular, subsidies on the broadband price are analyzed to see if this initiative has a significant effect in the broadband penetration. To carry out this study, machine learning techniques are used to identify different household profiles using the data obtained from a survey on access, use, and users of broadband Internet from Chile. Different policies are proposed for each group found, which were then evaluated empirically through Bayesian networks. Results show that an unconditional subsidy for the Internet price does not seem to be very appropriate for everyone since it is only significant for some households groups. The evaluation using Bayesian networks showed that other polices should be considered as well such as the incorporation of computers, Internet applications development, and digital literacy training.

1. Introduction

Stimulus for broadband adoption and usage can be achieved through policies from the supply-side or the demand-side. Most nations have ongoing programs for broadband roll out and the lack of availability nowadays is not reported as an influential factor in the adoption of broadband. For example in Rappoport, Kridel, Taylor, Duffy-Deno, and Alleman (2002), factors such as the household income and the level of education are good predictors for the acquisition of broadband services. In the case of Mexico (García-Murillo & Rendón, 2009) it has been found that the most important factor is the household income. For China, factors such as income, education, and penetration of fixed phones are relevant (Nam, Kim, Lee, & Duan, 2009). The presence of computers in homes is also a key factor as reported in Stanton (2004), and demographic factors such as gender, age, and education level of the parents as well (Oh, Ahn, & Kim, 2003).

From a demand-side point of view, recent reviews show that there have been more than 400 initiatives, most of these can be grouped in four areas (Hauge & Prieger, 2010): programs to mitigate price, programs to mitigate the lack of computer ownerships, programs to mitigate lack of digital literacy, and programs to mitigate perceived lack of value. For developing countries, such as Chile which presents a heterogeneous socio–economic population, it is not clear how the price factor influences in the acquisition of broadband services and how important is this factor, and if it is relevant for all the population.

Broadband Internet access is still not considered a necessity good in developing countries, although it has become more and more important. Due to this fact, the majority of households in these countries will destine most of their monthly income to pay for food, house payments (rents or mortgages), health insurance, childrens schools, etc. There have been recent studies which show that the development of Internet services is of great importance to a society due to its positive effects on economic growth (Katz & Suter, 2009; Crandall, Lehr, & Litan, 2007) and employment generation (Katz, 2009).

With this in mind, the governments of developing countries are keen on increasing broadband penetration rates. According to the latest OECD Broadband Portal (2012), Chile's fixed (wired) broadband subscriptions, per 100 inhabitants, is 12.2. Whereas, Chile's terrestrial mobile wireless broadband subscriptions, per 100 inhabitants, is 22.4. These rates are one of the highest in Latin America, nevertheless, when compared to countries of the OECD, they are one of the lowest, considering that the OECD has a penetration rate average of 26 for fixed broadband and 55 for mobile wireless broadband. This is an important matter since Chile became a member of the OECD in 2010, and therefore needs to improve its Internet penetration rate.

A straightforward hypothesis is that an unconditional broadband subsidiary campaign for everyone, can have a high impact in improving the broadband penetration rate. In this work, we try to empirically study if this is the right strategy as well as suggest other alternatives, which could favorably influence the broadband penetration.

Policy making through the use of machine learning seems to be a promising area, where decision-making can be carried out using

* Corresponding author. Tel.: +56 2 23311200.
E-mail address: gonzalo.ruz@uai.cl (G.A. Ruz).

0957-4174/$ - see front matter © 2013 Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.eswa.2013.06.039

historical information (data) in a multivariate way. This enables policy makers to take into account several aspects simultaneously, rather than designing policies based on individual dimensions or factors. Also, the ability to evaluate and predict the impact of the policies is straight forward through the models generated by machine learning techniques, as seen further in this paper. Examples of the use of machine learning techniques for policy making are as follows. In Kontogianni, Papageorgiou, and Tourkolias (2012), Fuzzy Cognitive Mapping are proposed as a supporting tool, for environmental policy makers, in the areas of participatory environmental scenario development, subjective risk analysis, and stated preference approaches in environmental valuation. Argumentation-based decision models were used in Bourguet, Thomopoulos, Mugnier, and Abecassis (2013) for the analysis of food quality in a public health policy. The use of self-organizing maps for devising future air pollution policies in Taiwan is presented in Li and Shue (2004). In Turkey, Kahraman and Kaya (2010) present a fuzzy multicriteria decision-making methodology for the selection among energy policies and Cinar and Kayakutlu (2010) for creating scenarios for energy policies using Bayesian networks.

The rest of the paper is organized as follows. Section 2 presents a brief description of the machine learning techniques used in the policy making process. Section 3 presents a clustering analysis to identify the profiles of households with no Internet in their homes. In Section 4, intra-cluster differences with the households that do have Internet in their homes is analyzed. Based on the results of Sections 3 and 4, policies are proposed in Section 5 to increment the broadband penetration rate. The impact of the policies for each cluster is empirically evaluated in Section 6. Discussion and final conclusions of this work are presented in Sections 7 and 8 respectively.

2. Background

To conduct this study, two machine learning techniques are used. First, the k-means algorithm is used for data clustering, then Bayesian networks for data classification. A description of these techniques is given below.

2.1. k-means

The well-known k-means clustering algorithm is used to discover natural groupings of a data set. The algorithm is as follows (Tan, Steinbach, & Kumar, 2005),

1. Select (randomly) k points as the initial centroids
2. **repeat**
3. Form k clusters by assigning all points to the closest centroid.
4. Recompute the centroid of each cluster.
5. **until** The centroids do not change.

One of the drawbacks of most clustering techniques, including the k-means, is that the number of clusters k must be specified a priori. To overcome this problem, several methods for selecting automatically the most plausible number of clusters have been developed. In this paper, we select k by computing a cluster validity function proposed by Pham, Dimov, and Nguyen (2005). This function is defined by,

$$f(k) = \begin{cases} 1 & \text{if } k = 1 \\ \frac{S_k}{\alpha_k S_{k-1}} & \text{if } S_{k-1} \neq 0, \quad \forall k > 1 \\ 1 & \text{if } S_{k-1} = 0, \quad \forall k > 1 \end{cases} \tag{1}$$

with

$$\alpha_k = \begin{cases} 1 - \frac{3}{4N_d} & \text{if } k = 2 \text{ and } N_d > 1 \\ \alpha_{k-1} + \frac{1-\alpha_{k-1}}{6} & \text{if } k > 2 \text{ and } N_d > 1 \end{cases} \tag{2}$$

where S_k is the sum of the cluster distortions (i.e., sum of the squared error between each center of the clusters and all the data points of that same cluster) when the number of clusters is k, N_d is the number of variables (i.e., the number of dimensions) and α_k is a weight factor. Basically, this function assumes that the data distribution is uniform; therefore, the cluster distortion for k clusters can be estimated with $k-1$ clusters. The ratio between these two measures will be close to one when no dense data regions are found when k is increased. If dense regions are detected, then S_k will be less than the estimated value $\alpha_k S_{k-1}$, therefore, $f(k)$ decreases. The idea is to find the k values that generate the smallest values of $f(k)$, since they can be considered as well-defined clusters.

2.2. Bayesian networks

Bayesian networks (BN) were introduced by Pearl (1988) they consist of two parts. A qualitative part that is a direct acyclic graph (DAG) where each node represents a discrete random variable and the edges represent probabilistic dependencies. The quantitative part is a conditional probability table, one for each node, which contains the conditional probability of the node conditioned to its parent nodes in the DAG. A key feature of a BN is that it satisfies the *Markov condition*, which states that every variable is conditionally independent of its nondescendents, given the set of its parents.

By combining both parts, a BN encodes the joint probability distribution. Given a BN of n random variables $\mathbf{X} = (X_1, \ldots, X_n)$ the joint probability distribution is computed by

$$P(\mathbf{X}) = \prod_{i=1}^{n} P(X_i | \Pi_{X_i}) \tag{3}$$

where Π_{X_i} represents the set of parent nodes (variables) in the DAG. One of the difficulties is that learning BNs from data is a difficult problem, in fact it has been shown that it is NP-complete (Chickering, 1996), nevertheless, there are many heuristics and approximations which makes this problem computationally tractable (Cooper & Herskovits, 1992; Heckerman, Geiger, & Chickering, 1995).

2.2.1. Bayesian networks for classification

In a classification task, a database with historical data is used to train a classifier to predict the outcome of a new example. The data consists in a matrix where each column represents an attribute (random variable) and each row a data example. The final column in the matrix contains a class label. The training process consists in using the data matrix to adjust the parameters of the classifier in order to reduce the error between the real output (class label) and the output of the classifier. There are many classifiers, such as, C4.5 decision tree, artificial neural networks, k-nn, support vector machines, etc.

One approach is using a probabilistic classifier, in this case, the class value for an example is found such that it maximizes the posterior probability of the class for a given set of assignments to the attributes. In other words, the class value for the hth example of the dataset, $X_1 = x_1^h, \ldots, X_n = x_n^h$, can be computed as

$$class_{value}(X_1 = x_1^h, \ldots, X_n = x_n^h) = \arg\max_k P(C = k | X_1 = x_1^h, \ldots, X_n = x_n^h). \tag{4}$$

The posterior probability can be computed using Bayes' theorem

$$P(C = k | X_1, \ldots, X_n) = \frac{P(C = k) P(X_1, \ldots, X_n | C = k)}{\sum_{k'} P(C = k') P(X_1, \ldots, X_n | C = k')}. \tag{5}$$

The denominator of the r.h.s. of (5) is constant with respect to the class and can be expressed as $1/\alpha$. So we only need to worry about how to compute the numerator, which is in fact the joint probability distribution of all the attributes and the class variable. This can be carried out using a BN with certain assumptions. The simplest approach is to consider that each attribute is conditionally independent of every other attribute. This rather 'naive' assumption yields the well-known naive Bayesian classifier (Duda & Hart, 1973),

$$P(C = k|X_1, \ldots, X_n) = \alpha P(C)\prod_{i=1}^{n} P(X_i|C). \tag{6}$$

Although this classifier has shown to perform rather well considering the strong independence assumption (Langley, Iba, & Thompson, 1992), when there are strong correlations amongst the attributes the naive Bayes classification performance is affected. To overcome this problem, a variant of the naive Bayes classifier, called the *Tree Augmented Naive Bayes classifier* (TAN), was introduced by Friedman, Geiger, and Goldszmidt (1997). In this model, each attribute has as a parent the class variable C and at the most, one other attribute, yielding tree structures that have $n - 1$ edges (without counting the edges from C to every attribute).

The first step to construct the TAN classifier is to compute the *conditional mutual information* between pairs of attributes, conditioned by the class variable. This measure is defined by

$$I(X;Y|Z) = \sum_{x\in X}\sum_{y\in Y}\sum_{z\in Z} P(x,y,z)\log\frac{P(x,y|z)}{P(x|z)P(y|z)}. \tag{7}$$

The conditional mutual information measures the information that Y provides about X when the value of Z is known.

The following step is to build a complete undirected graph. This is carried out by connecting an edge from each node (attribute) to every other node and assigning the weight of the edge that connects X_i with X_j by $I(X_i;X_j|C)$. Next, in order to obtain a tree structure, the *maximum weighted spanning tree* (MWST) is built using any well-known MWST procedure, such as Kruskal's algorithm (Kruskal, 1956). Finally, directions to the edges of the resulting tree can be added by choosing any attribute as the root and then setting the directions of all the edges to be pointing outwards from it.

Given that each attribute will have $\Pi_{X_i} = \{X_{j\neq i}, C\}$, except for the root attribute node that will have $\Pi_{X_i} = \{C\}$, the TAN classifier can be expressed as

$$P(C|X_1, \ldots, X_n) = \alpha P(C)\prod_{i=1}^{n} P(X_i|\Pi_{X_i}). \tag{8}$$

A summary of the learning algorithm for this type of Bayesian network is as follows.

1. Compute the conditional mutual information $I(X_i;X_j|C)$ between each pair of attributes $i \neq j$.
2. Build a complete undirected graph using the attributes as nodes and assign the weight of the edge that connects X_i to X_j by $I(X_i;X_j|C)$.
3. Apply the MWST algorithm.
4. Choose an attribute to be root and set the directions of all the edges to be outward from it.
5. Add a vertex node C and add an edge from C to every other attribute X_i.

3. Characterization of households with no Internet in Chile

During 2009, the Chilean Subsecretary of Telecommunications (Subtel) presented the study Survey on access, use, and users of broadband Internet in Chile (Subtel, 2009). This work consisted in carrying out a survey on the households in four regions of the country: Antofagasta, Valparaíso, Biobío, and Metropolitana. These regions concentrate approximately 65% of the total population of Chile. As part of the results from this survey, the following key factors were identified as been influential, for the households, in the decision of having or not broadband Internet in their homes: *income level in their homes*, *presence of children in school*, *gender*, *age*, *level of education*, *user or not of computers*, and *user or not of Internet*.

Table 1
Variables used in the clustering process to identify different household profiles.

Name	Description	Type
Family	Number of family members that live in the house	Integer
Sex	Male or female	Binary
Age	Age of household	Integer
Education	Education level (primary, high school, university)	Integer
Computer	Knows how to use a computer	Binary
Internet	Knows how to use the Internet	Binary
Income	Total income in the house measured as quintiles	Integer
Marital status	Marital status (single, married, divorced)	Integer

With this information, a cluster analysis can be used to identify different profiles of households, which do not have broadband Internet in their homes. There are 888 households with no Internet from the survey, each of them, is characterized by eight variables shown in Table 1.

The data mining open source package Weka (Hall et al., 2009) was used to carry out the k-means clustering with $k = 1, \ldots, 10$. For each k, ten independent runs were carried out, recording the clustering result with the lowest S_k. Then for each k, the $f(k)$ function was computed obtaining the results shown in Fig. 1.

From Fig. 1 we see that the most plausible values for k are 2, 4, and 9 (in that order). Given that these groups will be used to design policies, $k = 2$ partitions the data into two groups which are quite general, whereas $k = 9$ partitions the data into 9 groups which are too specific. Therefore, $k = 4$ was chosen for this study, which is a good compromise between the two results. The description of each cluster found is as follows.

3.1. Characterization of cluster 1

This group (C1) if formed by 317 female households, which are 55 years old in average and most of them are married. They have primary school education completed. They do not know how to use a computer nor the Internet. The income level in their homes corresponds to the second quintile and each home has 3.7 family members in average.

3.2. Characterization of cluster 2

This group (C2) if formed by 180 male households, which are 42 years old in average and most of them are married. They have high school education completed. They know how to use a computer and the Internet. The income level in their homes corresponds to the third quintile and each home has 3.6 family members in average.

3.3. Characterization of cluster 3

This group (C3) if formed by 238 male households, which are 55 years old in average and most of them are married. They have primary school education completed. They do not know how to use a computer nor the Internet. The income level in their homes corresponds to the second quintile and each home has 3.5 family members in average.

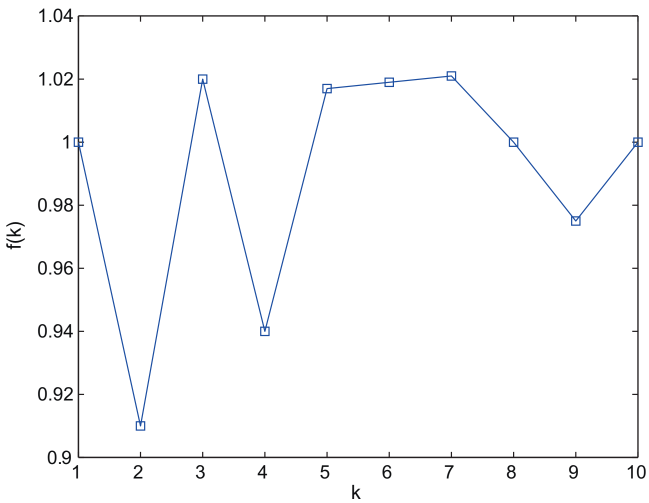

Fig. 1. Selection of k for the k-means. Lower values of $f(k)$, represent plausible values for k.

3.4. Characterization of cluster 4

This group (C4) if formed by 153 female households, which are 40 years old in average and most of them are married. They have high school education completed. They know how to use a computer and most of them know how to use the Internet. The income level in their homes corresponds to the second quintile and each home has 3.8 family members in average.

A summary of the four clusters appears in Table 2.

3.5. Accelerators and inhibitors for each cluster

Once the clusters are identified, we can analyze the answers that the household gave to the questions in the survey related to accelerators, which could drive a household to use the services of broadband Internet and the inhibitors which are currently influencing in the decision of not hiring broadband services. The accelerators for each cluster appears in Table 3, the inhibitors for each cluster related to the main reason for not having Internet at home appears in Table 4, and the main reasons why they have not used Internet in Table 5.

4. Intra-cluster differences between households with and without Internet

The four clusters described in the previous section were found using 888 households that reported that they did not have Internet in their homes. Additionally, there are 422 households which have Internet in their homes. Each of these households was characterized by the same eight variables described in Table 1, then, they were assigned to the closest cluster, that is, the cluster which presented the minimum distance between the cluster centroid and the household vector. With this procedure, 47 households were assigned to C1, 197 to C2, 32 to C3, and 146 to C4.

When analyzing the within differences amongst the eight variables for each cluster for households with and without Internet, we find that the two main differences in C2 and C4 are the income level and the level of education, been higher in both variables for the households with Internet in their homes, for the other remaining six variables there are no significant differences. For C1 and C3 the variables that are different are the income level and the number of family members, again been higher in both variables for households with Internet, for the other remaining six variables there are no significant differences.

In order to find additional reasons why households in a same cluster have or do not have Internet in their homes, 41 variables (questions) were selected from the survey (excluding the eight variables used to form the clusters) related to the use of Internet applications, Internet price, and access to technology. The decision tree called J48 (Weka's implementation of C4.5 algorithm (Quinlan, 1993)) was used to automatically select the most relevant variables to differentiate each example (household) in one of the two situations (with or without Internet). The identified relevant variables, per cluster, in descending order of importance, are as follows.

4.1. Variables that show intra-cluster differences between households with or without Internet in C1

(i) Do you have a computer desktop at home? (ii) Do you have a fixed line telephone? (iii) Do you have cable TV at home? (iv) How much is the maximum price you are willing to pay for broadband Internet at your home?

4.2. Variables that show intra-cluster differences between households with or without Internet in C2

Selected variables: (i) Do you use Internet at home or somewhere else? (ii) Do you have a computer desktop at home? (iii) Do you have a fixed line telephone? (iv) How many children do you have in scholar age?.

4.3. Variables that show intra-cluster differences between households with or without Internet in C3

Selected variables: (i) Do you have a computer desktop at home? (ii) Do you have cable TV at home?.

Table 2
Four different household profiles identified using k-means.

Name	Cluster1			Cluster2			Cluster3			Cluster4		
	Mean	S.D.	Mode	Mean	S.D.	Mode	Mean	S.D.	Mode	Mean	S.D.	Mode
Family	3.7	1.9	3	3.6	1.5	4	3.5	1.7	2	3.8	1.6	3
Sex	F	–	F	M	–	M	M	–	M	F	–	F
Age	55.5	14.9	–	42.1	12.7	–	55.4	14.7	–	40.1	12.7	–
Education	Prim	–	Prim	High	–	High	Prim	–	Prim	High	–	High
Computer	No	–	No	Yes	–	Yes	No	–	No	Yes	–	Yes
Internet	No	–	No	Yes	–	Yes	No	–	No	Yes	–	Yes
Income	2Q	–	–	3Q	–	–	2Q	–	–	2Q	–	–
Marital status	–	–	Mar	–	–	Mar	–	–	Mar	–	–	Mar

F = female, M = male, Prim = primary school education, High = high school education, 2Q = second quintile income level, 3Q = third quintile income level, and Mar = married.

Table 3
Actions which households perceive as accelerators to use Internet.

Action	Cluster 1(%)	Cluster 2(%)	Cluster 3(%)	Cluster 4(%)
None	47.3	16.1	45.8	17.6
Access to training	21.5	21.1	23.9	15.0
Cheaper desktop and laptops	11.7	13.3	12.2	15.7
Cheaper Internet access	11.0	35.6	6.7	32.0
Nearby free Internet access points	5.0	7.8	6.7	12.4
Nearby public Internet access points	1.9	3.3	1.3	3.2

Table 4
Inhibitors: The main reason for not having Internet at home for each cluster.

Reason	Cluster 1(%)	Cluster 2(%)	Cluster 3(%)	Cluster 4(%)
I do not have a computer	46.7	22.2	44.1	33.3
It is too expensive	24.3	32.8	24.8	29.4
I do not need it for now	9.5	11.1	7.9	5.2
I do not know how to use it	6.6	–	9.2	–
I am not interested for now	3.5	2.8	3.7	3.2
No Internet service provided where I live	0.3	1.7	0.8	1.3

Table 5
Inhibitors: The main reasons why the household has not used the Internet.

Reason	Cluster 1(%)	Cluster 2(%)	Cluster 3(%)	Cluster 4(%)
I do not know how to use it	57.7	7.8	50.4	11.1
I am not interest for now	17.4	13.3	24.8	9.8
I do not need it for now	12.9	13.9	14.7	14.4
I do not know for what it is useful for	5.4	–	–	–

4.4. Variables that show intra-cluster differences between households with or without Internet in C4

Selected variables: (i) Do you have a fixed line telephone? (ii) Do you have a computer desktop at home? (iii) Do you have a laptop at home? (iv) Do you have cable TV at home? (v) Do you use Internet at home or somewhere else? (vi) How often do you, or someone else for you, obtain certificates (civil registry, others) via Internet?

5. Proposed policies

To design the policies, the results from Sections 3 and 4: the accelerators and inhibitors for each cluster as well as the intra-cluster differences between households with and without Internet was taken into consideration. With this in mind, four areas have been identified:

1. Promotion of computers (desktops or laptops) incorporation.
2. Internet price subsidies.
3. Digital alphabetization and the promotion of usage.
4. Development of Internet applications.

It is important to point out that these policies are in accordance with recent experiences reported by Turk, Blazic, and Trkman (2008) for the case of EU countries, and Bouras, Giannaka, and Tsiatsos (2009) for OECD countries. Of course, each of these policies has a different impact and importance for each identified cluster. According to the results found in the previous sections, the order of importance (1 = very important, 2 = important, 3 = less important) in each cluster could be assigned as shown in Table 6.

6. Evaluation

To validate the effects that these policies would have on the households with no Internet, as well as to compare the level of importance assigned, in a qualitative manner, in Table 6 versus a more empirical quantitative manner, Bayesian network classifiers were constructed for each cluster. Specifically, the TAN classifier was constructed for each cluster to predict if a household has Internet or not at their homes. The prediction of the class variable C, which is 1 if the household has Internet or 0 if it does not, is based on the following variables, which represent the proposed policies mentioned before:

- Computer: desktop or laptop incorporation.
- Price: Internet price subsidies (10USD or 20USD).
- Heath: Internet applications (related to health).
- Training: Digital alphabetization and promotion of usage.
- Paperwork: Internet applications (related to paperwork, e.g. civil registry certificates).
- Income: total income in a home.

The last variable, Income, is used to measure the effect of broadband penetration due to the nations economic growth.

Once the TAN models were constructed for each cluster (using the Weka open source data mining package), we proceeded to measure the effect of each policy, independently, on households with C = 0, by changing the original values of the variables as a consequence of a policy. For example, in a specific cluster, to measure the effect of the incorporation of computers, households with Computer = 0 were changed to Computer = 1, then the output of the network (the prediction) was recorded and compared to its original situation, identifying the households that changed from C = 0 to C = 1 due to the effect of changing the Computer variable to 1. This procedure was carried out for each policy (variable), and each cluster.

To measure the performance of the classifiers, *n-fold cross validation* (Witten, Frank, & Hall, 2011) was used. In n-fold cross validation, the original data set is randomly partitioned in n equally sized groups. Then, $n-1$ partitions are used to train the classifier, and the remaining partition is used for testing. This process is repeated n times, so that each partition is used as a test set once.

Table 6
Policies per cluster and their level of importance.

Policies	Cluster 1	Cluster 2	Cluster 3	Cluster 4
Promotion of computers incorporation	(1)		(2)	(2)
Internet price subsidies	(1)	(2)	(3)	(3)
Digital alphabetization and the promotion of usage	(2)		(1)	
Development of Internet applications	(3)	(1)		(1)

(1)=very important, (2)=important, (3)=less important.

The correct classification result, on the test set, of each process is averaged to obtain a final estimation of the performance of the classifier. We use $n = 10$, as suggested in Witten et al. (2011).

7. Evaluation results and discussions

The resulting Bayesian network classifiers for each cluster using the TAN algorithm is shown in Fig. 2. Fig. 2(a) shows the Bayesian network structure for C1. As discussed in Section 2, in the TAN model each variable is dependent on the class variable C, also the following probabilistic dependencies can be visualized: Training depends of Price, Computer depends of Health, Price depends of Health, Paperwork depends of Health, Health depends of Income, and Income is the root and only depends of C. The structure of the network and the conditional probability tables learned from the data are used to compute the posterior probability (8) to carry out the classification of each example in the cluster. The performance of this classifier, using 10-fold cross validation is 91% of correct classifications.

Fig. 2(b) shows the resulting network for C2. The probabilistic dependencies for this cluster are as follows: Computer depends of Health, Training depends of Health, Health depends of Paperwork, Paperwork depends of Income, Income depends of Price, and Price is the root and only depends of C. The performance of this classifier, using 10-fold cross validation is 78% of correct classifications.

Fig. 2(c) shows the resulting network for C3. The probabilistic dependencies for this cluster are as follows: Paperwork depends of Health, Health depends of Training, Training depends of Price, Computers depends of Price, Price depends of Income, and Income is the root and only depends of C. The performance of this classifier, using 10-fold cross validation is 91% of correct classifications.

Finally, Fig. 2(d) shows the resulting network for C4. The probabilistic dependencies for this cluster are as follows: Computers depends of Paperwork, Paperwork depends of Health, Training depends of Health, Health depends of Income, Income depends of Price, and Price is the root and only depends of C. The performance of this classifier, using 10-fold cross validation is 80% of correct classifications.

The general structures of the Bayesian networks for each cluster are different; this gives a first insight that the policies affect in a different way for each group, although, if each variable is analyzed, we notice that some dependencies are repeated in some clusters. For example the Computer variable is dependent of the Health variable, given C, in C1 and C2, whereas, in C3 the Computer variable depends on the Price, given C, and in C4 the Computer variable depends on Paperwork, given C.

The impact of each policy, in each cluster, is summarized in Table 7. An interesting result that is observed straightaway is that, although the clustering results show that C1 and C3 are similar (the same for C2 and C4), the main difference is the gender (one has female households, the other male households), we notice that the policies affect differently in each cluster.

Further analysis of Table 7 shows that the hypothesis of an unconditional broadband subsidiary campaign for everyone would only have an important effect in C4, then in a less degree in C3. The Internet price subsidies have little effect in C2, and practically none in C1. The incorporation of computers impacts significantly in C1, then in C2, and in a less degree in C4. Applications, such as health related impacts in C4 and C3. It has little effect in C2, and has negative effect in C1 (remember this group does not know how to use Internet and is not interested).

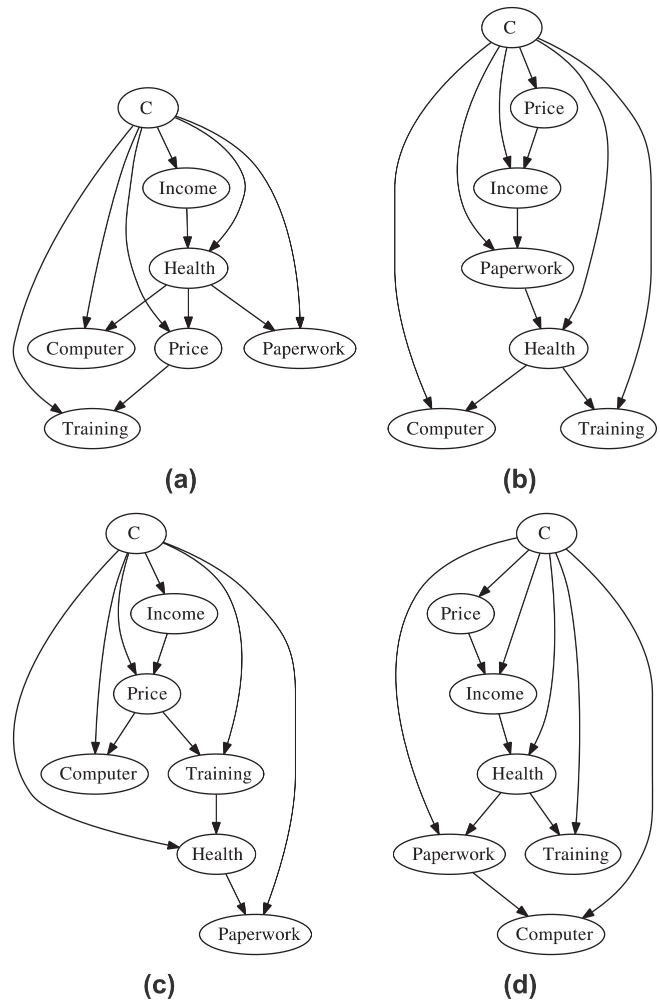

Fig. 2. TAN classifier representation for (a) cluster 1, (b) cluster 2, (c) cluster 3, and (d) cluster 4.

Applications related to paperwork have a positive impact in all the clusters, most significantly in C4. Digital alphabetization (Training) has a negative effect in C2 and C4. This is normal, since the households in these clusters already know how to use computers and the Internet.

For C3, training also has a negative effect since this group does not know how to use a computer nor the Internet and is motivated first to see useful applications before they move towards training. C1 is the only cluster which training could have a positive effect,

Table 7
Impact of the policies: Increase % of households with Internet per cluster.

Policies	Cluster 1(%)	Cluster 2(%)	Cluster 3(%)	Cluster 4(%)
Computers	43	17	6	11
Health	−12	1	11	19
Paperwork	6	5	11	20
Training	6	−42	−3	−20
Income	0.2	0.2	0	0
Subsidies for 10 USD	0.5	11	3	30
Subsidies for 20 USD	0.5	21	7	31
TAN's correct classification[a]	91	78	91	80

[a] Using 10-fold cross validation.

although not very high, compared to the incorporation of computers. The case where no policy is carried out, therefore hoping that the country's economic growth would allow to improve the households income, thus, increase the Internet penetration rate, is considered in the Income variable. For this, an increase of 3% in the households' income for 10 years was evaluated in each model. The results show that this income increase has hardly any effect in the increase of households with Internet.

8. Conclusions

Public policies to develop broadband and Internet penetration for developing countries, such as Chile, must consider aspects related to the different household types, because the available household income is not the only reason why they do not adopt such services. We have found other household factors which vary in importance depending of the type of household, which can be used to explain the possibility of broadband and Internet adoption. The household factors found are digital literacy (technology knowledge), income, age, sex, number of family members and education level. Using these factors (variables), the *k*-means clustering algorithm was used on data corresponding to households with no broadband and Internet services. Several numbers of clusters were tested. Finally, it was found that four clusters were appropriate for this study.

Once each cluster was characterized, households that do have Internet services were assigned to each cluster by minimum distances to the clusters centroids. We compared within each cluster the households with and without broadband and Internet service. The main variables that explain the differences in each cluster are the availability of computers, the knowledge in use of computers and Internet, and the broadband and Internet service price.

Based on such findings we developed a series of policies, some cross-sectional and others focused specifically for the different clusters. We show that each policy has a different impact depending on the household profile. An unconditional subsidy for the Internet price does not seem to be very appropriate for everyone since it is only significant for some households clusters. Other polices should be considered as well (simultaneously in some cases) such as the incorporation of computers, Internet applications development, and digital training.

The methodology used in this work can be followed to analyze data from other developing countries in order to design and evaluate policies to increase the adoption and usage of the Internet. Also, all the machine learning techniques employed in the analysis are available in the open source software Weka.

Acknowledgements

The authors thank the research assistants Mauricio Valle and Cristóbal Giadach for their contributions to this work and the Chilean Subsecretary of Telecommunications (Subtel) for financially supporting this research.

References

Bourguet, J., Thomopoulos, R., Mugnier, M., & Abecassis, J. (2013). An artificial intelligence-based approach to deal with argumentation applied to food quality in a public health policy. *Expert Systems with Applications, 40*, 4539–4546.
Chickering, D. M. (1996). Learning bayesian networks is NP-complete. In D. Fisher & A. Lenz (Eds.), *Learning from data* (pp. 121–130). New York: Springer-Verlag.
Cinar, D., & Kayakutlu, G. (2010). Scenario analysis using Bayesian networks: A case study in energy sector. *Knowledge-Based Systems, 23*, 267–276.
Cooper, G. F., & Herskovits, E. (1992). A Bayesian method for the induction of probabilistic networks from data. *Machine Learning, 9*, 309–347.
Crandall, R., Lehr, W., & Litan, R. (2007). The effects of broadband deployment on output and employment: A cross-sectional analysis of U.S. data.. *Issues in Economic Policy (The Brookings Institute)*, 1–34.
Duda, R., & Hart, P. (1973). *Pattern classification and scene analysis*. New York: John Wiley & Sons.
Friedman, N., Geiger, D., & Goldszmidt, M. (1997). Bayesian network classifiers. *Machine Learning, 29*, 131–163.
García-Murillo, M., & Rendón, J. (2009). A model of wireless broadband diffusion in latin america. *Telematics and Informatics, 26*, 259–269.
Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., & Witten, I. H. (2009). The WEKA data mining software: An update. *SIGKDD Explorations, 11*, 10–18.
Hauge, J., & Prieger, J. (2010). Demand-side programs to stimulate adoption of broadband: What works? *Review of Network Economics, 9*, 1–38 [Article 4].
Heckerman, D., Geiger, D., & Chickering, D. M. (1995). Learning Bayesian networks: The combination of knowledge and statistical data. *Machine Learning, 20*, 197–243.
Kahraman, C., & Kaya, I. (2010). A fuzzy multicriteria methodology for selection among energy alternatives. *Expert Systems with Applications, 37*, 6270–6281.
Katz, R. (2009). The impact of the broadband policy framework on jobs and the economy. *The Parliament, 293*, 2–3.
Katz, R., & Suter, S. (2009). Estimating the economic impact of the broadband stimulus plan. Working Paper, Columbia Institute for Tele-Information.
Kontogianni, A. D., Papageorgiou, E. I., & Tourkolias, C. (2012). How do you perceive environmental change? fuzzy cognitive mapping informing stakeholder analysis for environmental policy making and non-market valuation. *Applied Soft Computing, 12*, 3725–3735.
Kruskal, J. B. Jr., (1956). On the shortest spanning subtree of a graph and the traveling salesman problem. *Proceedings of the American Mathematical Society, 7*, 48–50.
Langley, P., Iba, W., & Thompson, K. (1992). An analysis of Bayesian classifiers. In *Proceedings of the 10th national conference on artificial intelligence (AAAI'92)* (pp. 223–228).
Li, S., & Shue, L. (2004). Data mining to aid policy making in air pollution management. *Expert Systems with Applications, 27*, 331–340.
Nam, C., Kim, S., Lee, H., & Duan, B. (2009). Examining the influencing factors and the most efficient point of broadband adoption in china. *Journal of Research and Practice in Information Technology, 41*, 25–38.
OECD Broadband Portal. (2012). Fixed and wireless broadband subscriptions per 100 inhabitants (june 2012). <http://www.oecd.org/sti/broadband/oecdbroadbandportal.htm>.
Oh, S., Ahn, J., & Kim, B. (2003). Adoption of broadband internet in Korea: The role of experience in building attitudes. *Journal of Information Technology, 18*, 267–280.
Pearl, J. (1988). *Probabilistic reasoning in intelligent systems: Networks of plausible inference*. San Francisco, CA: Morgan Kaufmann.
Pham, D. T., Dimov, S. S., & Nguyen, C. D. (2005). Selection of k in k-means clustering. *Proceedings of the Institution of Mechanical Engineers, Part C: Journal of Mechanical Engineering Science, 219*, 395–412.
Quinlan, J. R. (1993). *C4.5: Programs for machine learning*. Morgan Kaufmann.
Rappoport, P., Kridel, D., Taylor, L., Duffy-Deno, K., & Alleman, J. (2002). *Residential demand for access to the Internet. The International Handbook of Telecommunication Economics* (Vol. II). Cheltenham: Edward Elgar Publishers.
Stanton, L. (2004). Factors influencing the adoption of residential broadband connections to the internet. In *Proceedings of the 37th annual Hawaii international conference on system sciences (HICSS'04)* (p. 50128a).
Tan, P., Steinbach, M., & Kumar, V. (2005). *Introduction to data mining*. Addison-Wesley.
Witten, I. H., Frank, E., & Hall, M. (2011). Data mining: Practical machine learning tools and techniques. *3rd ed.*. Morgan Kaufmann.