Using a New Tool to Visualize Environmental Data for Bayesian Network Modelling

R.F. Ropero[1(✉)], Ann E. Nicholson[2], and Kevin Korb[2]

[1] Informatics and Environment Laboratory, Department of Biology and Geology, University of Almería, Almería, Spain
rosa.ropero@ual.es
[2] Faculty of Information Technology, Monash University, Melbourne, VIC 3800, Australia

Abstract. This paper presents the software Omnigram Explorer, a visualization tool developed for interactive exploration of relations between variables in a complex system. Its objective is to help users gain an initial knowledge of their data and the relationships between variables. As an example, we apply it to the water reservoir data for Andalusia, Spain. Two Bayesian networks are learned using causal discovery, both with and without the information gleaned from this exploration process, and compared in terms of the Logarithmic loss and causal structure. Even though they show the same predictive accuracy, the initial exploration with Omnigram Explorer supported the use of prior information to achieve a more informative causal structure.

Keywords: Omnigram explorer · Natural complex systems · Bayesian networks · Data visualization · Water management

1 Introduction

Bayesian networks (BNs) are finding rapidly increasing application in Ecology and Environmental Science, modeling complex natural systems [3]. The development process is correspondingly complex, frequently involving extensive expert elicitation processes combined with the collection and automated mining of data from multiple sources (see [1,4,7,8]).

As a part of the process, visualizing the data available to build the models, or again visualizing artificial data generated by the models developed, has an important role. Data visualization can assist in understanding the key relationships between variables in a system, assisting in both model construction and validation. Data visualization done well can also greatly simplify communications with non-expert stakeholders. Some common data visualization approaches include the scatter plots and parallel coordinates [2].

J.M. Puerta et al. (Eds.): CAEPIA 2015, LNAI 9422, pp. 175–184, 2015.
DOI: 10.1007/978-3-319-24598-0_16

Here we use the new visualization tool, Omnigram Explorer[1] (OE) [10], to investigate a BN model of water reservoirs in Andalusia, Spain. OE provides new ways to interact with data sets, allowing for a visual sensitivity analysis of, for example, the effect of causal variables on downstream effect variables under a variety of conditions. The original intention behind OE was that it be an *interactive* tool, coordinating through an API with a BN to explore different initial conditions and their consequences. That is still the intention for the future, but it is currently restricted to working with static data sets, produced by a BN or otherwise, although it will use a BN defined by a Netica (www.norsys.com) dne file in its display, if one is provided. Here we simply illustrate OE's value with static data sets.

2 Omnigram Explorer

OE was designed as a tool for interactive exploration of relations between variables in an agent-based simulation [10]. It draws upon ideas for visualization in the *Attribute Explorer* [9], where data is presented in a set of histograms, one per variable. For more detail information about the data requirements see the link: http://www.tim-taylor.com/omnigram/.

To begin, a data file and model definition are necessary. The data file containing a joint data sample are loaded and presented by OE in a graphical form (Fig. 1(a)). Each variable is represented by a histogram, showing its sample distribution, with a maximum of 20 bins. If a bin is empty (e.g., bin 0 in *Rainfall* node in Fig. 1(a)), a thin horizontal line is drawn at the base. A small circle represents the mean (or, if the user chooses, the median). The range of values is indicated by the horizontal bar under the histogram. The initial histogram represents all the values read from the data file in a plain form, but a subset of them can be highlighted in a *linking and brushing* process (in dark red color).

2.1 OE Interaction Modes

OE allows you to designate some variables as input or outputs and to use Bayesian network links to represent causal structure or other dependencies, as you wish. At present these features are for display only.

The power of this tool lies in its interaction modes, where a variable or subset of variables can be selected and their relation with the remaining variables explored. The selected variables are the "focus" of attention, which is indicated visually by a red square indicator in the corner of the node. Having selected a focus, OE has four different modes of interaction.

[1] Omnigram Explorer is an open-source tool developed in Processing (http://processing.org). The source code, executables (for Windows, Mac and Linux), documentation and related material are available at http://www.tim-taylor.com/omnigram/.

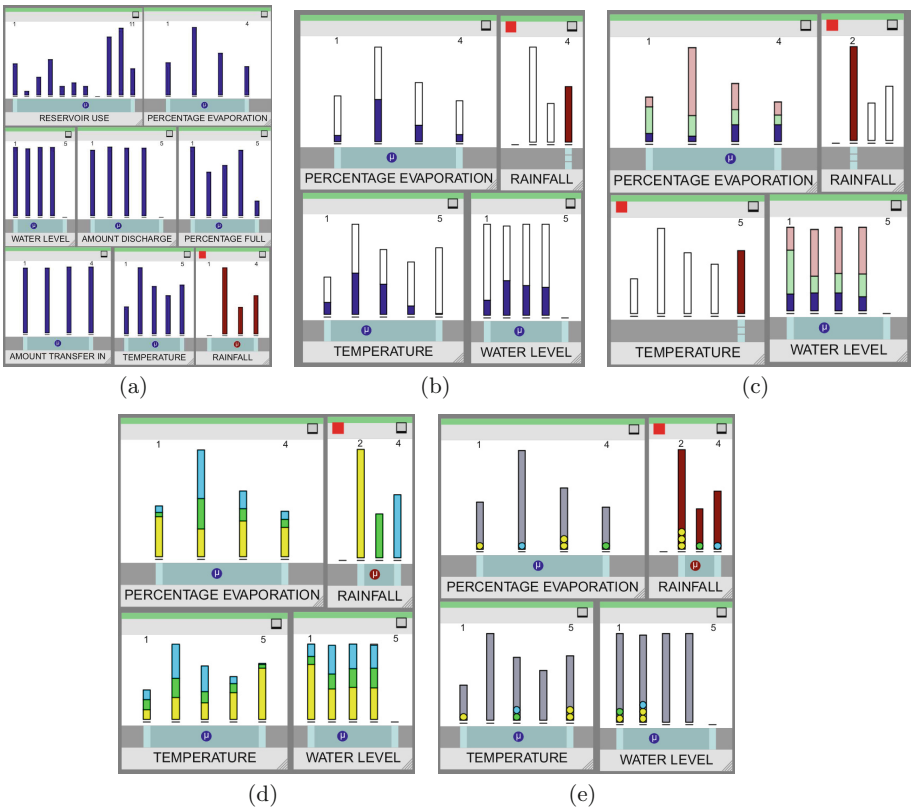


Fig. 1. Initial histograms for the reservoir example with the focus in *Rainfall* variable (a) and modes of interaction in OE for a subset of variables: Single node (b), Multi node (c), Omnibrushing (d) and Sample view (e) (Color figure online).

- Single Node Brushing (Fig. 1(b)), in which only one variable can be in the focus. When a range of values for that node is selected, all of the other variables are updated to show the corresponding sample values in their distributions (represented in dark blue). When changing the focal range, you can simultaneously watch the changes across the other variables, allowing you to intuitively discover the strength of dependencies between the variables. In the example of Fig. 1(b) the focus is on high levels of rainfall (red), and the distributions across other variables conditioned on that high level are displayed in blue.
- Multi Node Brushing (Fig. 1(c)) extends the previous interaction mode, with more than one variable in focus. When two or more variables are selected, OE indicates the ranges selected in red and shows the conditional distributions over other variables in dark blue. Samples which fail match one of the selected ranges are shown in light green; those which match all but two of the ranges are displayed in light red; white displays all other samples. The color, therefore,

shows how close a sample is to matching the conjunctive condition indicated by all the specified ranges in the focal variables. As in Single Node Brushing, the user can interactively change the range of focus nodes and watch the response of the rest of the variables, performing an interactive sensitivity analysis with the sample of the model or data which generated it.
- Omnibrushing (Fig. 1(d)) focuses on a single node. In this case, each focal bin is represented with a different colour. The remaining variables are updated to show for each bin what fraction of the data correspond to the focal bins.
- Sample View (Fig. 1(e)) again uses a single node, and the bins are represented by different colours. The difference is the way data is visualized. Rather than represent a conjunction of corresponding samples, each individual sample is represented itself as a small colored circle, simultaneously across all variables. The display iterates through samples, continuously lighting them up in a sequence. After being lit, a sample will slowly fade as other samples are selected, resulting in a rotating display of subsamples. How quickly new samples are selected and old ones fade is under the user's control.

3 Using OE to Understand Andalusian Reservoirs

3.1 The Problem

A prominent characteristic of the annual water cycle in Andalusia, Spain, is its irregularity. Rainfall alternates between heavy storms and long droughts. Historically, dam construction has been the main solution for both flood control and extending water availability, and there are now more than 1200 dams currently working in Spain. Apart from human water consumption and agriculture, the dams provide a minimum water flow during droughts, allowing biodiversity to be maintained in river beds.

International Panel of Climate Change (IPCC) predicts in Andalusia an increase in the annual average temperature and evaporation, along with a decrease in the rainfall, but with more frequent extreme weather events [5]. These changes will cause severe difficulties in Andalusian water management. New tools and techniques for understanding and managing the reservoirs are urgently needed [11].

3.2 Data Description

For our study we used the Water Quality Dataset from Andalusia (Environmental Information Network of Andalusia) and the National Environmental Statistics (National Government of Spain), from 1999 to 2008. From the reservoirs located in this region of Spain, we selected those that belong to the *Guadaliquivir* and *Guadalete-Barbate* river basin areas, which have no missing values, giving us complete data about 61 dams over the period.

The dataset consists of 6588 samples over the following eight variables (and their states, with continuous variables discretized using expert knowledge and taking into account the distributions of the variables). *Reservoir use* is a discrete

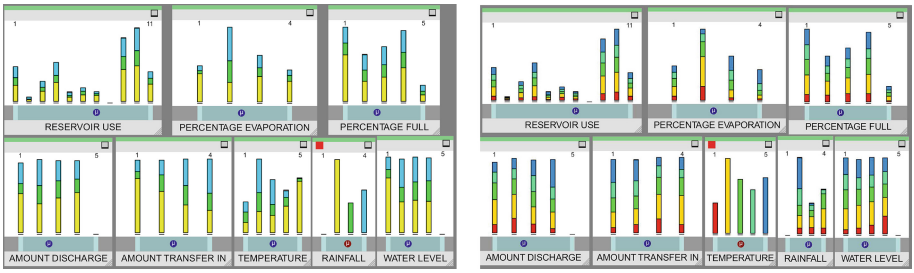
variable with 10 states representing the main use/s of each reservoir classified by the regional Government of Andalusia (1 Hydroelectric; 2 General regulation; 3 Irrigation; 4 Human consumption; 5 Industry; 6 No information; 7 Ecological; 8 Irrigation and other; 9 Irrigation and consumption; 10 Consumption and others). *Temperature* (less than 10; 10–15; 15–20; 20–25 and 25–30°) and *Rainfall* (less than 0.03; 0.03–0.06 and more than 0.06 m^3/m^2) represent the climatic conditions near the dam. *Percentage Evaporation* is the percentage of the reservoir capacity that evaporates (less than 0.029; 0.029–0.035; 0.035–0.93 and more than 0.93 %). *Water level* indicates the height of the water column (less than 92; 92–257; 257–497 and 497–1039 m.a.s.l.) whilst *Percentage Full* expresses the percentage of the reservoir capacity that is currently used (0–25; 25–50; 50–75; 75–100; more than 100 %, during an event of a storm the reservoir can exceed the dam capacity). Finally, dam management is represented by *Amount Discharge* and *Amount Transfer in*. *Amount Discharge* refers to the amount of water that is released when floodgates are opened for ecological, water consumption or regulation purposes (less than 0.13; 0.13–1.36; 1.36–7.08 and more than 7.08 m^3). By contrast, *Amount Transfer in* (less than 0.27; 0.27–1.42; 1.42–6.95 and more than 6.05 m^3) is the amount of water deliberately added to the reservoir, e.g., pumped in from another reservoir.

3.3 OE Data Exploration Prior to Modeling

Here we illustrate the value of OE in initial data exploration, prior to any use of machine learning or BN modeling. Some other tools such as Weka or R software are also available and useful. The aim of this paper is not to compare OE with them, but to present it as an alternative of traditional statistical packages. Figure 1(a) shows the OE view of the data before choosing an interaction mode. *Water Level*, *Amount Transfer in* and *Amount Discharge* present homogeneous distributions, with similar number of samples in each bin.

The goal is to better understand the variables *Percentage Full* and *Water Level* in Andalusia and to predict their behavior under several future scenarios of management decision and those being predicted by the IPCC. So, first we explored the behavior of the system when *Rainfall* is altered. Using Omnibrushing (Fig. 2(a)) we could easily see that lower values of *Rainfall* are associated with higher *Temperature* values and are also associated with lower values of *Percentage Full*. However, the highest values of *Rainfall* are not particularly correlated with higher values of *Percentage Full*.

After we explored the distribution of the *Rainfall* variable in relation to the others, we used Single Node Brushing to explore the changes when we selected the lowest *Rainfall* value and moved through to the highest value (Fig. 3), attempting to identify correlations between the variables. There is a clear negative relation between *Rainfall* and *Temperature* and a clear positive relation with *Percentage Full*, *Water Level* and *Amount Transfer in*. However, the relationships with *Percentage Evaporation* are more ambiguous. When *Rainfall* values are higher, *Percentage Evaporation* tend to be more prevalent in the second bin.



(a) Rainfall (b) Temperature

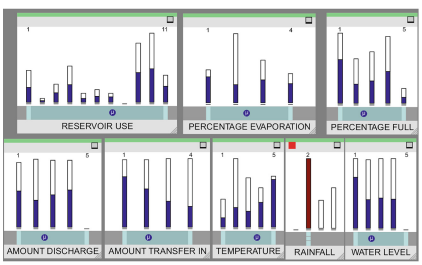
Fig. 2. Omnibrushing for *Rainfall* and *Temperature*

Another variable of prime interest is *Temperature*. As with *Rainfall*, we use Omnibrushing (Fig. 2(b)) and Single Node Brushing (Fig. 3) to explore its relation with the rest of the variables. First, we can see that medium values are more prevalent in the rest of the variables than both extremes (bins 1 and 5). When we focus on a subset, bins 1 and 2 (corresponding to temperatures lower than 15°), we find that samples are fairly flat except for lower *Percentage Evaporation* and slightly higher values of *Rainfall*. If we move now to the highest bin (temperatures above 25°), more changes are evident. The sample size is markedly smaller, so inferences must be less certain, but this smaller sample shows low rainfall and higher water discharge, presumably to combat drought conditions.

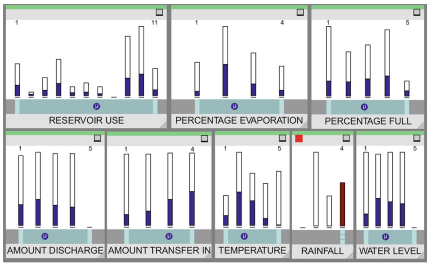
Next we followed the same procedure with *Percentage Full*. One thing we observed was that both *Amount Transfer in* and *Amount Discharge* behave in the same way with respect to *Percentage Full* (Fig. 3(e) and (f)) and that the relation between all three is positive. We computed the Pearson correlation between *Amount Transfer in* and *Amount Discharge* conditioned on *Water Reservoir*, which was a very high 0.95. This suggests some redundancy between the two variables *Amount Transfer in* and *Amount Discharge*; however, we have already observed that they behave in *opposite* ways in high temperature conditions (Fig. 3).

Finally, we took advantage of the Multi Node Brushing and explored the system when both *Rainfall* and *Temperature* nodes were in focus. One of the possible scenarios for the future of Andalusia combines an increase in the annual temperature with a decrease in average rainfall. Using OE, we checked the sensitivity of the system to this change. With the *Rainfall* node focused on the lowest bin, we ran the focus on *Temperature* from the medium values (bin 3) to the highest value (Fig. 4). The remaining variables showed changes as the focus moved, allowing us to do interactive sensitivity analysis.

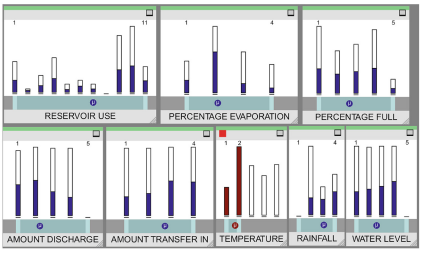
With this exploration process we gained some initial understanding of how the variables are related, but also some idea the system's causal structure. *Rainfall* and *Temperature* are clearly inversely related, whilst *Rainfall*, *Percentage Full* and *Water Level* are positively related. *Percentage Evaporation* is also related with both *Rainfall* and *Temperature*, but the relations seem to be more



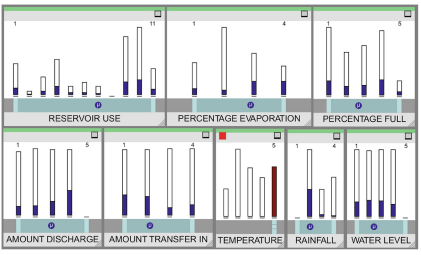
(a) *Rainfall*: Lowest values



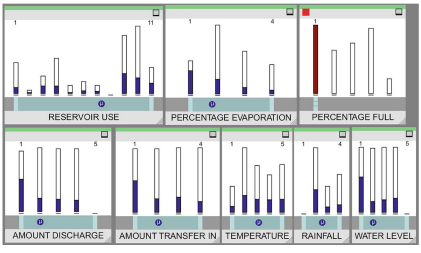
(b) *Rainfall*: Highest values



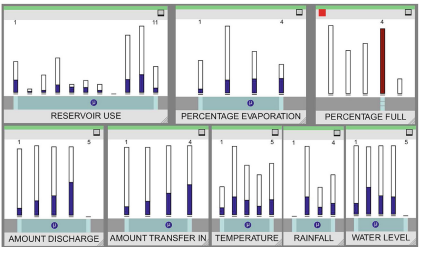
(c) *Temperature*: Lowest values



(d) *Temperature*: Highest values

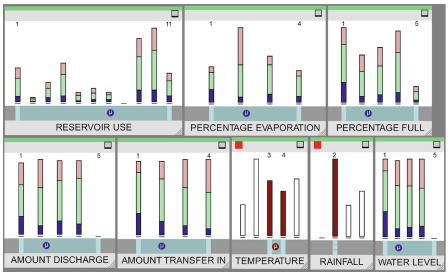


(e) *Percentage Full*: Lowest values

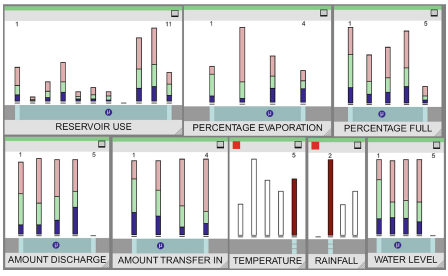


(f) *Percentage Full*: Highest values

Fig. 3. Single Node Brushing for *Rainfall*, *Temperature* and *Percentage Full* variables, focus on the lowest and highest values.



(a) *Temperature*: medium values



(b) *Temperature* :highest values

Fig. 4. Multi Node Brushing given the lowest value of *Rainfall*, for medium (a) to the highest (b) values of *Temperature*

complex. So, these relations should be included in the model. In both cases, *Rainfall* and *Temperature* seem to act as a posible cause of *Percentage Full*, *Percentage Evaporation* and *Water Level*, so they should appear in the network as parent of them. Also, given a fixed *Percentage Full*, *Amount Discharge* and *Amount Transfer in* provide similar information and should be considered closely related in the model.

3.4 BN Learning

We used the causal discovery program CaMML[2] (Causal discovery via Minimum Message Length) to learn causal structure (causal BNs) from the available data. CaMML uses a Bayesian metric (MML score) and stochastic search to find the model, or set of models, with the highest posterior probability given the data (see [4] Chap. 9, or [6]).

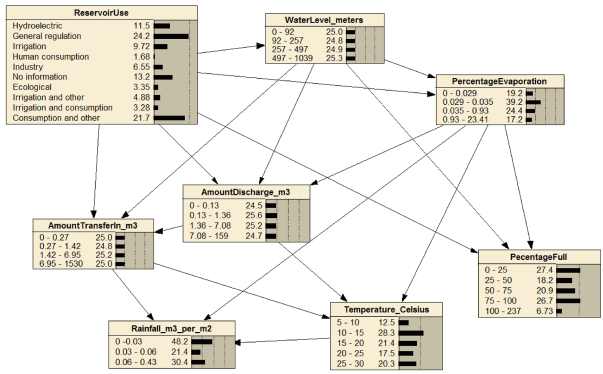
CaMML supports prior information about the structure of the model, such as what variables should be linked (Priors), or the partial (or total) order of variables (Tiers). The idea of using priors is to assist the discovery process with common sense background knowledge or genuine expert opinion, or, in this case, with what we think we have learned from data exploration. Inspired by OE, we tried the following Tiers and Priors (with varying degrees of confidence):

- *Priors:* There should be the following links: from *Rainfall* to *Percentage Full*, from *Percentage Evaporation* to *Percentage Full*, and from *Water Level* to *Percentage Full*.
- *Tiers:* Variables in the model should follow this structure: *Reservoir Use* in the first level, as a root node; in a second level *Rainfall* and *Temperature* as parent of *Percentage Evaporation*, *Amount Discharge* and *Amount Transfer ir* that are positioned in a third level; and, finally, *Percentage Full* and *Water Level*.

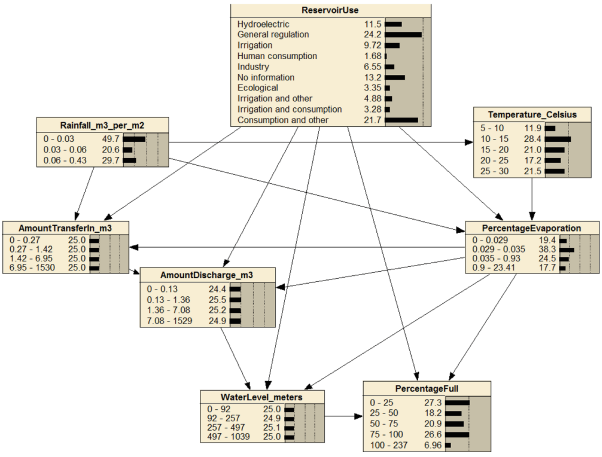
We ran CaMML on the data both with and without these priors and tiers. A *10-fold Cross Validation* was carried out to calculate the Logarithmic Loss using *Percentage Full* as target variable, to check what kind of predictive advantage the exploratory work might have for causal discovery. Figure 5 shows the models that were learned and their Logarithmic Loss values. This metric is exactly the same in both models (with and without priors and tiers), so we can consider them equally predictively adequate. However, including the information from OE yields a more useful structure from the environmental point of view.

Percentage Full and *Water Level* are our target variables, what we might like to influence with water management decisions (e.g., changing in *Reservoir use* or *Amount Transfer in*) or predict in response to climatic change scenarios (e.g., hypothetical changes to *Rainfall* and *Temperature*). For either use, the causal structure of the model without OE information does not allow us to model appropriately the effects on *Percentage Full* not *Water Level*, since, for

[2] https://github.com/rodneyodonnell/CaMML.



(a) Initial (Log.Loss: 1.1798+-0.022)



(b) Final (Log.Loss: 1.1798+-0.022)

Fig. 5. Netica displays and the Logarithmic Loss (Log.Loss) for the BNs learned both with (Final) and without (Initial) information from OE.

example, the climate change variables are child variables with respect to the rest of the model, while *Water Level* is spuriously shown as a causal factor for other variables. Reordering these variables in causally meaningful tiers was suggested by both common sense and the OE visualization process.

4 Conclusion

This paper describes the software Omnigram Explorer (OE) and its application to analysing and modeling water reservoirs in Andalusia. The initial exploration of the data with OE allowed the users to achieve a better understanding of the data, with the resultant causal structure better fitting the aims of modeling. The interactive graphical interface provides users with an easy and intuitive way to explore the data, as well as assisting the communication of results to non-specialists.

Acknowledgements. This work has been supported by ARC grant number DP110101758R. F. Ropero is supported by the FPU research grant, AP2012-2117, funded by the Spanish Ministry of Education, Culture and Sport.

References

1. Coreau, A., Treyer, S., Cheptou, P., Thompson, J.D., Mermet, L.: Exploring the difficulties of studying futures in ecology: what to do ecological scientists think? Oikos **119**, 1364–1376 (2010)
2. Heer, J., Bostock, M., Ogievetsky, V.: A tour through the visualization zoo. Commun. ACM **53**(6), 59–67 (2010)
3. Kelly, R., Jakeman, A.J., Barreteau, O., Borsuk, M., ElSawah, S., Hamilton, S., Henriksen, H.J., Kuikka, S., Maier, H., Rizzoli, E., Delden, H., Voinov, A.: Selecting among five common approaches for integrated environmental assessment and management. Environ. Model. Softw. **47**, 159–181 (2013)
4. Korb, K.B., Nicholson, A.E.: Bayesian Artificial Intelligence. CRC Press, Boca Raton (2011)
5. Méndez-Jiménez, M.: Estudio Básico de Adaptación al Cambio Climático (2012)
6. O' Donnell, R.: Flexible Causal Discovery with MML. Ph.D. thesis, Faculty of Information Technology (Clayton). Monash University, Australia, 3800, September 2000
7. Parrott, L.: Hybrid modelling of complex ecological systems for decision support: recent successes and future perspectives. Ecol. Inf. **6**, 44–49 (2011)
8. Ricci, P.F., Rice, D., Ziagos, J., LA Jr., C.: Precaution, uncertainty and causation in environmental decisions. Environ. Int. **29**, 1–19 (2003)
9. Spence, R., Tweedie, L.: The attribute explorer: information synthesis via exploration. Interact. Comput. **11**, 137–146 (1998)
10. Taylor, T., Dorin, A., Korb, K.: Omnigram explorer: a simple interactive tool for the initial exploration of complex systems. Proceedings of the European Conference on Artificial Life 2015, MIT Press, 381–388 (2015)
11. Teegavarapu, R.S.V.: Modeling climate change uncertainties in water resources management models. Environ. Model. Softw. **25**, 1261–1265 (2010)