UNIVERSIDAD DE ALMERÍA

DEPARTAMENTO DE BIOLOGÍA Y GEOLOGÍA

# Redes Bayesianas Híbridas: Una Herramienta Estadística en Ecología y Ciencias Ambientales


*Author:*
**Rosa M. Fernández Ropero**

*Advisors:*
**Dr. Pedro Aguilera Aguilera**
**Dr. Rafael Rumí Rodríguez**

*In which we try to interpret the present, understand the past, and perhaps predict the future, even when very little is crystal clear.* Russell and Norvig, 2002

# Acknowledgements

People often say *time flies*, but I was not conscious of that until now. It seems like yesterday when I read the first paper about Bayesian networks and thought the most clever option was to run away as far as possible. However, I kept on keeping on, and through out these years, an incredibly large number of people support me both academically and personally.

First and foremost, I would like to thank my advisors, Rafael Rumí and Pedro Aguilera for believing in me. Their encouragement has helped me to overcome the difficulties I found during this pre-doctoral period. In particular, I wish to thank Pedro for his persistence and determination in conducting this research. Also, I thank Rafa for his patience and friendly attitude which makes it easier for me to deal with this mathematical context.

I would particularly like to highlight Antonio Fernández for his valuable help and really good moments shared over more than 5 years working together. I will remember all the times he went downstairs to answer my S.O.S. messages.

I shall cherish great memories of the members of the Department of Mathematics with whom I share unforgettable moments: Antonio, María, Pepe, Darío and Rafa. Special thanks to Ana for being a source of support and sharing common interests and goals along these years.

I was fortunate to have the opportunity of visiting the following research groups in the last two years. In this sense, I would like to thank the members of the Deparment of Computing Sciences, SIMD $i^3$A, University of Castilla - La Mancha, for their really warm welcome. In particular, to Julia Flores for her friendliness and later collaboration. Thanks to Ann Nicholson, Kevin Korb and the rest of members of the Information and Technology Faculty, Monash University. They made my stay in Melbourne really interesting and enriching. I also wish to thank Linda van der Gaag, Silja Renoij and the rest of the members in the Department of Information and Computing Science, Utrecht University, for their hospitality and their contribution during my short but fruitful stay.

A nivel personal son muchas las personas cuyo apoyo veo reflejado de forma directa o indirecta en estas páginas.

A mis amigas de la carrera: Pili, Esther, Veli y Maribel, por todas las risas, anécdotas, viajes, tardes de biblioteca y tetería que hemos pasado juntas. Por el esfuerzo de conservar nuestra amistad pese al tiempo y la distancia, y por todos los momentos que aún nos quedan por vivir.

A Joe, por ser un excelente profesor que con el tiempo se ha convertido en un gran amigo. We'll keep on keeping on.

A mis padres, por su incondicional apoyo, a pesar de no tener muy claro aún a qué me dedico. En especial, a mi padre, por ayudarme a crear mis propias alas, y a mi madre, por inspirarme a volar cada vez más alto. A mi hermana Eva, quien con su ejemplo

me hace ver a diario que las piedras en nuestro camino son solo una excusa para los débiles, y una oportunidad para los valientes.

A mi familia, tanto presente como ausente, por todo lo que he aprendido de ellos. A mi abuelo Paco, allí donde estés, por ser el precursor de mi vocación. A mi familia política, por dejarme formar parte de sus vidas, en especial a Pilar, por su paciencia y su ayuda.

Y por último, aunque quizás el más importante, a Sergio, a quien esta Tesis también ha supuesto ciertos sacrificios. La vida es demasiado corta para hacer planes que quizás nunca se cumplan, pero mientras vivamos cada segundo juntos con intensidad, las promesas llegarán solas.

Finally, thanks to the foreign reviewers who will allow this thesis to achieve International Mention.

This work has been supported by the Spanish Ministry of Education, Culture and Sport through an FPU research grant, AP2012-2117.

# Resumen

La sociedad y los ecosistemas han evolucionado de forma conjunta a lo largo de la historia estableciendo una estrecha relación. Mientras que los ecosistemas proveen de bienes y servicios a la sociedad, esta, mediante sus acciones y decisiones, afecta a la estructura y funcionamiento de los mismos a distintas escalas espacio-temporales. La gestión de este complejo sistema de interacciones supone un desafío desde el punto de vista científico - técnico, ético, político. Por todo ello, se requiere de un enfoque interdisciplinar en las políticas de gestión y en los procesos de toma de decisión. El concepto de Socioecosistema definido como *un sistema integrado del hombre en la naturaleza*, aporta un nuevo marco conceptual integrado y unitario.

Dentro de la modelización ecológica, son diversas las técnicas y herramientas estadísticas que tratan de representar y modelizar el Socioecosistema desde diversas perspectivas. Desde la estadística tradicional pasando por algunos métodos basados en técnicas de Inteligencia Artificial, ecólogos y expertos en Ciencias Ambientales han tratado de obtener modelos capaces de manejar la complejidad de estos sistemas, así como incluir en los mismos conceptos como la incertidumbre o la probabilidad.

En esta Tesis se propone la aplicación de las Redes Bayesianas a la modelización ambiental y ecológica, y en concreto, a la modelización del Socioecosistema. Definidas al comienzo de la década de los 90, se han aplicado con éxito en áreas como Medicina y Ciencias de la Vida. Sin embargo, su aplicación en Ecología y Ciencias Ambientales es escasa y centrada en determinados aspectos, dejando aún sin explorar gran parte de su potencial.

El principal objetivo de esta Tesis es, por tanto, *el estudio de la aplicación de Redes Bayesianas híbridas como una herramienta probabilística en la modelización ecológica*, desglosado en cuatro objetivos secundarios que se corresponden con los cuatro capítulos principales. A lo largo de esta memoria se describen con detalle los conceptos básicos sobre los que se apoya esta nueva herramienta, con la finalidad de aportar un marco metodológico entendible por expertos en ecología y medio ambiente que no estén familiarizados con este tipo de técnicas.

Si bien el objetivo no es el de realizar una exhaustiva comparación con otras técnicas estadísticas aplicadas en el mismo campo, las Redes Bayesianas aportan ciertas ventajas frente a metodologías más clásicas. En primer lugar, su estructura visual basada en la Teoría de Grafos, permite que los modelos aprendidos sean fácilmente interpretados por expertos y actores sociales, permitiendo su aplicación en los procesos de toma de decisión y gestión de recursos naturales. Además, su naturaleza probabilística permite obtener los resultados como distribuciones de probabilidad en lugar de un valor absoluto. A diferencia de otras técnicas, los resultados obtenidos mediante Redes Bayesianas pueden ser interpretados de una manera más variada y detallada. A partir de las distribuciones de probabilidad obtenidas, distintos estadísticos como la media o la varianza pueden ser calculados. Además, es posible calcular la probabilidad de un determinado valor, o de un rango de valores, lo cual permite, por ejemplo, obtener la

probabilidad de que un embalse alcance un valor por encima del umbral de seguridad. Esta ventaja supone un mejor manejo de la incertidumbre asociada al modelo. Una de las mayores diferencias con respecto a otras técnicas, es la posibilidad de incluir variables tanto discretas como continuas en el mismo modelo. Mediante el uso del modelo *Mixture of Truncated Exponential*, ambos tipos de variables son introducidos de forma simultánea en el mismo modelo sin necesidad de ningún tipo de modificación en la estructura del mismo.

Una vez que se estudian las relaciones entre un conjunto de variables de un socioecosistema, a menudo surge la necesidad de determinar su comportamiento ante un cambio. Mediante el proceso de inferencia probabilística, las Redes Bayesianas son capaces de analizar las repercusiones de dicho cambio. Si bien esta es una propiedad común a otras técnicas, las Redes Bayesianas de nuevo aportan una ventaja sobre las demás al permitir incluir esta nueva información tan solo en aquellas variables que tengamos evidencia de un cambio, dejando que el resto se actualicen durante el proceso de inferencia.

Por último señalar que problemas tanto de caracterización, clasificación y regresión pueden ser abordados por las Redes Bayesianas tanto para el caso de datos estáticos, como series de datos temporales.

Por todo ello, las Redes Bayesianas híbridas constituyen una herramienta novedosa y con un gran potencial para su aplicación en la modelización ecológica y ambiental, haciendo frente a los principales desafíos de la modelización del Socioecosistema.

# Abstract

Society and ecosystems have co-evolved together along history. From the beginning of human societies, ecosystem support us with natural resources needed for our own development, and absorb the produced waste. At the same time, societies impact on ecosystem on loca, regional and global spatial and temporal scales. Managing cause and effect relationships among nature and society supposes a challenge from technical, ethical and political point of view. For that reason, an interdisciplinar framework is needed. The term Socioecosystem is defined as *an intregated system of human in the nature*.

Modeling these complexity is still a challenge, but several statistical tools try to represent and model Socioecosystems from different point of view. From traditional statistical techniques to some Artificial Intelligence based methods, ecologist and environmental sciences experts have tried to obtain methods able to manage the complexity involved in that social-natural systems, also, probability and uncertainty concepts.

In this Thesis, we propose the use of a new statistical model, Bayesian network, for Socioecological modeling. Defined at the beginning of the nineties, they have been successfully applied in Health and Life sciences. However, their application in Ecology and Environmental Sciences is still scarce and focused just on specific aspect, without taking advantage of their potentials.

The main objective of this Thesis is *the study of the applicability and contribution of hybrid Bayesian networks (hBNs) as a statistical tool for ecological and environmental modeling*, divided into four secondary goals developed in detail in the four main chapters. Throughout this manuscript, a deep explanation about the basic concept of this methodology is shown, with the aim of providing a theoretical framework for experts in ecology and environmental sciences.

Even when the idea is not to compare with other methodologies, hBNs present a set of advantages over others traditional tools. Firstly, the qualitative part of hBNs based on the Graph Theory, allow the models to be easily interpreted by experts and stakeholders, and make them appropriate for decision making processes and natural resource management. Besides, the quantitative part based on probability theory provide results in terms of probability distributions rather than a value. From these probability distributions, a set of measurements such as mean or standard deviation can be calculated. Also, the probability of a specific value or a range can be computed, which allow, for example, to obtain the probability of a dam reaching a value over the security threshold. This advantage involves a better uncertainty management. Secondly, the ability to deal simultaneously with discrete and continuous variables is one of the most important advantages. Thanks to the use of *the Mixture of Truncated Exponential models* both types of variables are included in the same model without any data preprocessing or changes in the model structure.

Once the relationship between variables have been modeled, their behaviour when a change happens need to be studied. Throughout the *inference* process, hBNs are able to analyzed that changes. Even when this property is common to other techniques, hBNs overcome them since information about hte change is included just in a set of variables, whilst the rest are updated during the *inference* process.

Finally, hBNs are able to deal with characterization, regression and classification problems both with static and temporal data.

For that reasons, Bayesian networks are a powerful and novelty tool to be applied in ecological and environmental modeling, able to overcome the main challenges of Socioecosystem modeling.

# Thesis Details

The main body of this Thesis is based on the following papers:

1. Ropero, R. F.; Flores, M.J.; Rumí, R. and Aguilera, P. A. (2016). Applications of hybrid Dynamic Bayesian networks to water reservoir management. Submitted to Environmetrics. pp. 1 - 19.
2. Ropero, R. F.; Rumí, R. and Aguilera, P. A. (2016). Modelling uncertainty in social-natural interactions. Environmental Modelling & Software 75, 362 - 372.
3. Ropero, R. F., Aguilera, P.A. and Rumí, R. (2015). Analysis of the socioecological structure and dynamics of the territory using a hybrid Bayesian network classifier. Ecological Modelling 311, 73 - 87.
4. Ropero, R. F., Nicholson, A. and Korb, K. (2015). Using a new tool to visualize environmental data for Bayesian network modelling. CAEPIA'15. Albacete, Spain. 9-12 November.
5. Ropero, R. F., Aguilera, P. A., Fernández, A. and Rumí, R. (2014). Regression using hybrid Bayesian networks: Modelling landscape - socioeconomy relationships. Environmental Modelling & Software 57, 127 - 137.

In addition to the main papers, the following publications have also been published:

1. Flores, M. J.; Nicholson, A. E.; Ropero, R.F. (2016). Dynamic OOBNs applied to Water Management in Dams. IEEE International Conference on Knowledge Engineering and Applications, ICKEA. Singapore, 28 - 30 September.
2. Ropero, R. F., Flores, M.J., Rumí, R., Aguilera, P.A. (2016). Modelling time in species distribution models: a dynamic Bayesian network approach. 5th International ECOSUMMIT, Le Corum, Montpellier, France. 29 August - 1 September.
3. Ropero, R. F., Aguilera, P. A., Fernández, A. and Rumí, R. (2014). Redes bayesianas: una herramienta probabilística en los modelos de distribución de especies. Ecosistemas 23, 1 - 6.
4. Ropero, R. F., Aguilera, P.A. and Rumí, R. (2014). Soft Clustering based on Hybrid Bayesian networks in Socioeclogical Cartography. Hybrid Artificial Intelligence Systems, 9th International Conference, HAIS, Salamanca, Spain, 11 - 13 June. pp. 607 - 617.
5. Aguilera, P. A. , Fernández, A., Ropero, R. F. and Molina, L. (2013). Groundwater quality assessment using data clustering based on hybrid Bayesian networks. Stochastic Environmental Research & Risk Assessment 27, 435 - 447.
6. Ropero, R. F., Maldonado, A., Aguilera, P. A., Fernández, A., Rumí, R. and Salmerón, A. (2013). Discrete vs. Hybrid Bayesian network in ecological modelling. INTECOL Congress, Excel, London, UK, 18 - 23 August.

7. Fernández, R., Willaarts, B. A., Fernández, A., Rumí, R. and Aguilera, P. A. (2012). Social structure- land use- water flows: modelling the relationships by discrete Bayesian networks. Proceedings of the 6th International Congress on EMSs, Leipzig, Germany, 1 - 5 July. pp. 1992-1999.

8. Aguilera, P. A., Fernández, A., Fernández, R., Rumí, R. and Salmerón, A. (2011). Bayesian networks in environmental modelling. Environmental Modelling & Software 26, 1376 - 1388.

# Contents

# List of Abbreviations

| | |
|---|---|
| **ABM** | Agent Based Models |
| **AI** | Artificial Intelligence |
| **AML** | Agricultural Mediterranean Landscape |
| **BNs** | Bayesian networks |
| **CaMML** | Causal Discovery via Minimum Length |
| **CBW** | Consumptive Blue Water |
| **CG** | Conditional Gaussian |
| **CPT** | Conditional Probability Table |
| **GEC** | Global Environmental Change |
| **GIS** | Geographical Information System |
| **DAG** | Directed Acyclic Graph |
| **DBNs** | Dynamic Bayesian networks |
| **hBNs** | Hybrid Bayesian networks |
| **MLR** | Multiple Linear Regression |
| **MTE** | *Mixture of Truncated Exponential* |
| **NB** | Naïve Bayes |
| **NPGW** | Non-Productive Green Water |
| **OE** | Omnigram Explorer |
| **PGW** | Productive Green Water |
| **RBW** | Runoff Blue Water |
| **rmse** | Root Mean Squared Error |
| **SC** | Systemic Change |
| **SES** | Socio-Ecological System |
| **TAN** | Tree Augmented Naive Bayes |

# Motivation, Objectives and Structure

## Motivation

Nowadays, the impact societies have on the biosphere, and the inter-relationships between different natural and social subsystems is clear. National and International Governments are aware of about the necessity of including both social and natural aspects into management plans and politics, which makes interdisciplinary studies indispensable. Besides, the spread of information technologies, and Geographical Information Systems, takes us to a new era of information in which large amounts of data are available.

Bayesian networks have been developed in the last 30 years and successfully applied in such important areas as Health or Life Sciences. From the theoretical and real life applications point of view, Bayesian networks have demonstrated their powerfulness when dealing with different types of data (discrete, continuous and hybrid), sources of information (empirical data, expert opinion, literature information), problems to face (classification, regression, future scenario study) and, also, time series data.

In ecology and environmental modeling they started to be applied at the beginning of the nineties with a few papers per year, and the number of papers increased in the following years. However, their application in that field is still scarce and partial. Currently, just around 50 papers per year can be found in literature and most of them are applied to specific areas, like water research and ecology. Despite the advantages of using original continuous data, the majority of papers discretized them and used a small set of all the available algorithms and softwares. Also, both classification and regression problems are scarcely solved by Bayesian networks, which are mainly used for studying scenarios of change or the relationships between the variables involved in the model.

In this sense, further efforts are needed to encourage experts in ecology and environmental sciences to use this statistical tool, expand its application to other areas (*i.e.* biological conservation, land use management or fisheries) and deal with original continuous or hybrid data. Several reviews and papers about how to use Bayesian networks in environmental modeling are available, but their framework are often focused on discrete data. This Thesis aims to provide experts in ecology and environmental science with a set of theoretical framework, applications, recommendations and tools for a proper application of hybrid Bayesian networks (including discrete and continuous variables simultaneously).

This Thesis is quite ambitious since it tries to merge two different areas; on the one hand, Ecology and Environmental Sciences through the concept of socio-ecological Systems; on the other hand, Mathematics and Computer Science through the use of

hybrid Bayesian networks. The expected main audience of this Thesis is experts in ecology and environmental modeling interested in applied BNs, so basic concepts and theoretical framework of Bayesian networks needs to be explained in detail. This can make readers from mathematics and computer science areas feel some parts are redundant or unnecessary to be described.

## Objectives and Structure

The main objective of this Thesis is *the study of the applicability and contribution of hybrid Bayesian networks (hBNs) as a statistical tool for ecological and environmental modeling*. To achieve this goal, four different secondary goals were proposed:

1. Study of characterization problem through hBNs.
2. Study of regression techniques through hBNs.
3. Study of non-supervised classification techniques based on hBNs.
4. Study of dynamic models based on hBNs.

For each secondary goal, a set of applications has been proposed: the study of Systemic Change in a socio-ecological System, Landscape-Society interaction in a socio-ecological System, socio-ecological cartography and water reservoir management.

This manuscript is divided into three parts. **Part I** contains the *Introduction* divided into two chapters: Chapter 1 describes the ecological framework on which this Thesis is based, whilst Chapter 2 deals with the theory behind Bayesian networks: definition, their adaptation for hybrid domains and an overview about their development and applications in several areas, mainly in environmental modeling.

**Part II** corresponds to the main part of this Thesis, where Chapters 3, 4, 5 and 6 each deals with one of the four secondary goals. They correspond to the main problem in which hBNs can be applied following the flowchart shown in Figure 1. The first step in BNs modeling is to decide the objective of the model among three main options: Characterization, Regression and Classification (Figure 1 a)). For all these options, structure and parameter have to be learnt. It can be carried out following the flowchart shown in Figure 1 b). A step beyond these static models, is the extension of BNs to deal with time series according to Figure 1 c) and presented in Chapter 6.

Chapter 3 deals with *Characterization* problem with hBNs. In this chapter, the theory behind structural learning through automatic algorithm and expert knowledge is presented and how the concept of *d-separation* is appropriate to manage Systemic Change study. The environmental application is focused on the discovery of relationships between subsystems (both natural and social) in a socio-ecological System and how a disturbance can be propagated through the system and have its impact checked in the rest of the variables. Also, uncertainty in environmental modeling is analyzed.

Chapter 4 presents an application based on a *Regression* problem. The relation between society and landscape is modeled through a hBNs regression model. In this case, fixed structure is applied and their potential is demonstrated against other techniques (traditional regression methodologies, and discrete BN model). Besides, the relationships between socioeconomic variables and landscapes are deeply studied and whether they

are directly proportional or not is established. Finally, two scenarios of social evolution are included and their impact on the landscape is observed.

Chapter 5 deals with the *Classification* purpose. In this case, hBNs are applied to identify the socio-ecological structure of a territory and obtain a map of socio-ecological sectors in an extended region. The proposed approach divides the problem in such a way that this complex task can be easily tackled allowing it to be clearly understood by experts. Also, a scenario of Global Environmental Change is included with the aim of studying the impact of this global phenomenon in a Mediterranean region.

Chapter 6 is focused on *Dynamic* or temporal models through hBNs. In this chapter, the water reservoir system of Andalusia is modeled under a dynamic framework with the aim of predicting the temporal behavior of the reservoir capacity. Firstly, a comparison between static and dynamic models is carried out, and later on the chapter deals with the main approaches for learning and performing inference in Dynamic BNs (DBNs).

Finally, **Part III** presents Chapter 7 in which the conclusion of this Thesis are drawn.

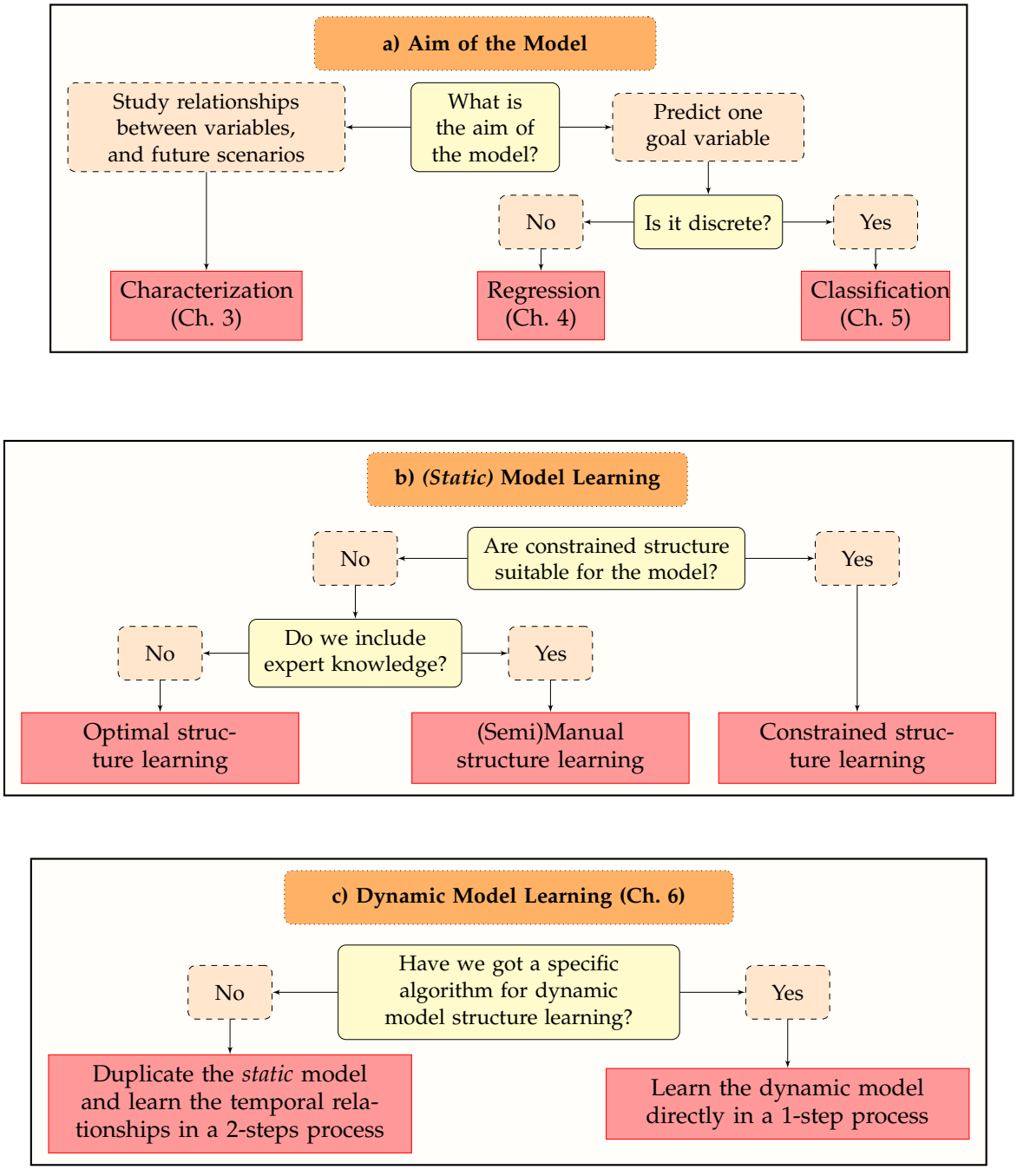



FIGURE 1: Flowchart for BNs modeling process in which each Thesis goal is pointed out. Ch., Chapter.

# Part I

# Introduction

# Chapter 1

# Ecological Framework: Socio-Ecological Systems

In this Chapter the ecological framework based on the concept of Socio-ecological Systems is explained. From their initial motivation and development through classical statistical tools until today, this concept has hardly been studied using models that allow complexity to be taken into account.

Human society and natural systems have co-evolved along history (Norgaard, 1988). From the beginning of human societies, ecosystems support us with natural resources needed for our own development, and absorb the produced waste. Before the Industrial Revolution, societies impacted on ecosystem in a slight way, but the development of new technologies and industrial processes, together with the increase in the human population, have led to an important degrade of impact on the ecosystems on local, regional and global scales (Liu *et al.,* 2007). Nowadays, effects of society in ecosystems are clear and visible, and managing cause and effect relationships among nature and society supposes a challenge from several points of view: technical, ethical, political and ecological.

In this sense, several international projects (like the Intergovernmental Panel on Climate Change (IPCC, 2014), or the Millennium Ecosystem Assessment (M.E.A., 2003)) demand an integrated management of natural systems, and Ecology is applied with the aim of bringing a scientific background into political discussion and decision-making (Carpenter & Folke, 2006). Relationships between natural and social systems have been deeply studied from the ecological perspective, *i.e.*, how human activities affect the natural processes and the structure of ecosystems. Recently, researchers are commonly also focused on how changes in ecosystems affect human wellbeing and the generation of ecosystem services, and try to relate both natural and social systems (Martín-Lopez *et al.,* 2009). This integration stage a conflict between ecological sustainability *versus* economic development and growth (Eliott, 2013) since ecologists have focused on the ecosystems considering society as external agents; whilst social sciences have treated ecosystems as an external influence (Barnard & Elliott, 2015). For a successful sustainable development, three dimensions should be considered: economy, society and the environment (Wang *et al.,* 2011; Dawe & Ryan, 2003; Young, 1997).

From this new integrated point of view, the concept of Socio-Ecological System (Figure 1.1) can be defined in different ways (Martín-Lopez *et al.,* 2009):

**SES definition 1:** *An integrated system of human in the nature* (Anderies *et al.,* 2004)

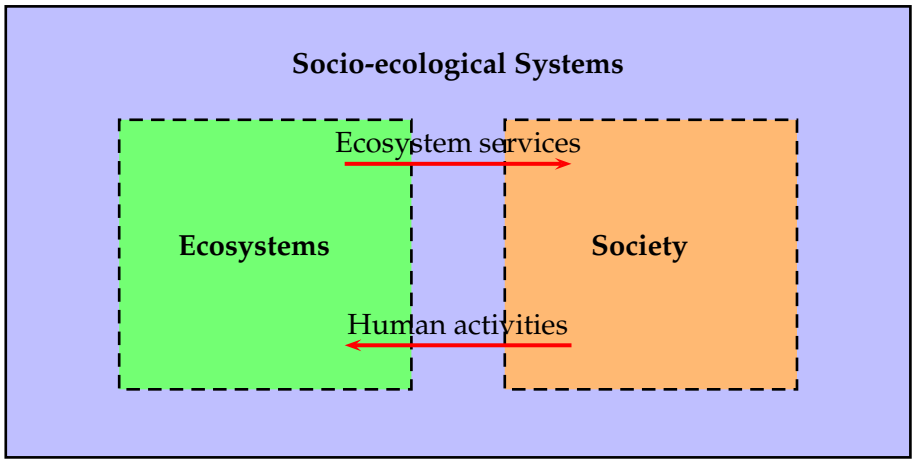



FIGURE 1.1: Flowchart of the components of Socio-ecological Systems and their relations.

**SES definition 2:** *An ecological system linked to and affected by the social system* (Folke, 2006)

Under this framework, nature and society are considered as a clearly related system and so any delimitation between them is artificial and arbitrary (Berkes & Folke, 1998). When all these links between social and ecological systems are identified, the overall system is really complex. Instead, relationships between them operate at different spatial and temporal scales (Liu *et al.*, 2007; Anderies *et al.*, 2004). This complexity has been widely studied from the ecological perspective (Cadenasso *et al.*, 2006; Holling, 2001).

SES exhibits a set of properties that do not come from human or natural systems separately, but stemming from the interactions between them (Liu *et al.*, 2007). Some ecological properties that need to be defined for the SES modeling are vulnerability and resilience (Young *et al.*, 2006).

**Vulnerability:** *the degree in which an ecosystem changes due to both internal or external changes.*

**Resilience:** *the ability of an ecosystem to maintain similar structure and functions after a change.*

In the case of a SES the term vulnerability needs to be extended to also include disturbances in the overall system due to an alteration in relationships, rather than a specific change in a set of variables (Liu *et al.*, 2007; Walker *et al.*, 2002). When this kind of disturbance affects an a SES, it may be altered or not depending on its own resilience.

Modeling these complex and heterogeneous SES is still a challenge (Filatova *et al.,* 2013; Filatova & Polhill, 2012; Jorgensen, 1999). Even when the term SES is novel, attempts to model both social and natural systems are not new.

Traditional statistical techniques have extensively been applied (Hong *et al.,* 2016; Atuo *et al.,* 2015; Van Holle *et al.,* 2014; You *et al.,* 2014). The most extensively used are regression models, including logistic regression. Some examples can be found in literature. In this sense, Salvati & Carlucci, 2015 study the effect of grazing over agro-forestry systems including socioeconomic variables as important factors. In the paper of Sterzel *et al.,* 2014 how socio-ecological vulnerability is related to armed conflict in global drylands on a subnational level is studied. Vu *et al.,* 2014 explores the socioeconomic factors that determine land degradation in Vietnam. Schmitz *et al.,* 2012 determine the effect of protected areas over landscapes dynamics and socioeconomic development. Beverly *et al.,* 2011 include social components in modeling changes in the annual wildfire activity in Canada in relation to Northern Hemisphere climate variability. Bellassen *et al.,* 2010 model the organic carbon sequestration potential from different agricultural intensification process in Senegal.

Other techniques applied are classification trees and cluster analysis. Staudhammer *et al.,* 2015 applied classification trees to examine the spatial distribution and Socio-ecological predictors of invasive plants in different ecosystems. Nair *et al.,* 2016 used cluster analysis to characterize regional landscapes typologies, whilst Harlan *et al.,* 2013 worked on the prediction of social and environmental factors for heat deaths under a climate change framework.

As an extension of these classic methodologies, Agent Based Models (ABM) have evolved and been applied into SES modeling. Initially developed as a computer science paradigm called object-oriented programming, their evolution was feed by several areas (Bousquet & Le Page, 2004; Langton, 1988). Nowadays, an ABM involves the creation of several virtual objects with autonomous behavior (agents) to represent real actors and their interactions between one and other. In SES modeling, they are mostly applied to represent the institutional and governance structures, which is crucial to understand how policy and organizations provide feedback to agent behavior (Rounsevell *et al.,* 2012). The main advantage of these kinds of models is their ability to include the behavior of social actors in a more realistic way, also, combining with a dynamic heterogeneous representation of the spatial environment (Filatova *et al.,* 2013).

Some examples of its application are found in Verhoog *et al.,* 2016 in which a biogas infrastructure effect is studied in The Netherlands; Mena *et al.,* 2011 simulates how the dynamic of land use change impacts household farms in the Amazons; Rebaudo *et al.,* 2011 use ABM as a tool for modeling the interaction between pest invasion and farmers in an agricultural landscape in the tropical Andes; or Bousquet & Le Page, 2004 that explores the necessity of a new shift in the paradigm toward a clear and explicit integration of society into these models.

An extended revision of ABM was done by Filatova *et al.,* 2013 pointing out its main challenges and An, 2012 in which its advantages and disadvantages are highlighted. One of the most important disadvantages is to compare different agent-based models due to the high variability of methodologies and frameworks applied. Besides, modeling human behavior and decisions making processes are still considered as a challenge despite the advances in that field (Smajgl *et al.,* 2011), as well as how to include biophysical processes (Matthews *et al.,* 2005).

Models described above were mainly developed and applied from the ecological and environmental sciences. In contrast, social sciences deal with this coupled human-nature systems through the use of some Social Theories (Borges *et al.,* 2014; Poppenborg & Koellner, 2013; Wauters & Mathijs, 2012). The most commonly used is the Theory of Planned Behavior (Ajzen, 1991) which describes Human Behavior as a result of behavioral intention, subjective norms and perceived behavioral controlled attitudes. For example, in Deng *et al.,* 2016 this theory is used to analyze the factors affecting the intention of farmers for ecological conservation via payment of ecosystem services.

Under this social perspective Solstrand, 2015 studies several conservation centers and theoretical framework in Iceland and Norway to check the challenges of current policy and management actions. Besides, some models are based on purely physical modeling approaches that try to describe the biophysical basis of human societies (Fischer-Kowalski, 2011; Suh *et al.,* 2010; Miller & Blair, 2009).

In this Thesis, we propose the use of a new statistical model, Bayesian networks, for SES modeling. There are some attempts to model SES using BNs (Drees & Liehr, 2015; Naranjo-Madrigal *et al.,* 2015). The next chapter includes an extended overview of literature.

# Chapter 2

# Bayesian networks in Environmental Modeling

In this Chapter, an introduction to Bayesian networks (BNs) is shown which includes their definitions, types of problems they can deal with, their adaptation to hybrid domains and their extension to dynamic models. Their main concepts are explained in order to provide a general knowledge about this new tool necessary to understand the rest of the Thesis. Finally, an overview about their application in literature is included.

## 2.1 Bayesian networks definition

Bayesian networks were firstly defined by Judea Pearl who designed an algorithm for efficiently computing probabilities in the mid-1980s (Pearl, 1986) and proposed it as a novel approach to apply probability Theory for reasoning with uncertainty in knowledge - based systems (Pearl, 1988b). Also called belief networks or Bayesian belief networks, their potential usefulness and applications were immediately identified by the Artificial Intelligent (AI) community but the strong mathematical concepts and structural limitations in Pearl's algorithms limited their spread to other areas (Charniak, 1991). When these structural and conceptual limitations were overcome, BNs became really popular for dealing with uncertainty domains (Andersen *et al.,* 1990; Jensen *et al.,* 1990b; Shenoy & Shafer, 1990). During the nineties, their potential applications were dramatically expanded for several reasons; *i)* machine learning techniques were developed and allowed BNs to be directly learnt from datasets (Cooper & Herskovits, 1992; Spirtes *et al.,* 1993), *ii)* they were proposed for pattern recognition or classification tasks giving robust and accurate results in comparison to others well known classifiers (Friedman *et al.,* 1997), and *iii)* the introduction of hybrid domains in which both discrete and continuous variables can be included in the same BNs models (Lauritzen, 1992).

They were defined by Jensen & Nielsen, 2007 as:

**Bayesian network:** *A Bayesian network consists of the following:*

- *A set of variables and a set of directed edges between variables.*
- *Each variable has a finite set of mutually exclusive states.*

- *The variables together with the directed edges form an acyclic directed graph (DAG); a directed graph is acyclic if there is no directed path $A_1 \rightarrow \ldots \rightarrow A_n$ so that $A_1 = A_n$.*
- *To each variable $A$ with parents $B_1, \ldots, B_n$, a conditional probability table P(A | $B_1$, $\ldots$, $B_n$) is attached*

The DAG configures the qualitative part, whilst the probabilities assignments the quantitative part of a BNs. Both qualitative and quantitative components are explained in detail below.

### 2.1.1 Qualitative component

The qualitative component represents the structure of the model, which is based on the well known Graph Theory. A BN structure is defined as a direct acyclic graph (DAG) in which variables are represented as nodes, and the presence of an edge linking two variables indicates the existence of a statistical dependence between them.

For a clear understanding about the mathematical concepts involved in BNs modeling, a toy and simple example about an illegal dumping in a pool is presented. This example does not focus on the ecological process modeling from an ecological point of view, since more variables and processes should be included to be modeled appropriately.

*Suppose we are hiking in the countryside and plan to picnic close to a pool, but the water has a bad smell and green color. We want to know if this situation is due to an illegal dumping, or to natural reasons. We know that an illegal dumping provokes an increase of nutrients in the water, with a quick development of weed as a consequence, which makes the water turn green. This implies a change in the ecological conditions of the pool and the water starts to rot. Also, the dumping could provoke some lather in the water surface. However, a decrease in the river flow that feeds the pool can also involve the accumulation of nutrients, so the process is natural.*

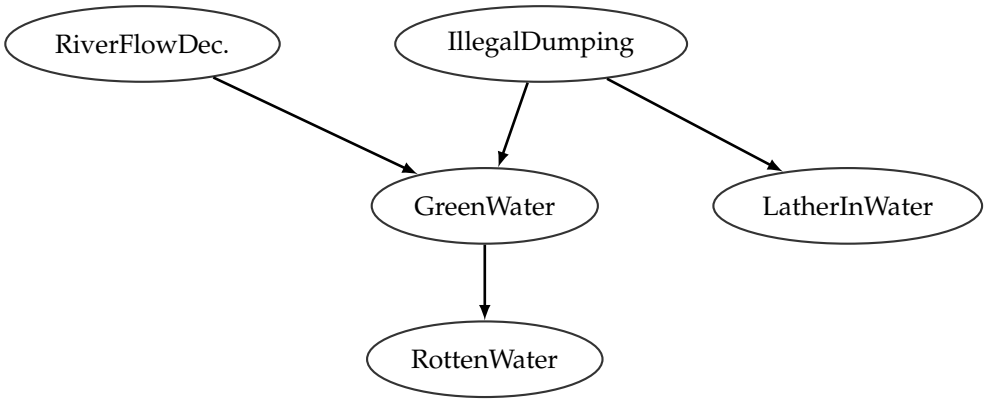



FIGURE 2.1: The qualitative part for the *Illegal dumping* example.

Figure 2.1 shows the DAG for this example in which each node represents one of the random variables with two possible states: True or False. This qualitative structure allows us to see the variables included in the model and their causal relationships and infer what is going to happen (if there is an illegal dumping, water will be green), or discover causes from observed effects (if there is lather and a green color, probably the cause is an illegal dumping). Besides, from this information we can determine what variables are important for a certain one with no mathematical calculation involved

(for the variable *RottenWater*, the variable *LatherInWater* is not directly related and has no direct influence over it). This is related to the concepts of *d-separation* (explained in detail in Chapter 3) and *Markov blanket* (Friedman *et al.*, 1997). For a node $X_i$ its *Markov blanket* is a set of nodes composed of $X_i$'s parents, its children and the parents of its children apart from $X_i$. For example, for variable *GreenWater*, the *Markov blanket* is the set of variables composed of: *RiverFlowDec.*, *IllegalDumping* and *RottenWater*.

In general, three types of relations can be identified in a DAG, enough to explain how information flows through the network (Figure 2.2):

1. Serial connections (Figure 2.2 a)). In the example, the variable *RiverFlowDec.* has a direct influence on *GreenWater*, which in turn affects *RottenWater*. So, information flows from *RiverFlowDec.* to *RottenWater* and viceversa. However, if we see the water is green (variable *GreenWater* is true), information about the river flow (if it has been reduced or not) is not important for our belief about *RottenWater*.

2. Converging connections (Figure 2.2 b)). Variable *GreenWater* is directly influenced by both *IllegalDumping* and *RiverFlowDec.* In this case, these last two variables are irrelevant to each other. If we know nothing about the color of the pool (we haven't arrive to check it yet), neither variables have relationships between them. But, if we have some knowledge about the color, and there is probably no river flow decrease, then this will affect my belief about *IllegalDumping*: some village up in the river is illegally dumping in the river which provokes the change in the water color, which also reduces the probability of a decrease in the river flow.

3. Diverging connections (Figure 2.2 c)). *IllegalDumping* directly influences both *LatherInWater* and *GreenWater*. In this case, information flows from *LatherInWater* to *GreenWater* and viceversa. If we see lather in the water surface, then the probability of variable *GreenWater* would rise. But, if we know that there is an illegal dumping (we check the information on the news), any information about the lather is irrelevant to our belief of the water color and viceversa (maybe the lather has not appeared yet).

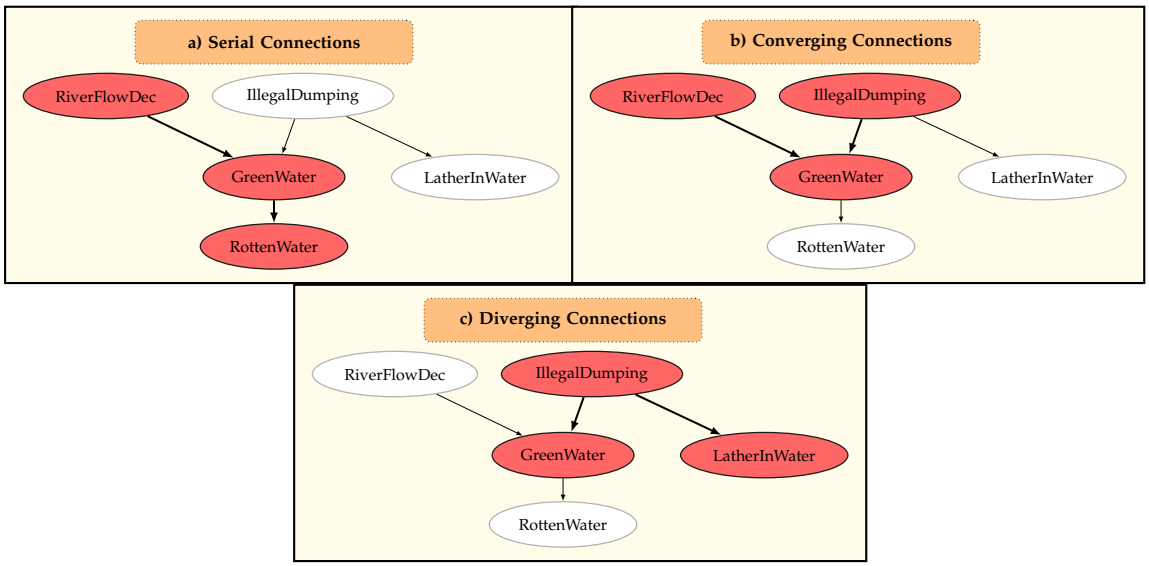



FIGURE 2.2: Types of connections in a DAG structure: serial (a), converging (b) and diverging (c).

One of the main advantages of BNs is their associated DAG structure, which determines the (in)dependence relationships between the variables. Also, it makes this model easy to interpret and understand by experts and stakeholders who play an importan role in real life problems modeling (Voinov & Bousquet, 2010).

### 2.1.2 Quantitative component

However, the causal relationships shown in the DAG structure are not absolute, sometimes even when there has been an illegal dumping, the lather does not immediately appear, or the water needs some days to suffer a color change in a significant way. So, for example, what is the probability of finding lather in water after an illegal dumping? To solve this question, the Probability Theory is included. Associated with the qualitative structure, there are a set of numerical functions representing the strength of these relationships between the variables.

In a network we can differentiate between two types of nodes: a root nodes are those that have no predecessors (*RiverFlowDec.* and *illegalDumping*), and child nodes are those which have predecessors or parent nodes (*RiverFlowDec.* is the parent of the child node *GreenWater*). Firstly, the probability distributions of all the root nodes is included, and for the rest of the nodes, the probability is expressed as a conditional probability giving all possible combinations of their parents. Taking into account the structure given by the DAG structure and using the rule of the probability, we can express the probability of a variable as

$$p(x_1, \ldots, x_n) = \prod_{i=1}^{n} p(x_i \mid pa(x_i)) \tag{2.1}$$

This is what allows complex models to be expressed in an adequate way. In the *Illegal Dumping* example it means that the probability of any child can be easily calculated from the probability of its parents. In Figure 2.3 some examples are shown: the probability of the river flow decrease is 0.1 (10% that we previosly calculated from the data), and 0.2 for an illegal dumping. So, for the variable *GreenWater* the probability of being true when there is no decrease in the river flow (Rfd = F) but an illegal dumping (Id = T) is equal to 0.7 (the same explanation for the rest of the values). These probability values can be estimated from the data or even elicited by expert knowledge. This is explained later in this Thesis.

The definition provided by Jensen & Nielsen, 2007 focused on discrete BN in which the quantitative part is expressed as a conditional probability table since the variables have a finite number of states. In the case of continuous variables, relationships are quantified as density functions.

## 2.2 Bayesian networks learning

The definition of the aim of the model is a key point (Chen & Pollino, 2012). Figure 1 shows the decision making process for BNs modeling in which we can see BNs are a versatile tool which can be adapted to different kinds of problems: *Characterization*, *Regression* and *Classification*. These three types of models are extensively explained in

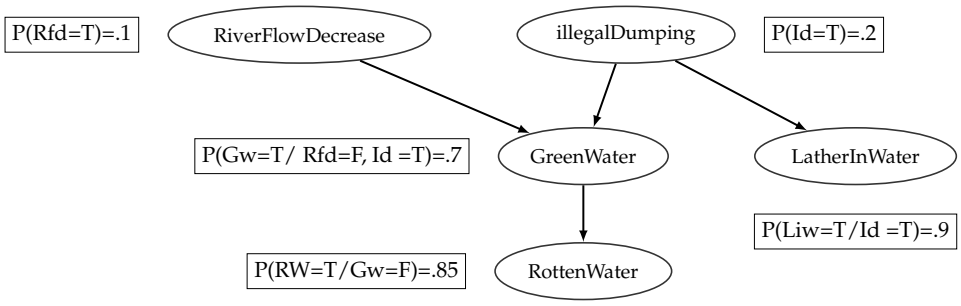



FIGURE 2.3: Some examples of the quantitaive part for the *Illegal Dumping?* example. Adapted from Charniak, 1991. T, True; F, False; Rfd, RiverFlowDec.; Id, illegal Dumping; Wi, GreenWater; Liw, LatherInWater; RW, RottenWater.

Chapters 3, 4 and 5, respectively. Also, they can be extended to deal with temporal dataset, configuring a Dynamic BNs (see Chapter 6).

Firstly, the aim of our model needs to be identified, and two options are available: *i)* predict one goal variable, which has to be estimated as accurately as possible, and *ii)* use the model for studying the relationships between variables and also predicting some future scenarios. In the first case, we need to distinguish between two possible cases relevant to the nature of the target variable. If it is discrete, the problem we are facing is *classification* and the goal variable is called the *class variable*, whilst it is a *regression* when this variable is continuous. In the second case, we are dealing with a model for *Characterization* purpose.

Once the aim of the model is clear, the next step is the decision about the model structure and learning process. The first question to answer is if the structure should be constrained or not. Constrained structures are mainly developed for classification and regression purposes and their main advantages are that they are defined for accurately predicting the goal variable values, but not the distribution of the rest of the variables, called features. They usually have fewer links than non constrained structures which implies a smaller number of parameters to be estimated but yielding appropriate results. Some examples of constrained structures are the naïve Bayes (NB) (Minsky, 1963), TAN (Friedman *et al.*, 1997), kDB (Sahami, 1996) and AODE (Webb *et al.*, 2005), in which the possible relationships between variables are restricted. These structures are automatically learnt from the data and no information from stakeholders or experts is needed for establishing the relations between variables. But they can be used for variable selection and other pre-processing steps.

In contrast, if we decide not to use a constrained structure, we need to learn the optimal structure using data or expert information, or both, by automatic or manual approaches, or a combination of them. This will be deeply explained in Chapter 3.

Static BNs have demonstrated their ability to provide robust and accurate results, but, if temporal series data are available and we want to take advantage of their temporal behavior, a step beyond is the use of dynamic BNs. In such case, we need to face the following question: Have we got a specific algorithm for dynamic model learning? In case we have it, we execute the model structure and parameters learning using that algorithm. However, they can still be difficult to apply for experts in other areas, so another option is to use static BNs algorithms for dynamic model learning. In this case, firstly a static model is learnt, and in a second step, it is repeated and linked through a

set of temporal relationships that represents the transition from one time to the next. In this way, a complex model is learnt with some sub-networks, each one for a different time. Both cases are deeply explained in Chapter 6.

## 2.3 Bayesian networks inference.

Once the model is learnt, BNs allow new information to be included in the variables as new values or *evidence*, and update the probability values of the rest of the variables by what is called *inference* or *probability propagation*. If we denote the set of *evidences* as **E**, and its values as **e**, then the inference process consists of calculating the posterior distribution, $p(x_i/\mathbf{e})$, for each variable of interest $X_i \notin \mathbf{E}$.

In this way, BNs can compute the effects given the causes (What is the probability of having rotten water given a high probability of a weed increase?), and even the causes given their effects (If we know there is a weed increase, what is the probability of being provoked by illegal dumping?) (Malekmohammadi *et al.,* 2009; Uusitalo, 2007; Getoor *et al.,* 2004).

Several algorithms have been defined and proposed for efficiently computing these pro-bability values both exact or approximately (Madsen & Jensen, 1999; Shenoy & Shafer, 1990). On the one hand, exact inference algorithms obtain the posterior probability distribution in an exact way, usually based on the idea of performing the computations locally. Some examples are the fusion algorithm (Pearl, 1988b), variable elimination method (Zhang & Poole, 1996) or the junction tree algorithm (Jensen *et al.,* 1990a).

However, obtaining exact values of probabilities is difficult and computationally costly, mainly when the network is so complex (Cooper, 1990). Thus, approximate algorithms are also proposed. They can be divided into two main groups:

- Methods based on simulation. These methods are based on the Monte Carlo methodology to simulate a sample of the variables in the network and estimate the probability distributions from it (Salmerón *et al.,* 2000).
- Deterministic methods. In this group several ideas have been proposed such as the Penniless algorithm based on the Kullback-Leibler cross entropy as a measure of the error of approximation (Cano *et al.,* 2002; Cano *et al.,* 2000), replace the lowest probability values with zeros to reduce the information complexity (Jensen & Andersen, 1990), simplify the network structure avoiding weak relationships (Kjærulff, 1994), or focus on the most probable configurations (Santos & Shimony, 1994).

## 2.4 Bayesian networks for hybrid domains.

BNs were initially developed to deal with discrete variables in which the results are expressed as Conditional Probability Table (CPT). Thus, a wide range of algorithms, software and applications are easily found in literature (Aguilera *et al.,* 2011; Marcot *et al.,* 2006). However, applications in ecology and environmental sciences involve the use of continuous variables, or even, a mix between discrete and continuous variables in the same dataset. In these cases, the available algorithms require conti-nuous data to

be transformed into categorical or discrete variables. This data discretization often implies a loss of information and accuracy (Uusitalo, 2007), and supposes a challenge in environmental modeling through BNs. Some attempts to solve this task using several methodologies can be found in literature - entropy minimization, equal width, equal frequency, deterministic equations, k-means, ChiMerge, among others- that try to obtain the most suitable thresholds for each variable (Christofides *et al.,* 1999; Kozlov & Koller, 1997; Dougherty *et al.,* 1995). Other techniques try to incorporate ecological knowledge to offer an objective approach for the discretization process (Lucena-Moya *et al.*, 2015).

In spite of these improvements in the classical discretization methods, dealing with continuous or hybrid data is usually the best solution. Estimation of the BN's parameters directly from the original data returns a model which can report more specific and accurate solutions for the proposed objectives (Ropero *et al.,* 2016). Following this idea, some models were proposed to represent probability distributions in hybrid BNs.

The first proposal was the *The Conditional Gaussian* (CG) model (Lauritzen, 1992) in which both continuous and discrete variables can be incorporated with no prior transformation. This model requires the joint distribution of the continuous variables, for each configuration of the discrete ones, follows a multivariate Gaussian, which is not always true in the case of environmental data. Also, CG model imposes some restrictions to the DAG structure, where a discrete variable can not have continuous parents.

These restrictions have induced the development of other alternatives. The first alternative was the *Mixture of Truncated Exponentials* (MTE) model (Rumí, 2003; Moral *et al.,* 2001) which does not impose any restrictions on the network structure. They were defined by Moral *et al.,* 2001 and deeply developed by Rumí, 2003. This model overcomes the main limitation when dealing with hybrid BN, since it provides us with a common structure to represent both discrete and continuous variables in such a way that all the computations needed to perform probability propagation in the model can be done using the same structure. Also, this model is implemented in the software Elvira (Elvira-Consortium, 2002) including algorithms for *characterization*, *regression* and *classification* in the same package.

In contrast discretization, in which the domain of the variable is divided into several intervals and approximated by a constant function, when other functions with better properties are used, the accuracy of the model is improved (Rumí, 2003). This is the idea behind the MTE models, in which exponential functions are used due to their great fitting power. A step beyond this was the proposal of the *Mixture of Polynomials* model (Shenoy & West, 2011; Shenoy & West, 2009) and the *Mixture of Truncated Basis Functions* model (Langseth *et al.,* 2012) in which both used other functions were used.

One of the objectives of this Thesis is to encourage ecologists to apply BNs to real problems. Even when MoP and MoTBFs models have been proposed as promising solutions, they: *i)* are still under development, what implies that *ii)* algorithms are not totally developed for regression, classification and inference and also, *iii)* they are spread in different software packages which makes it difficult for ecologists to apply. These limitations are not present in MTE, that is the reason why they are used to represent pro-bability distributions all through this Thesis.

### 2.4.1 *Mixture of Truncated Exponential* Models

With the aim of performing inference in BN, the probability distributions need to be restricted, marginalized and combined, which can be easily done in MTE models. So, during this process, where the posterior distributions of the variables are obtained given some evidence, the intermediate functions are not necessarily density functions. Therefore a general function called *MTE potential* needs to be defined (Moral *et al.*, 2001):

**MTE potential** *Let $X$ be a mixed $n$-dimensional random vector. Let $\mathbf{Z} = (Z_1, \ldots, Z_d)^T$ and $\mathbf{Y} = (Y_1, \ldots, Y_c)^T$ be the discrete and continuous parts of $X$, respectively, with $c + d = n$. We say that a function $f : \Omega_X \mapsto \mathbb{R}_0^+$ is a* Mixture of Truncated Exponentials potential (MTE potential) *if one of the next conditions holds:*

i. $Z = \emptyset$ and $f$ can be written as

$$f(x) = f(y) = a_0 + \sum_{i=1}^{m} a_i \mathrm{e}^{\{b_i^T y\}} \tag{2.2}$$

for all $y \in \Omega_{\mathbf{Y}}$, where $a_i \in \mathbb{R}$ and $b_i \in \mathbb{R}^c$, $i = 1, \ldots, m$.

ii. $Z = \emptyset$ and there is a partition $D_1, \ldots, D_k$ of $\Omega_{\mathbf{Y}}$ into hypercubes such that $f$ is defined as

$$f(\mathbf{x}) = f(y) = f_i(y) \quad \text{if } \ y \in D_i,$$

where each $f_i$, $i = 1, \ldots, k$ can be written in the form of Equation (2.2).

iii. $Z \neq \emptyset$ and for each fixed value $z \in \Omega_Z$, $f_z(y) = f(z, y)$ can be defined as in ii.

**MTE density** An MTE potential $f$ is an *MTE density* if

$$\sum_{z \in \Omega_Z} \int_{\Omega_{\mathbf{Y}}} f(z, y) dy = 1.$$

A *conditional MTE density* can be specified by dividing the domain of the conditioning variables and specifying a MTE density for the conditioned variable for each configuration of splits of the conditioning variables. We can see this in the following example.

**Example** Consider two continuous variables $Y_1$ and $Y_2$. A possible conditional MTE density for $Y_1$ given $Y_2$ is the following:

$$f(y_1 \mid y_2) = \begin{cases} 0.28 + 0.01\mathrm{e}^{1.03y_1} + 0.02\mathrm{e}^{0.01y_1} & \text{if } 0 \leq y_1 < 1,\ 1 \leq y_2 < 3, \\ 0.02 + 0.02\mathrm{e}^{1.01y_1} + 0.12\mathrm{e}^{0.09y_1} & \text{if } 1 \leq y_1 < 3,\ 1 \leq y_2 < 3, \\ 0.49 - 0.12\mathrm{e}^{0.59y_1} - 0.24\mathrm{e}^{-0.08y_1} & \text{if } 0 \leq y_1 < 1,\ 3 \leq y_2 < 4, \\ 0.07 - 0.02\mathrm{e}^{-0.23y_1} + 0.62\mathrm{e}^{-0.23y_1} & \text{if } 1 \leq y_1 < 3,\ 3 \leq y_2 < 4. \end{cases}$$

In the same way as in discretization, the more intervals used to divide the domain of the continuous variables, the better the MTE models accuracy will be, but the complexity increases. Furthermore, in the case of MTE, using more exponential terms within each interval substantially improves the suitability to the real model, but again more complexity is assumed (Morales *et al.*, 2006). There are different approximation techniques that can be applied to obtain the result as MTE densities. In this Thesis, we have

followed the scheme presented in Moral *et al.,* 2003 and Rumí *et al.,* 2006, which produces MTE functions with high fitting power and low computational complexity. From these results, we are able to get approximations that are accurate, yet simple enough to allow the use of an exact inference algorithm (Rumí & Salmerón, 2007).

Model learning and inference are both addressed through this Thesis using Elvira software (Elvira-Consortium, 2002) and following the approach of Rumí *et al.,* 2006 to estimate the corresponding conditional distributions. Let $X_i$ and $Y$ be two random variables, and consider the conditional density $f(x_i \mid y)$. The idea is to split the domain of $Y$ by using the equal frequency method with three intervals. Then, the domain of $X_i$ is also split using the properties of the exponential function, which is concave, and increases over its whole domain. Accordingly, the partition consists of a series of intervals whose limits correspond to the points where the empirical density changes between concavity and convexity or decrease and increase.

At this point, a $5$-parameter MTE is fitted for each split of the support of $X$, which means that in each split there will be 5 parameters to be estimated from data:

$$f(x) = a_0 + a_1 e^{a_2 x} + a_3 e^{a_4 x}, \quad \alpha < x < \beta\,, \tag{2.3}$$

where $\alpha$ and $\beta$ define the interval in which the density is estimated.

The reason for using the $5$-parameter MTE lies in its ability to fit the most common distributions accurately, while the model complexity and the number of parameters to estimate is low (Cobb *et al.,* 2006). The estimation procedure is based on least squares (Romero *et al.,* 2006; Rumí *et al.,* 2006). In the case of models with more than one conditioning variable see Moral *et al.,* 2003 for more details.

## 2.5 Overview

A keyword search in ISI Web of Knowledge was carried out to illustrate the evolution of BN development and applications. The terms *Bayesian network* and *belief network* were used in the search from January 1990 to today (the last search done on $29^{th}$ June, 2016) resulting in $10301$ papers. Figure 2.4 shows the distribution of these papers in different research areas according to the ISI Web of Knowledge. Two research areas stand out from the others, *Mathematic and Computer Sciences* and *Engineering*, which can be considered mainly as areas in which the theoretical background, algorithms and software are developed (Lim *et al.,* 2016; Butz *et al.,* 2016; Boudali & Dugan, 2005; Ratnapinda & Druzdzel, 2015; Cheng *et al.,* 2002; Heckerman, 1997; Zhang & Poole, 1996; Heckerman *et al.,* 1995).

The rest of the research areas draw together several scientific fields in which BNs are directly applied to real life problems: *Health Sciences*, *Ecology and Environmental Sciences*, *Life Sciences*, *Social Sciences* and *Others* (that combine the field of Music, Business and Economy fields). Figure 2.5 shows the evolution over time of these research areas from 1990. Again, both *Mathematic and Computer Sciences* and *Engineering* show a quick and remarkable evolution against the rest of the areas.

The first applied areas were *Health science* and *Life Sciences* (van der Gaag, 1996) in which BNs have been successfully applied (Chai *et al.,* 2014; Mumford & Ramsey, 2014; Lee & Abbott, 2003). A skimming of papers included in these areas reveals that BNs

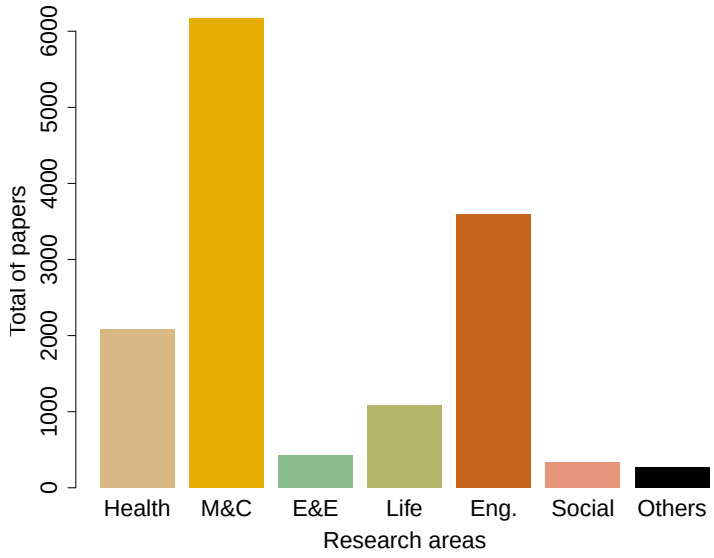



FIGURE 2.4: Number of paper in Research Areas of ISI Web of Knowledge from 1990 to nowadays. M&C, Mathematical and Computer Sciences; E&E, Ecology and Environmental Sciences; Eng., Engineering.

are much more advancely applied than in other scientific fields. For example, dealing with temporal data through the extension of BN, called Dynamic BN (explain in detail in Chapter 6), is still a challenge in Ecology and Environmental Science (Molina *et al.*, 2013), whilst an extensively applied tool in health and life sciences (Baur & Bozdag, 2015; Marini *et al.*, 2015; van Gerven *et al.*, 2008; Zou & Conzen, 2005).

Applied papers in *Ecology and Environmental Sciences* began in 1990 with one or two papers per year, but it was not until the beginning of 2000 when they started to be more frequently applied. In Aguilera *et al.*, 2011' s paper a thorough study was carried out taking more than 120 papers of this research area and analyzing their modeling goal, model structure, learning and validation process and softwares used. In this section, this study is updated with around 260 papers published during the period of time 1990-2016.

Firstly, papers are classified into several topics according to the ISI Web of Knowledge information and the content of each paper. Figure 2.6 shows the results of this classification in which *Water research* topic is prominent with respect to the rest. This is explained by the advantages that BNs provide as a tool for Decision Support System which encouraged scientifsic to apply them under the Integrated Water Resource Ma-nagement context (Castelletti & Soncini-Sessa, 2007b; Henriksen *et al.*, 2007). This has led to the application of BNs in some European projects such as the FP5-MERIT (Bromley *et al.*, 2005) or the NeWater (Henriksen & Barlebo, 2008). *Ecology and Environmental Science* topic is also remarkable, probably due to its more general and open contents where papers about general ecology or even socio-ecological systems are included (Young *et al.*, 2011; Milns *et al.*, 2010; Ticehurst *et al.*, 2007; Pollino *et al.*, 2007). *Biodiversity and Conservation* is the third topic in number of papers mainly for the use of BNs in species distribution and habitat suitability models (Boets *et al.*, 2015; Roberts *et al.*, 2013; Laws & Kesler, 2012). There are several topics with no more than 10 papers,

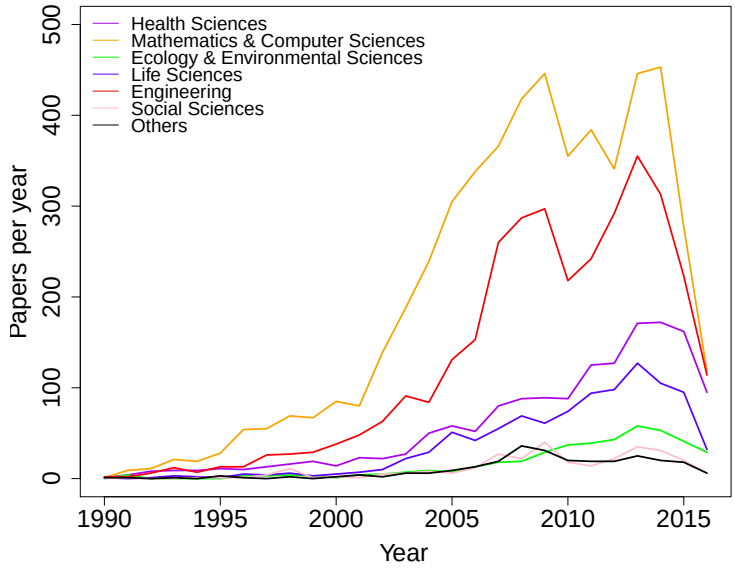



FIGURE 2.5: Evolution of the number of papers per year in each Research Area of ISI Web of Knowledge.

*Agriculture, Forestry, Fisheries, Marine and Freshwater Biology, Geology* and *Meteorological and Atmospheric Sciences*, where BNs have been scarcely applied nor tested. Special mention for the case of *Land Use* topic where despite the scarce number of papers, they are applying BNs with Geographic Information Systems improving predictive models about land use changes (Celio *et al.*, 2014; Aitkenhead & Aalders, 2009; Kocabas & Dragicevic, 2009; Grêt-Regamey & Straub, 2006). Finally, *General* topic refers to those papers that do not build any BN model but discuss their ability, potential applications in certain environmental problems, or even about how to learn and validate it, or measure uncertainty from BNs (Tiller *et al.*, 2013; Chen & Pollino, 2012; Marcot, 2012; Smith *et al.*, 2011; Lynam *et al.*, 2007; Pollino *et al.*, 2007; Uusitalo, 2007; Marcot *et al.*, 2006; Ricci *et al.*, 2003; Varis & Kuikka, 1999).

A thorough study of these papers reveals that BNs are partially used in Environmental Science and Ecology, where this powerful tool is mainly applied for characterization purpose (Figure 2.7), using discrete or discretized data (Figure 2.8). With respect to the modeling learning process, this review shows that a high percentage of paper using experts (alone or mixed with data) for learning the structure and parameters of the BNs (Figure 2.9), also, encouraged by several papers (Caley *et al.*, 2013; Castelletti & Soncini-Sessa, 2007a; Walton & Meidinger, 2006).

This tendency can be explained because there is a high percentage of papers used software with the need that data be discrete, and have an intuitive learning toolbox. In this way the author may not provide more information about the model learning scheme the software act as a black box. Besides, BNs have been used as a tool, not the objective of the paper, so the focus is on the environmental problem, not the details about the structure and parameters of the model. However, this lack of information about the model itself implies the non-reproductibility of the models presented.

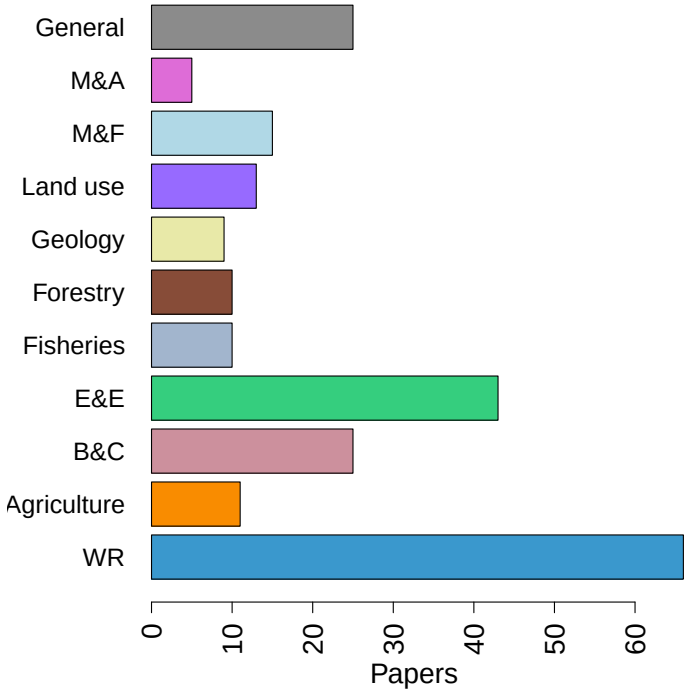



FIGURE 2.6: Total number of papers per topics in the *Ecology and Environmental Science* research area according to the classification done from the ISI Web of Knowledge. M&A, Meteorological and Atmospheric sciences; M&F, Marine and Freshwater Biology; E&E, Ecology and Environmental sciences; B&C, Biodiversity and Conservation; WR, Water research.

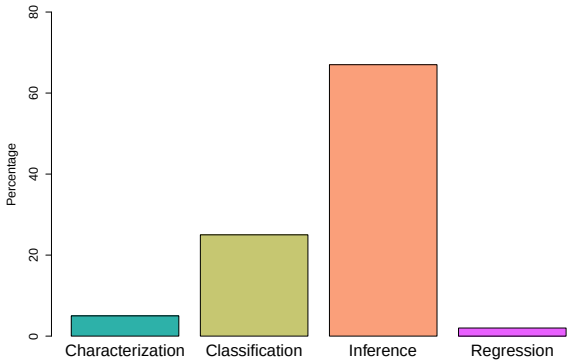



FIGURE 2.7: Aim of the model in the papers reviewed in the *Ecology and Environmental Science* research area. Papers included in the topic *General* are not considered.

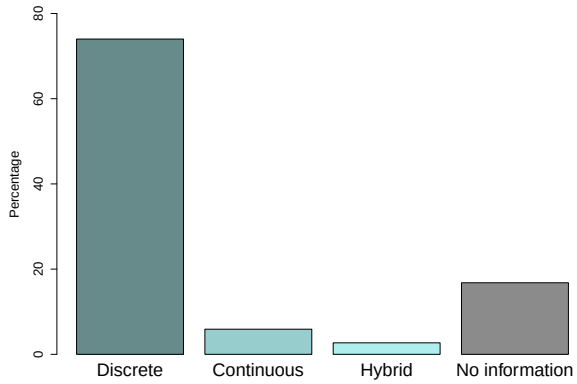



FIGURE 2.8: Percentage of papers that use discrete, continuous or hybrid data in the *Ecology and Environmental Science* research area. No information include those papers in which the information about the data used is not available. Papers included in the topic *General* are not considered.

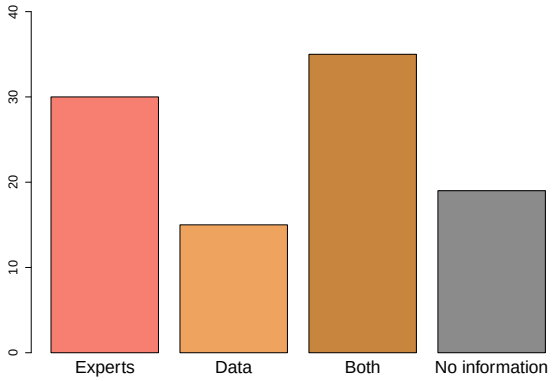



FIGURE 2.9: Use of expert, data or both during the model learning step in the papers reviewed in the *Ecology and Environmental Science* research area. Papers included in the topic *General* are not considered.

In general, further efforts are needed to expand the application of BNs in environmental sciences. Whilst in other areas this model is widely applied, the application in eco-logy is scarce and focused on a particular aim, without taking advantage of its strength. One potential of BNs hardly explored in ecology is their extension to deal with time series called Dynamic BNs. Fewer than 3% of the papers reviewed use this dynamic version of BNs, and those that applied it do not include continuous variables.

As a summary, most papers that applied BNs in ecology and environmental modeling use discrete or discretized data with the aim of learning models for characterization. Besides, experts and stakeholders are included in the model learning process more often than structural learning algorithms. The small percentage of papers that deal with hybrid domains (Figure 2.8) and dynamic datasets, encourage the aim of this Thesis. Trough this dissertation, BNs are based on hybrid datasets and, even when experts are included, model learning has been carried out with different algorithms that obtain the structure from the data, or even, fixed and constrained structures have been applied. Finally, not only characterization purpose was solved, but also, classification and regression problems.

# Part II

# Hybrid Bayesian networks in SES modeling

# Chapter 3

# Characterization: Modeling Uncertainty in Social-Natural Interactions

Characterization is the most usual problem in which BNs are applied in. In this Chapter, an Andalusian watershed is modeled as a SES and the interactions between social and natural subsystems (land use and water flow components) are studied using a hybrid BNs. The aim is to provide a new methodology to model systemic change in a socio-ecological context. Two endogenous changes, agricultural intensification and the maintenance of traditional cropland, are proposed. Besides, a methodology for scenarios comparison using the probability of the tails is presented. As a conclussion, intensification of the agricultural practices leads to a rise in the rate of immigration to the area, as well as to greater water losses through evaporation; whilst maintenance of traditional cropland hardly changes the social structure, while increasing evapotranspiration rates and improving the control over runoff water.

## 3.1 Introduction: Bayesian networks for characterization

In the mathematical field *Characterization* is the process of identifying the collection of properties that distinguish one object from the others. In our context, BNs for *Characterization* purpose aims to explore the behavior of the system modeled, and the nature and strength of the relationships between the variables. Besides, through the *inference* process, this model is able to study the behavior of the (socio)ecosystem modeled under different scenarios (Dyer *et al.,* 2014; Keshtkar *et al.,* 2013; Vilizzi *et al.,* 2013). BNs for *Characterization* purpose is the most applied model objective. As Figure 3.1 shows, they have been widely used in *Water Research*, *Ecology and Environmental Science* and *Biodiversity & Conservation* topics. In this sense, they were used for modeling scenarios of Climatic and Global change in different ecosystems and watershed (Dyer *et al.,* 2014; Mantyka-Pringle *et al.,* 2014; Webster & McLaughlin, 2014), changes in management plans of groundwater systems or species conservation (Shenton *et al.,* 2014; Tiller *et al.,* 2013), or environmental features that impact on species distribution patterns (Meineri *et al.,* 2015).

To model this kind of problems, first step is to learn the network structure, and two approaches are available: automatic and manual (or a mix of the two).

Automatic approach involves using training algorithms which calculate the optimal structure from the data, which can be tackled by different ways. On one hand, PC algorithm (Spirtes *et al.,* 1993) is based on conditional independence tests for discovering the (in)dependence relationships between the variables. By contrast, K2 method (Cooper & Herskovits, 1992) and its adaptation to MTEs (Romero *et al.,* 2006) propose looking for the optimal structure of a BN is an optimization problem in which the best solution is a set of all BNs, and select what best represent the dataset. In order to check if one is better than another, these methods used several measurements like the Bayesian Information Criterion (Schwarz, 1978). Once the structure is learnt, the parameters of the model are estimated from the data.

Manual approach means the use of experts for both structural learning and parameter estimation. One of the most important advantages of BNs is their ability to include this kind of information, what makes it a more commonly applied approach (Figure 3.2) than the automatic one. Using experts provides with several advantages from the natural resource management point of view (Voinov & Bousquet, 2010): stakeholders are included into management plans, and also encourage them to feel a part of the process, which provides a common language for interaction and finally, leads to more consensus and easier management decision. However, if the problem includes a wide number of variables, this approach can be so complex. In general, manual approach has been only used with discrete data (Figure 3.3), and several methodologies have been developed for modeling through experts. Some examples of them are the Participatory and Integrated Planning methodology, proposed by Castelletti & Soncini-Sessa, 2007a, or the Public Participation process, by Henriksen *et al.,* 2007. In these models, the process starts with an initial phase of problem discovering knowledge in which variables, and even their states, are identified. Next step consists on a set of phases more or less complex of surveys or participatory methodologies among the different stakeholders involve in the problem. Finally, model is presented to the participant (to the same that learn it, or new participant) and discussed.

Finally, a mix between both approaches are also used. In that case, the structure is usually learnt using expert knowledge but the parameters are estimated from the data, or also, some relations are estimate using expert (for discrete variables). Besides, experts are used to evaluate and improve the network obtained by the automatic approach.

## 3.2 Modeling uncertainty in social-natural interactions

SES is a complex system of interactions among nature and society operating at different spatial and temporal scales (Cadenasso *et al.,* 2006; Folke, 2006; Anderies *et al.,* 2004). Both natural and social systems contain several subsystems that interact between them and, also, have interactive subsystems. Such complexity supposes a challenge in the field of ecological modeling (Filatova *et al.,* 2013; Filatova & Polhill, 2012). Besides, SES is not a static system and two types of disturbances can be identified:

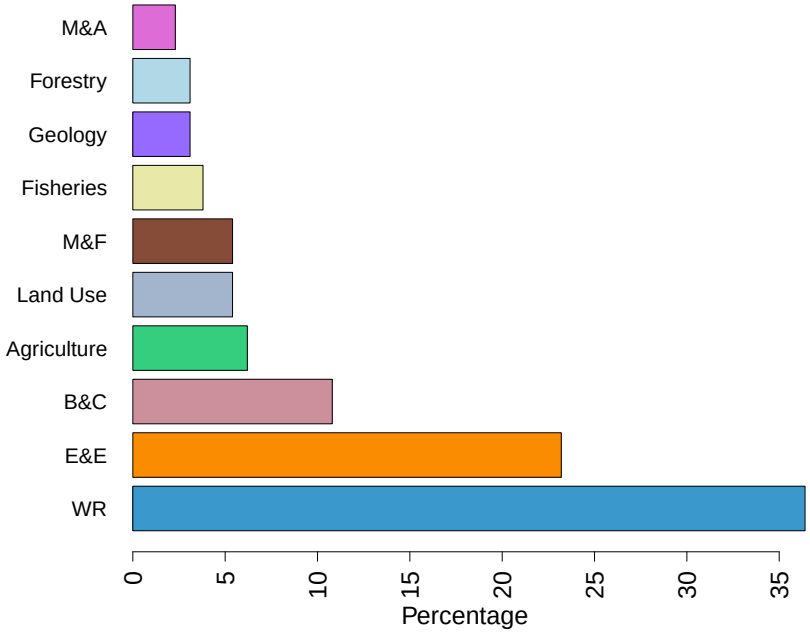



FIGURE 3.1: Percentage of papers that applied BNs for *Characterization* in each topic. M&A, Meteorological and Atmospheric sciences; M&F, Marine and Freshwater Biology; E&E, Ecology and Environmental sciences; B&C, Biodiversity and Conservation; WR, Water research

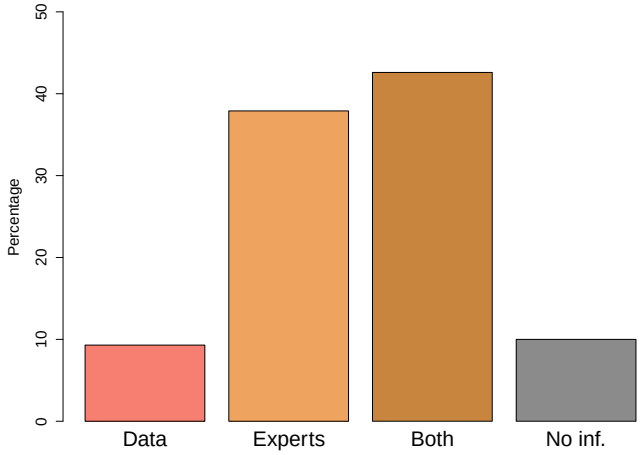



FIGURE 3.2: Percentage of papers that applied BNs for *Characterization* for each learning process.

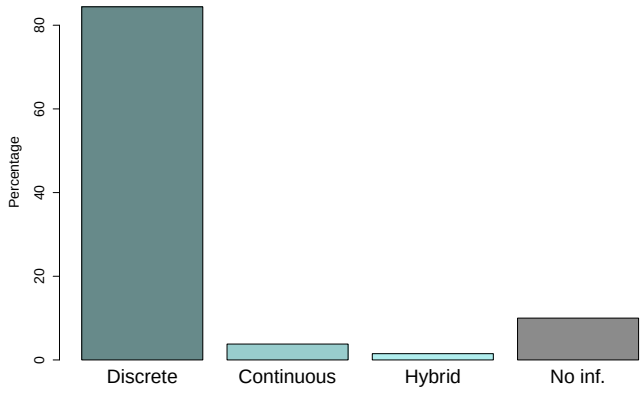



FIGURE 3.3: Percentage of papers that applied BNs for *Characterization* for each type of data

- External disturbance: a hazard event or shock (such as floods or earthquakes) or an extreme change in an input variable (*i.e* an extreme increase in the temperature) that provokes a rapid change in the system properties.
- Internal disturbance: a gradual change in the system's components that provokes a reorganization of the natural and social systems.

If these disturbances involve a fundamental change in the interactions within a system which leads to a shift in the state of the systems to another with new properties, we are dealing with a Systemic Change (SC) (Kinzig *et al.,* 2006; Filatova & Polhill, 2012).

In the context of SES and SC modeling, two main challenges can be identified (Filatova & Polhill, 2012): (i) accommodate the study of SC while taking uncertainty into account (Clark, 2002), and (ii) represent the new state of the system after SC has been propagated (Filatova & Polhill, 2012). The concept of uncertainty is widely used in ecology and environmental science but, the definition is sometimes not clear and also associated with the terms of error, risk or ignorance (Refsgaard *et al.,* 2007). Spite their importance, mainly when the model will be used to support decision making, there is a lack of understanding about its definition and characteristics (Walker *et al.,* 2003).

Under this perspective, several methodologies have been developed and defined to deal with uncertainty (Filatova *et al.,* 2013; Warmink *et al.,* 2010; Ricci *et al.,* 2003). In this Chapter our objective is focus on dealing with uncertainty in SES and SC modeling.

Graphically the SES can be represented as a network of nodes (social and natural components), with a number of links between them. When a hazard event occurs or a component undergoes a gradual change, the change can be propagated through the entire system by means of cause-effect interactions between the components of the SES. These types of interactions are subject to the uncertainty inherent in the system (Clark, 2002; Refsgaard *et al.,* 2007) which can be modelled using probability theory (Ricci *et al.,* 2003; Walker *et al.,* 2003; Refsgaard *et al.,* 2007; Warmink *et al.,* 2010).

Since BNs are modelled by means of probability distributions, uncertainty can be estimated more accurately than by using models which only consider mean values (Uusitalo, 2007). They allow a system to be represented both in its current state (*a priori*), and

*a posteriori*, once the change has been propagated through the system, using the probability distribution functions of the variables and the *inference* process. Their main purpose is to provide a framework for efficient reasoning about the system they represent, in terms of updating information about unobserved variables, when new information (changes to a single or several observed variables) is incorporated to the system.

However, not every change included into a component of the system (one or more variables) will lead to SC because some components may be conditionally independent. BNs are able to represent the independencies in the graph in a natural way, which makes them a highly appropriate tool to study SC. One of the main features of SC is that every change introduced into the system affects all the components (set of variables) involved in the system, rather than just some of them. This feature is difficult to model using more classical statistical tools; in contrast, the type of connections in the BN graph implicitly encodes this kind of situations. Not all inputs to the model would lead to a SC. Using the d-separation concept (Pearl, 1988a) it is possible to select the variables that connect different parts of the graph, allowing the SC to propagate all through the network. In the context of SC study and evaluation, the concept of *d-separation* is crucial, for that reason it is again in this chapter.

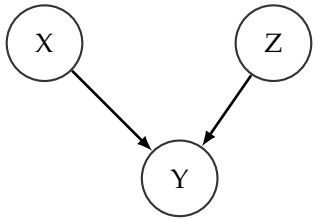



FIGURE 3.4: Example of two variables $X$ and $Z$ d-separated by $Y$

Figure 3.4 shows a simple example of the d-separation concept. In this situation, variables $X$ and $Z$ are d-separated by $Y$ i.e., $X$ and $Z$ are independent, if we do not know the exact value of $Y$, so an input in the model only for variable $X$ will not affect variable $Z$ (and viceversa) and so that input will not promote a systemic change. Figure 3.4 represents a very simple BN but the concept of d-separation is the same for larger BNs: given several parents for $X$ and $Z$ and children for $Y$ forming different components, then as long as $Y$ and its descendant are unknown, $X$ and $Z$ are independent, *i.e.*, any change in $X$ or its parents is not propagated to $Y$ or its parents (For more information see Pearl, 1988a and Jensen & Nielsen, 2007 Section 2.2.)

In this chapter, the aim is to demonstrate the ability of hybrid BNs to model SC considering a Spanish catchment as a SES. To identify SC a new methodology is proposed, which considers the tails of the probability distribution functions, and statistical tests were carried out to differentiate between different states of the system. By this means, this methodology provides the expert with a set of tools to help assess SC taking into account uncertainty.

### 3.2.1 Methodology

Figure 3.5 outlines the methodology followed in this chapter divided into four different steps: *i)* data collection, *ii)* model learning, *iii)* evidence propagation, and *iv)* analysis

of results. Elvira[1] software (Elvira-Consortium, 2002) was used to obtained the parameters of the model from the data and to carry out the evidence propagation.

### Study area

The study area comprises the catchment of the river Adra in south-eastern Spain (Figure 3.6). It is bounded to the north by the Sierra Nevada, to the south by the Mediterranean Sea, to the west by the Sierra de Gádor, and to the east by the Sierra Filabres. It occupies 74.400 Ha, and supports an estimated population of 124.000 people distributed over fourteen municipalities.

Figure 3.6 shows the main land uses of the watershed. In the North, the landscape of the Sierra Nevada mountain range is characterized by dense woodland, mainly oaks and conifers species, mixed with Mediterranean scrubland. This configuration is the results of several episodes of deforestation during the 19th century (García-Latorre & Sánchez-Picón, 2001), when Adra watershed supported an important mining activity. In this upper reaches, the relief of the mountains allows several patches of Mediterraneam woodland to be kept. Also, it provokes that socioeconomy in this upper area was characterized by small municipalities accommodating an ageing population with a high rate of migration. Moving down to the south, landscape is replaced by mixed - in which several small patches of rainfed, irrigated and scrub coexist - and irrigated croplands, and the population is also slightly younger, but still with a high movement. In the foothills the landscape also presents traditional croplands which configure an heterogeneous landscape of olive, almond and groves with small patches of woodland and scrub.

At the west of the area, we found the foothills of Sierra de Gádor where the landscape is totally different. Here, land use is mainly comprise by rainfed and mixed croplands whilst the socioeconomy is characterised by depopulation and an older population.

In the east of the study area, the landscape is composed by scrubs and some patches of woodland whose configuration was determined by historical trends in the 19th century (mining and the deforestation of natural forest) (García-Latorre & Sánchez-Picón, 2001). Finally, in the middle and the south of the catchment, landscape is mainly composed by scrubland and human infrastructures and the most important and biggest municipalities. Also, some intensive agriculture with greenhouses are located around the municipalities. In this area, population is younger than in the upper reaches, and also the economic systems is richer. Immigration rate is significant given the incoming of a new workforce to the greenhouses.

### Data collection

Table 3.1 shows the main statistics of the continuous socioeconomic and water flow variables in the data set. Taking into account socio-economic characteristics of the study area (Camarero *et al.*, 2009), three representative variables (ageing, emigration and immigration rates) were selected. Data on these variables were obtained for each municipality from the Andalusian Multiterritorial Information System [2] (Figure 3.5 *i)*).

[1] This is a free software based on JAVA. It can be found in *http://leo.ugr.es/elvira*

[2] http://www.juntadeandalucia.es/institutodeestadisticaycartografia/sima/index2-en.htm

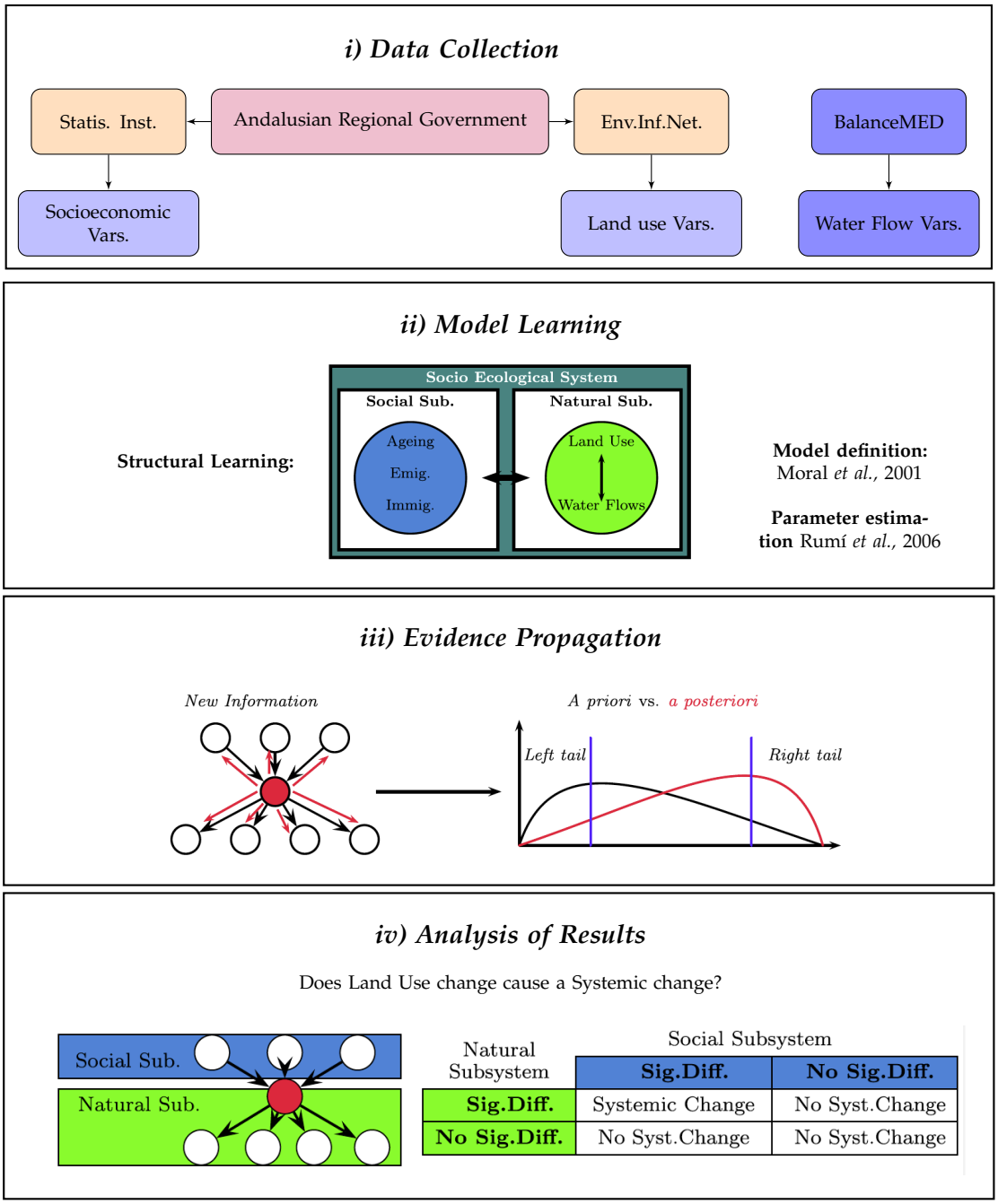



FIGURE 3.5: Outline of the methodology. Statis.Inst., Statistical Institute; Env.Inf.Net., Environmental Information Network; Vars., Variables; Sub., Subsystem; Sig.Diff., significant difference in statistical test; Syst.Change., Systemic Change; Emig, emigration rates; Immig, immigration rates

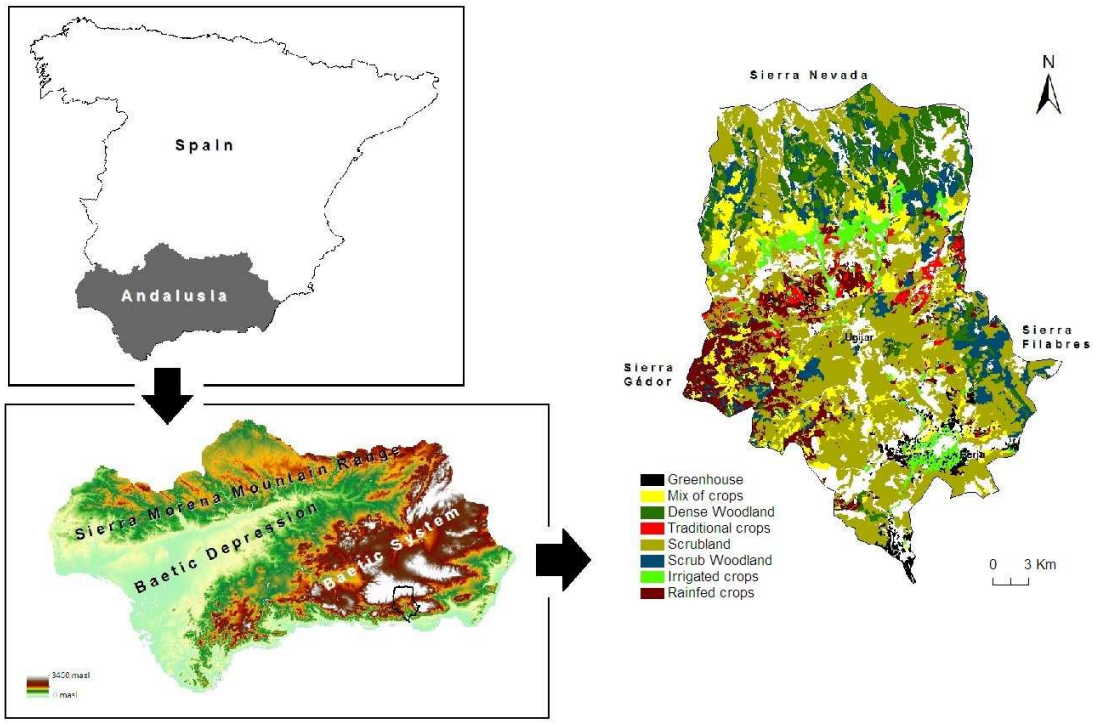



FIGURE 3.6: Study Area

The ageing component was expressed as the percentage of people older than 65 years old, while emigration and immigration rates were calculated as percentages of the total population.

The BalanceMED model (Willaarts, 2009; Willaarts *et al.*, 2012) was applied to calculate the water flows. This is a semi-deterministic model developed to quantify hydrological functioning in Mediterranean catchments using long time series of monthly rainfall and potential evapotranspiration data. It is based on the concepts of blue and green water (Falkenmark, 1997; Falkenmark & Folke, 2002):

**Blue water:** *is the amount of rainfall that exceeds the soil's storage capacity and feeds rivers, lakes and aquifers.*

**Green water:** *it refers to the rainfall that infiltrates into the root zone of the soil to support the primary productivity of natural and agricultural systems through evapotranspiration.*

The model assumes that a fraction from the total precipitation is intercepted by vegetation or soil and evaporates directly as a Non Productive Green Water (NPGW). Another fraction from the total precipitation can be intercepted on impermeable surfaces and is returned to the atmosphere as Consumptive Blue Water (CBW). The remaining precipitation reaches the soil and is taken up by plants and transpired, this portion is termed Productive Green Water flow (PGW). When the infiltrated water exceeds the soil storage capacity, it can either percolate or drain as Runoff Blue Water (RBW). In the specific case of greenhouse crops, we consider that the concept of PGW is not applicable since the crops are irrigated from groundwater flows rather than from direct precipitation. Moreover, evaporative flows are difficult to evaluate under a greenhouses cover. For that reason, in this specific case, we focus on CBW when considering greenhouse crops as the land use.

TABLE 3.1: Summary statistics of the continuous variables in the data set. SD, standard deviation.

| Variable | Minimum | Maximum | Mean | SD |
|---|---|---|---|---|
| Ageing (%) | 11.84 | 32.00 | 23.14 | 6.12 |
| Emigration rates (%) | 1.47 | 3.59 | 2.62 | 0.57 |
| Immigration rates (%) | 0 | 4.27 | 2.03 | 1.23 |
| PGW (mm) | 0 | 459.90 | 216.40 | 77.60 |
| NPGW (mm) | 0 | 346.70 | 62.10 | 55.91 |
| CBW (mm) | 0 | 765.5 | 58.48 | 176.51 |
| RBW (mm) | 0 | 1032 | 257.10 | 129.97 |

TABLE 3.2: Land uses selected and total percentage of each one in the study area.

| Land use | Description | % Surface |
|---|---|---|
| Scrub | Land devoid of trees and with more than 20% of scrub | 40.00 |
| Mixed crops | Mixture of crops (irrigated and rainfall-fed) with patches of natural vegetation | 14.69 |
| Rain-fed crops | Herbaceous and woody crops fed by rainfall | 10.93 |
| Dense woodland | Forest land with more than 50% of tree cover (conifers and oak) | 9.14 |
| Dense scrub woodland | Land with a tree cover of between 5 and 50%, and more than 50% of scrub | 7.15 |
| Greenhouses | Intensive greenhouse crops under plastic cover | 6.71 |
| Disperse scrub woodland | Land with a treecover of between 5 and 50%, with 20 to 50% of scrub | 5.73 |
| Traditional crops | Mixture of patchwork of olive, almond groves, grapevines, subsistence croplands, and forest (conifers and oak) | 3.57 |
| Irrigated crops | Herbaceous and woody crops with permanent irrigation infrastructure | 2.08 |

Nine land uses representative of the study area landscape were selected (Table 3.2). These data were obtained from the Land Use and Land Cover shape file 2007, from Andalusian Regional Environmental Information Network using ArcGis v.9.3.1 (ESRI, 2006) (Figure 3.5 *i)*). They are expressed as a discrete variable which represents the presence of each land uses as a percentage.

### Model learning

In this example, the structure of BNs was learnt taking into account expert knowledge and literature. Adra catchment is considered as a SES with a social and a natural systems divided into two main components: Land use and water flows. In literature it is widely recognized that global socio-economic changes affect regional and local socio-economic structures (Lambin *et al.*, 2001; Foley *et al.*, 2005) and lead to changes in the landscape (Schmitz *et al.*, 2005; Caillault *et al.*, 2013) and in the structure and functionality of natural ecosystems (Matson *et al.*, 1997; Foley *et al.*, 2005; Rudel *et al.*, 2009). One of the main effects of these changes relates to the behaviour of water flows (Scanlon *et al.*, 2005; Maes *et al.*, 2009; Toda *et al.*, 2010; Park *et al.*, 2014).

The concepts of green and blue water flows were defined to introduce the whole water cycle into water management plans (Falkenmark, 1997; Rockstroem, 2000). Green and blue water flow through natural subsystems across the landscape, participating in several ecological processes; as a result, there is a clear interaction between land use and green and blue water flows (Willaarts *et al.*, 2012). The characteristics of soil and the type and cover of vegetation determine the amount of water that evaporates back to the atmosphere, infiltrates into the soil or flows away as runoff (Falkenmark, 2003; Willaarts *et al.*, 2012).

In the model learning step, the structure of the BN is defined taking into account this theoretical background (Figure 3.5 *ii)*) and shown in Figure 3.7. Natural and social subsystems are connected through causal interactions where land use is influenced by the social subsystem and affected to water flows. Elvira software (Elvira-Consortium, 2002) was used to learn the parameters of the model from the data.

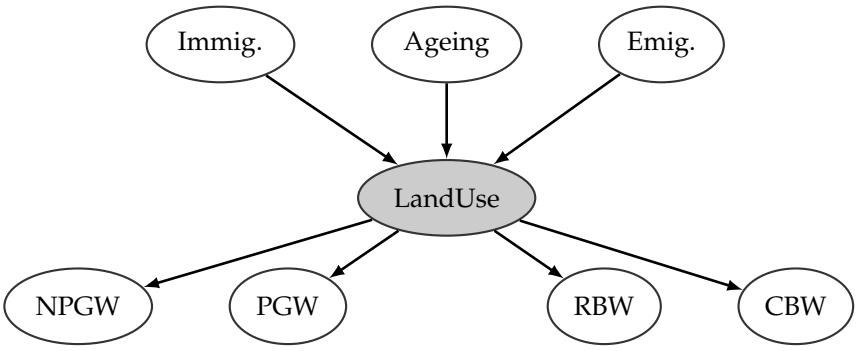



FIGURE 3.7: Qualitative part of Hybrid BN. By their nature every variable except one (*Land Use*) were continuous. Emig, emigration rates; Immig, immigration rates.

**Evidence propagation and analysis of results**

After model learning, next step is evidence propagation or inference (Figure 3.5 *iii)*). Since the network learnt is not so complex, an exact inference algorithm is applied, in this case Shenoy-Shafer algorithm (Shenoy & Shafer, 1990) which is able to deal with hybrid model and is specifically adapted to the MTE model (Rumí & Salmerón, 2007).

The evidence (new information) represents the change in one component of the SES which is propagated through the system and evaluated if it causes the SC (Figure 3.5 *iii)*). In this example, the influence of a gradual change in the Land Use component on simultaneously the social subsystem and the water flow component is studied.

The current state of the system, *"a priori"*, reflects the probability in the case where no new information is added to the system. Once the *"a priori"* situation is studied, the evidence is introduced as the presence of one of the states of the land use variable. Two different endogenous changes are selected: presence of traditional cropland and presence of greenhouses, which address both the land use trends that are observed in the study area.

Once the change is introduced into the model as evidence, we can take advantage of the versatility of BNs to obtain detailed results. The change alters the interactions in the

model, leading to changes in the distribution of each variable. The mathematical relationships that govern the interactions are expressed in a BN by means of conditional probability distributions, which are difficult to interpret. In contrast, the behaviour of the variables, both *a priori* and *a posteriori* is expressed through univariate probability distributions, which are much easier to interpret, especially for environmental systems. For this reason, changes are commonly quantified in terms of the mean value of the variable, which sometimes is not the most appropriate statistic to represent a probability distribution because it does not allow the overall behaviour of the variable to be tracked.

For a more comprehensive study of the results, how the proposed changes are propagated to the water flow component and to the social subsystem can be measured looking at the tail values of the probability distribution of each variable. In any probability distribution, the tails are highly relevant because they show the probability of the extreme values of the variable - in this case, the very high or very low water flows, and the very high or very low emigration and immigration rates and ageing. Firstly, the threshold values of the tails need to be defined since there are no references in literature to identify what constitutes an extreme value. For this purpose, a k-means clustering (Anderberg, 1973; Jain *et al.*, 1999) with 3 clusters was performed dividing the original data in three groups, according to their similarity. The first group was considered as the left tail, the second group as the centre of the distribution, and the third group as the right tail. Accordingly, the upper and lower thresholds were determined as the points that separated the first and the third cluster from the second. Once the thresholds were obtained, we computed the cumulative probability of both the left (lower) and right (upper) tails of all the water flow and social variables. As an example, Figure 3.8 shows the computation of these cumulative probabilities and the degree of change for a variable X in the *a priori* and *a posteriori* scenarios. Using the k-means clustering method, the left tail threshold was determined as $X < 37$, and the right tail threshold was determined as $X > 59$. Then the cumulative probability of the tails were computed, and we can see for example that $P(X > 59) = 0.34$ *a priori*, but it decreases to $0.13$ *a posteriori*.

**Validation of the model**

Validation of BNs depends on the aim of the model (Aguilera *et al.*, 2011). In the case of *Characterization*, experts or comparison with models that try to solve the same problem are appropriate validation techniques. In this example, there are no other models relating socioeconomy-land use and green and blue water flows under a SES framework. Therefore, validation by experts is considered to be an appropriate option.

However, as a way of validating the conclusions drawn from the results obtained, several goodness-of-fit tests were performed. To determine whether there were significant differences between the variables *a priori* and *a posteriori* (that is, between the different states of the system) a sample of size 1000 from each of the *a priori* and *a posteriori* probability distribution functions were simulated and a two-sided Kolmogorov-Smirnov test at a 0.05 level of significance was carried out. If significant differences are found between the system state *a priori* and *a posteriori* in both social and natural subsystems, the change introduced in the model can be considered as a SC (Figure 3.5 *iv)*).

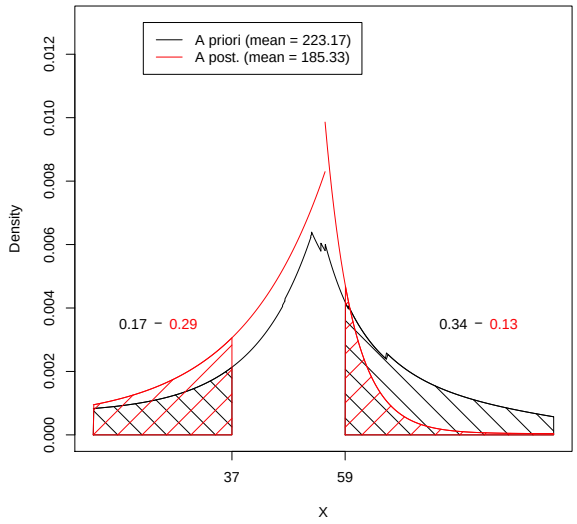



FIGURE 3.8: An example of the probabilities of the left (P($X < 37$)) tail, and the right (P($X > 59$)) tail in variable X *a priori*, in black, and *a posteriori*, in red, and its mean values between brackets. As we can see, mean value decreases which indicates the shift of the distribution function to the left. Probabilities in the tails show us that the change of the function is much more intensive in the right tail than in the left.

### 3.2.2 Results and Discussion

Table 3.3 shows the mean and standard deviation values of the variables both in the current situation (*a priori*), and after the two land use changes were simulated (*a posteriori*). Tables 3.4 and 3.5, show the probability in the tails and the p-values of the two-sided Kolmogorov-Smirnov tests respectively. Figures 3.9, 3.10, and 3.11 show the probability distributions of social and water flow variables in the current situation, and under both land use change scenarios.

#### *A priori*

*A priori* shows the current situation without any change introduced in the system. The ageing variable has a mean value of $21.65\%$, while emigration and immigration rates are $2.56\%$ and $1.87\%$, respectively (Table 3.3). Social variables are more probable in the left tail (Table 3.4, Figure 3.9).

Likewise, the probability of green water, (both productive and non productive), are more probable in upper values (Table 3.4, Figure 3.10), with mean values of $223.17$ mm for PWG, and $123.04$ mm for NPGW (Table 3.3). RBW has the same behaviour, with a mean value equal to $454.74$ mm (Table 3.3), and increased probabilities of falling into the right tail, $0.53$ (Table 3.4, Figure 3.11). By contrast, CBW probabilities increase in the left tail, with $0.64$ of probability (Table 3.4, Figure 3.11), and $209.32$ mm mean value (Table 3.3).

TABLE 3.3: Mean and standard deviation (SD) values of water flow and social variables obtained from the *a priori* and *a posteriori* probability distribution.

| | | | ***A posteriori*** | |
|---|---|---|---|---|
| **Variable** | **Statistics** | ***A priori*** | **Greenhouses** | **Traditional** |
| Ageing (%) | Mean | 21.65 | 17.94 | 21.88 |
| | SD | 5.82 | 3.53 | 5.85 |
| Emigration (%) | Mean | 2.56 | 2.57 | 2.51 |
| | SD | 0.58 | 0.58 | 0.56 |
| Immigration (%) | Mean | 1.87 | 1.92 | 1.68 |
| | SD | 1.16 | 1.16 | 1.30 |
| PGW (mm) | Mean | 223.17 | - | 284.39 |
| | SD | 95.19 | - | 76.25 |
| NPGW (mm) | Mean | 123.04 | 173.97 | 151.85 |
| | SD | 95.58 | 87.32 | 74.85 |
| CBW (mm) | Mean | 209.32 | 559.09 | 124.48 |
| | SD | 216.89 | 200.33 | 207.91 |
| RBW (mm) | Mean | 454.74 | 537.35 | 260.75 |
| | SD | 235.43 | 336.78 | 209.91 |

TABLE 3.4: Threshold of left and right tails, and probability values in the tails of water flow, and social variables for the current situation (*a priori*) and both land use changes *(a posteriori)*. As an example, in the ageing variable, 0.39 *a priori* is the probability of having fewer than 19.08% of people older than 65 years old, while 0.17 is the probability of having more than 28.44% of people older than 65 years old. The thresholds in social variables are expressed as a percentage of the population; thresholds for water flows are in mm. PGW under greenhouse is not calculated since evaporative flow is considered as CBW.

| | | | **Probability** | | |
|---|---|---|---|---|---|
| **Variable** | **Threshold** | | ***A priori*** | **Greenhouses** | **Traditional** |
| Ageing | Left tail | 19.08 % | 0.39 | 0.62 | 0.38 |
| | Right tail | 28.44 % | 0.17 | 0.004 | 0.18 |
| Emigration | Left tail | 2.58 % | 0.52 | 0.52 | 0.56 |
| | Right tail | 3.06 % | 0.25 | 0.25 | 0.21 |
| Immigration | Left tail | 1.18 % | 0.33 | 0.31 | 0.39 |
| | Right tail | 2.91 % | 0.21 | 0.22 | 0.17 |
| PGW | Left tail | 138.53 mm | 0.17 | - | 0.049 |
| | Right tail | 251.61mm | 0.24 | - | 0.70 |
| NPGW | Left tail | 46.41 mm | 0.36 | 0.09 | 0.04 |
| | Right tail | 115.73 mm | 0.45 | 0.72 | 0.62 |
| CBW | Left tail | 140.9 mm | 0.64 | 0.049 | 0.80 |
| | Right tail | 506.53 mm | 0.21 | 0.67 | 0.11 |
| RBW | Left tail | 216.37 mm | 0.15 | 0.26 | 0.50 |
| | Right tail | 400.83 mm | 0.53 | 0.59 | 0.20 |

TABLE 3.5: P-values of Kolmogorov-Smirnov test among simulated values from *a priori* and *a posteriori* distribution functions. *The distribution functions are significantly different at a 0.05 level of significance.

| Variable | Greenhouses | Traditional |
|---|---|---|
| Ageing | $2.2e^{-16}$ * | 0.2634 |
| Emigration | 0.8593 | 0.001227 * |
| Immigration | 0.00060* | 0.1205 |
| PGW | - | $2.2e^{-16}$ * |
| NPGW | $2.2e^{-16}$ * | $2.2e^{-16}$ * |
| CBW | $2.2e^{-16}$ * | $2.2e^{-16}$ * |
| RBW | $7.05e^{-16}$ * | $2.2e^{-16}$ * |

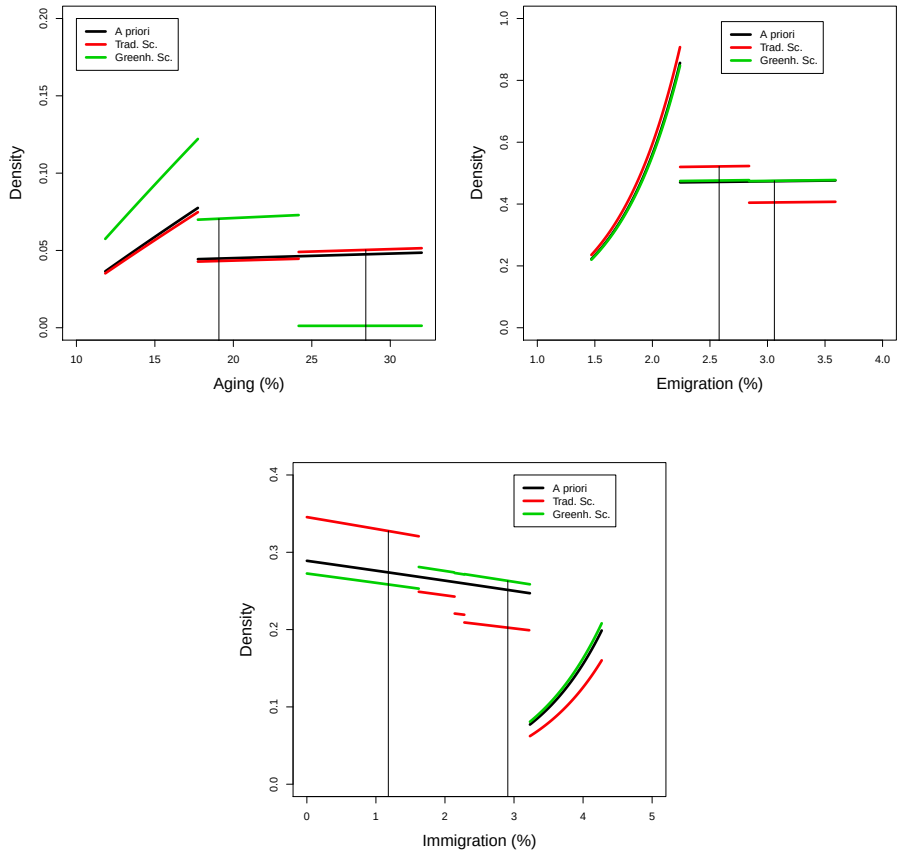



FIGURE 3.9: Probability distribution of social variables *a priori* and after both land use changes *(a posteriori)*. The vertical black lines represent the threshold values of the tails of the variables. Note that probability functions are defined as a piecewise function using MTEs.

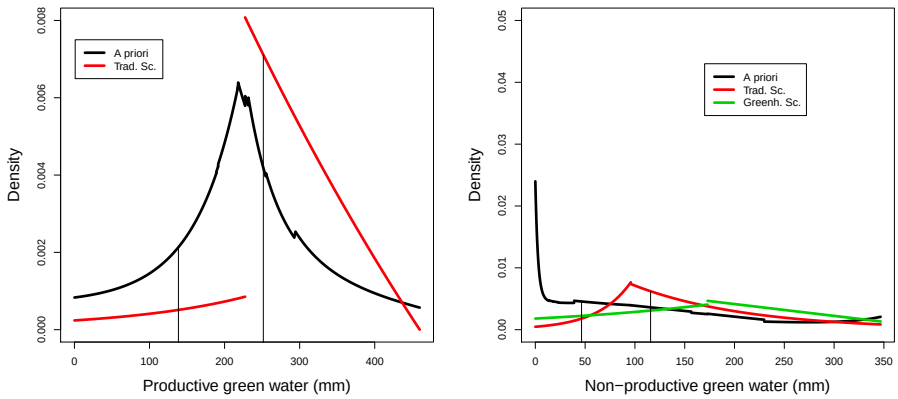



FIGURE 3.10: Probability distribution of green water flow variables *a priori* and after both land use changes *(a posteriori)*. PGW under greenhouse is not calculated since evaporative flow is considered as CBW. The vertical black lines represent the threshold values of the tails of the variables. Note that probability functions are defined as a piecewise function using MTEs.

In this state of the system, the population is ageing, and both emigration and immigration rates are low. The structure of the landscape determines that RBW and PGW are the main water flows.

**Scenario: Intensive agriculture with greenhouses**

Intensive agriculture with greenhouses is one of the most important economic activities in the south-east of Spain, and it can impact both social and natural subsystems (IEC, 2004). In the study area, greenhouses are mainly located in the lower reaches, where population is characterized by a significant immigration rate.

Under the first scenario of an increase in intensive agriculture, ageing mean value decreases from 21.65% to 17.94% (Table 3.3). A look at probability values in the tail shows the change more clearly than the mean value. The decreasing trend in this variable is more noticeable when the tails of the distribution are studied. This indicates that there is little probability of a population with greater than 28% over 65 (in the left tail, probability decreased from 0.17 to close to zero (Table 3.4, Figure 4.8)). On the other hand, immigration mean value increases from 1.87% to 1.92% (Table 3.3). In this case, the probability values of the tails are also small (Table 3.4, Figure 3.9). Furthermore, both variables show significant differences between the *a priori* and the new scenario (Table 3.5). By contrast, as Figure 3.9 shows, emigration rates hardly changes (Table 3.3, and 3.4) which is confirmed by the two-sided Kolmogorov-Smirnov test (Table 3.5).

The incoming young population has the effect of reducing the extreme values of the ageing variable (*i.e.* the proportion of people over 65 falls). By contrast, emigration hardly changes (*i.e.*, the departure of people looking for a job elsewhere is virtually unchanged). However, the behavior of the socioeconomic subsystems changes as a result of the intensification (García-Álvarez-Coque, 2002). These results concord with numerous studies made by different Spanish economic entities (CCA, (Colección Comunidades Autónomas), 2007; IEC, 2004; García-Álvarez-Coque, 2002); which show that

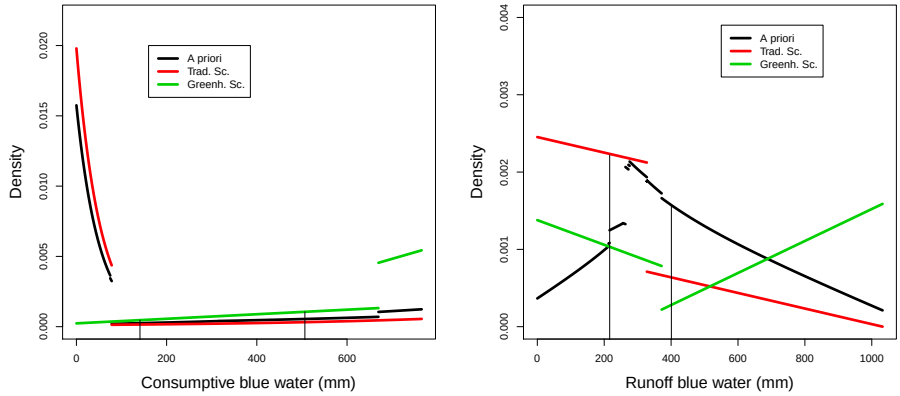



FIGURE 3.11: Probability distribution of blue water flow variables *a priori* and after both land use changes *(a posteriori)*. The vertical black lines represent the threshold values of the tails of the variables. Note that probability functions are defined as a piecewise function using MTEs

the agricultural intensification in the south-east of Andalusia has led to an increased influx of foreigners, mainly young people, to work in the greenhouses. This has reversed the trend of an increasingly ageing population, and has also led to an increase in the birth rate so changing the social structure of the area.

In the case of water flow, PGW is not calculated since evaporative flows from a greenhouse surface is difficult to evaluate and separate from CBW. For that reason, from greenhouse surface evaporative flow is considered as CBW. The means of the rest of water flow variables increase (Table 3.3), and there are significant differences in the distribution functions between *a priori* and under this scenario (Table 3.5). If only mean values were taken into account, the behavior of blue and green water flows can be considered similar. However, there is a marked difference if the probabilities in the tails are considered. In the case of NPGW, the right tail probability increases from $0.45$ to $0.72$ (Table 3.4, Figure 3.10) which only emphasizes the *a priori* behavior, *i.e.* higher water flows are more likely than lower ones. But in CBW, a marked change in the trend is predicted. The probability of the left tail (extremely low flows) decreases from $0.64$ to $0.049$; while probability of right tail values (extremely high flows) increases from $0.21$ to $0.67$. Vegetation cover around the greenhouses is often eliminated (to avoid invasion of pests into the greenhouses), so evaporation rates from the bare soil and plastic surfaces (greenhouses cover)-, described as NPGW and CBW, respectively- increase. As agriculture intensifies, NPGW becomes more important. However, while CBW in the *a priori* situation was low, the increase in greenhouse cover increases CBW quite significantly.

Similarly, the mean RBW value increases from which one might expect an increase in the right tail probability, and a decrease in the left tail. However, the probability in both tails increases (Table 3.4 and Figure 3.11). This shows a peculiar behaviour in the variable, which means that both high and low extremes values of runoff become more probable than *a priori*, whilst the moderate values are less probable.

**Scenario: Traditional agriculture**

In our study area, traditional croplands comprise a mixture of woody rain-fed crops of olive, almond groves and grapevines with patches of herbaceous subsistence crops and natural vegetation managed in a traditional way. This heterogeneous pattern of traditional land use has been promoted as an alternative management system, which can bring economic and environmental benefits (Schmitz *et al.*, 2005; Anderson *et al.*, 2009). Such croplands are found mainly in the Sierra de Gádor foothills, which is a landscape characterized by an ageing population and depopulated municipalities.

The land use change introduced into the SES is expressed as the greater presence of this traditional croplands. Mean ageing and immigration hardly change, nor does their probability distributions with respect to the *a priori* situation (Tables 3.3 and 3.5, Figure 3.9). By contrast, the emigration variable shows a significant difference between the *a priori* and the new scenario, with slightly higher probabilities in the left tail (Table 3.4 and 3.5, Figure 3.9). The presence of traditional croplands does not imply a new incoming population, nor the emigration of young people and so neither ageing nor immigration change significantly from their *a priori* values. Although emigration changes (Table 3.5), it does not imply an alteration of the global behavior of the socio-economic subsystem (CCA, (Colección Comunidades Autónomas), 2007).

For the four water flow variables, there are significant differences between the *a priori* situation and this scenario (Table 3.5). Both PGW and NPGW mean values increase, from $223.17$ mm to $284.39$ mm, and from $123.04$ mm to $151.85$ mm respectively (Table 3.3). Again, the study of the tails provides additional information about whether the extremes of the distribution become more or less pronounced. As we can see in Table 3.4, the probability of high PGW in the right tail shifts from $0.24$ to $0.70$ while the mean shows a more moderate increase. This means that, under this second scenario, extremely high PGW flows are 46% more probable than *a priori*. Given that traditional croplands are a mixture of woody and herbaceous crops and scrub, with areas of forest, the PGW is higher (Willaarts, 2009) because the evaporative demand of woody vegetation is higher than for herbaceous. Patches of scrubland and woodland, as well as the olive and almonds groves increased, and imply an increase in the PGW flow.

In the same way, the NPGW left tail (the probability of extremely low NPGW) decreases from $0.36$ to $0.04$. However, the shift in the mean is proportionally less, from $123.04$ mm to $151.85$ mm. These traditional systems are characterized by an absence of bare solid, and a presence of herbaceous crops, which explain the increases in NPGW (Rockstroem, 2000). At the same time, it involves a markedly decreases in CBW from $209.32$ mm to $124.48$ mm, with a shift in the left tail from $0.64$ to $0.80$ (Table 3.4).

In contrast, the RBW left tail probability increases from $0.15$ to $0.50$, giving more information about the extent of the change in this variable (Table 3.4, Figure 3.11). In this case, the evidence introduced in the model implies the change in the tendency from an *a priori* situation where runoff was quite probable in higher values (probability of the right tail decrease from $0.53$ to $0.20$), to a situation in which lower values are more pro-bable. Agriculture heterogeneity involves a tighter control over RBW since the structure of this Mediterranean multifunctional rural landscape with its patchwork of different types of land use, together with the presence of mature ecosystems next to exploited plots, favours this control of runoff (De-Lucio-Fernández *et al.*, 2003; Anderson *et al.*, 2009).

## 3.3 Conclusion

In this Chapter, the aim is to show the applicability of BNs to model SES in general and SC, in particular. Using BNs for a *Characterization* purpose, the relationships between different components of a SES can be modeled and easily interpreted through the qualitative part of the network. Besides, modifications in the interactions between the different components of a SES can be assesed through the study of the variables since they are affected by any change in the interactions and also, are easier to interpret.

The versatility of BNs allows several statistics to be calculated from the results of the variables. In this case, mean values and the probability of the tails were calculated. Although mean values provide clear information about the behavior of the variables, the tails allow the extent of the changes to be assessed.

BNs are able to deal with probability propagation, since new information can be introduced into one or more components of the natural or social subsystems and the effects over the rest of the SES can be inferred. Therefore, the current situation and the new system state can be easily compared because the model results can be displayed together in a single graph showing changes in probability distribution (see Figures 3.9, 3.10, and 3.11). In summary, the probabilities are updated when new information is incorporated into the model and they can be analyzed to evaluate the systemic change in SESs.

The results demonstrate how water flows are modified when soil and natural vegetation cover are lost due to the intensive agriculture activity. The increase in evaporative losses reduces the water available for human and agricultural supply (thus, in semiarid regions such as this, efforts need to focus on optimizing water use and minimizing water losses). Moreover, increase in runoff flows can alter soil structure due to increased erosion. Agricultural intensification leads to greater homogeneity in the landscape, and a loss of connectivity (the capability of the landscape to facilitate biophysical flows), (Taylor *et al.,* 1993), which implies poorer control of the nutrient and water cycle (De-Lucio-Fernández *et al.,* 2003). Agricultural intensification is a gradual trend that significantly modifies both natural and social subsystems, creating a new state in the system. Thus, it can be considered as a SC from the expert's point of view.

By contrast, in terms of whether an increase in traditional croplands can be described as a systemic change or not, we can say that SC can be defined as a fundamental change which involves a shift in the system state to another with new properties (Kinzig *et al.,* 2006; Filatova & Polhill, 2012). The model indicates that this does not happen under this second scenario and so increasing traditional agriculture cannot be considered as a SC from this point of view.

# Chapter 4

# Regression: Modeling Landscape - Socioeconomy Relationships

Regression through BNs is not often applied in environmental sciences. The aim is to give a prediction of a response variable given the value of some feature variables. Multiple linear regression models are more generally used. However, they have a number of limitations: (1) all feature variables must be instantiated to obtain a prediction, and (2) the inclusion of categorical variables usually yields more complicated models. Hybrid BNs are an appropriate approach to solve regression problems without such limitations, and they also provide additional advantages. In this Chapter, landscape - socioeconomy relationships are modeled by BNs for different types of data (continuous, discrete or hybrid), and is compared with Multi-Linear Regression. Three models relating socioeconomy and landscape are proposed, and two scenarios of socioeconomic change are introduced in each one to obtain a prediction.

## 4.1 Introduction: Bayesian networks for regression

In the study of environmental systems, it is common to find problems in which the goal is to predict the value of a continuous variable of interest depending on the values of some other features, facing with a *regression* problem (Hastie *et al.*, 2009):

**Regression:** *Let have a set of variables $Y, X_1, \ldots, X_n$, regression analysis consists on finding a model $g$ that explains the response variable $Y$ in terms of the feature variables $X_1, \ldots, X_n$, so that given a full observation of the features $x_1, \ldots, x_n$, a prediction about $Y$ can be obtained as $\hat{y} = g(x_1, \ldots, x_n)$.*

BNs have been proposed for regression purposes adding some advantages in comparison with traditional methodologies, since it is not necessary to have a full observation of the features to give a prediction for the response variable, and the model is usually richer from a semantic point of view.

However, there are just few attempts in literature to solve regression problems by BNs in environmental science, and these are focused on discrete feature variables or Gaussian distributions unable to handle discrete and continuous variables simultaneously without constraints on the structure (Malekmohammadi *et al.*, 2009; Pérez-Miñana *et*

*al.,* 2012). In real life problems, features can be either continuous or discrete, what adds a challenge to the traditional methodologies. For that reason, regression model based on BNs with the approximation of the joint distribution by an MTE was proposed (Morales *et al.,* 2007). Just only two papers deal with regression using MTE models in environmental modeling (Maldonado *et al.,* 2016; Maldonado *et al.,* 2015).

A BN can be used as a regression model for prediction purposes if it contains a continuous response variable $Y$ and a set of discrete and/or continuous feature variables $X_1, \ldots, X_n$. Thus, in order to predict the value for $Y$ from $k$ observed features, with $k \leq n$, the conditional density

$$f(y \mid x_1, \ldots, x_n), \tag{4.1}$$

is computed, and a numerical prediction for $Y$ is given[1] using the expected value as follows:

$$\hat{y} = g(x_1, \ldots, x_n) = \mathbb{E}[Y \mid x_1, \ldots, x_n] = \int_{\Omega_Y} y f(y \mid x_1, \ldots, x_n) dy, \tag{4.2}$$

where $\Omega_Y$ represents the domain of $Y$.

Note that $f(y \mid x_1, \ldots, x_n)$ is proportional to $f(y) \times f(x_1, \ldots, x_n \mid y)$, and therefore, solving the regression problem would require a distribution to be specified over the $n$ variables given $Y$. The associated computational cost can be very high. However, using the factorisation determined by the network, the cost is reduced.

Although the ideal would be to build a network without restrictions on the structure, for *regression* (and *Classification*) purposes, constrained structures were defined. Their objective is to accurately estimate the distribution of the response variable rather than the joint probability distribution of all features. NB was slightly explained in the Introduction (Chapter 2), but it is thoroughly described below since it is needed in this application.

The extreme case of constrained structures, is the so-called NB structure (Friedman *et al.,* 1997; Duda *et al.,* 2001). It consists of a BN with a single root node and a set of features having only the response variable as a parent (Figure 4.1). Its name comes from the naive assumption that the features are considered independent given the response variable $Y$. This strong independence assumption is somehow compensated by the reduction in the number of parameters to be estimated from data, since in this case, it holds that

$$f(y \mid x_1, \ldots, x_n) \propto f(y) \prod_{i=1}^{n} f(x_i \mid y), \tag{4.3}$$

which means that, instead of one $n$-dimensional conditional distribution, $n$ one - dimensional conditional distributions are estimated. Despite this extreme independence assumption, the results are competitive with respect to other models (Friedman *et al.,* 1997).

[1]Note that in the BN framework, a prediction of $Y$ can be obtained even when some of the variables are not observed.

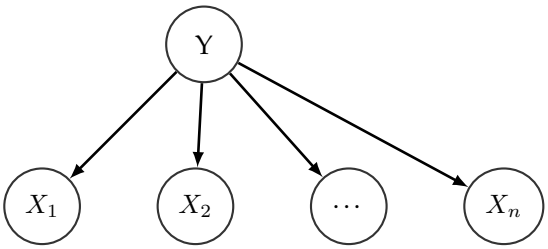



FIGURE 4.1: Structure of a *naïve Bayes* model.

However, if some variables are highly correlated, the error in the regression would decrease if any dependence between them could be included in the network (*i.e.*, links between features). There are several structures in which each feature is permitted to have more parents beside $Y$, for instance, TAN (Friedman *et al.,* 1997), FAN (Lucas, 2002), $k$DB (Sahami, 1996) or AODE (Webb *et al.,* 2005). These models are richer but an increase of complexity is assumed instead, both in the structure and the probability learning.

In general, including more variables does not necessarily increase the model accuracy since some variables are not informative for the prediction task, and therefore including them in the model provides noise to the predictor. Thus, a priori selection of the features would increase the model accuracy, and also decrease its complexity (less variables implies less parameters to be estimated). There are different approaches to the feature selection problem:

- The *filter* approach (Ben-Bassat, 1982), which in its simplest formulation, consists in establishing a ranking of the variables according to some measure of relevance with respect to the class variable, usually called *filter measure*. Then, a threshold for the ranking is selected and those variables below that threshold are discarded.
- The *wrapper* approach (Kohavi & John, 1997) proceeds by constructing several models with different sets of feature variables, and selecting the model that gives the highest accuracy.
- The *filter-wrapper* approach (Ruiz *et al.,* 2006) is a mixture of the above two options. First of all, the variables are sorted using a filter measure and then, using that order, they are included only if they increase the accuracy of the current model.

## 4.2 Modeling landscape - socioeconomy relationships

Under the SES framework, landscape and socioeconomic structures maintain a constant and reciprocal interaction configuring a “co-evolving system" (**Schmitz03**; Norgaard, 1984; Turner *et al.,* 1988; Lacitignola *et al.,* 2007). Thus, socioeconomic processes, as drivers of change (Burgi *et al.,* 2004), are the main cause of changes in land uses, *i.e.*, it determines the structure, function and dynamics of landscapes (Bicik *et al.,* 2001; Wu & Hobbs, 2002). In Mediterranean areas, this co-evolution is easy to observe in the landscape patterns where human-nature interaction has determined the so-called “agro-silvo-pastoral" systems (Sánchez-Picón *et al.,* 2011; Schmitz *et al.,* 2003; García-Latorre & Sánchez-Picón, 2001). These systems integrate the agricultural and livestock

infrastructures within the natural system in a traditional way. This systems support and important biodiversity rates (Pineda & Montalvo, 1995).

However, European agricultural landscapes have been undergoing significant changes associated with intense and rapid socio-economic changes (Nikodemus *et al.*, 2005; Strijker, 2005). In Europe, and particularly in Spain, socioeconomic development has led to a notable migration of the rural population to the city, and the depopulation of the countryside which supposes the increase of the scrubland surface. In some occasions this economic change also involves the substitution of the culture heritage from extensive traditional systems to more intensive agricultural systems. These tendencies imply a reduction in biodiversity rates, a loss in the ecological connectivity and less control of the physic-chemical flows into ecosystems.

Modeling environment-human relationships are becoming increasingly important and it has been applied in decision-making processes (Wang & Zhang, 2001; Serra *et al.*, 2008; Milne *et al.*, 2009; Celio *et al.*, 2014). More specifically, the relationships between landscape structure and socioeconomy have been formalized through Multiple Linear Regression (MLR) (**Schmitz03**; Schmitz *et al.*, 2005). This procedure provides a dependence model with a limited number of socioeconomic variables, which themselves can account for much of the variation in the landscape structure.

The objective of this Chapter is to develop a regression model based on a hybrid BN that can be applied to study landscape - socioeconomy relationships. In literature there are no studies about this relationship based on hybrid BNs. In this case, we are facing with an internal disturbance in the SES; social system follows a gradual change that shift the system from one point to another with different characteristics. In particular, this Chapter is focused on the relations between socioeconomic change and the structure of the landscape, but the methodology explained can be extrapolated to any other problem into the SES modeling framework. Also, continuous, hybrid and discrete BNs approaches are compared among them in terms or error rate and with a MLR methodology. The idea is just to overview that the BN-based solution is coherent, but not to provide an exhaustive comparison of the two approaches. Finally, two scenarios of socioeconomic changes were evaluated.

### 4.2.1 Methodology

Figure 4.2 outlines the methodology followed in this chapter divided into five different steps: *i)* Data collection, *ii)* Model learning, *iii)* Model validation, *iv)* Relations between features and response variable and *v)* Scenarios of socioeconomic change. Elvira software was used in the learning and validation processes, and for the evidence propagation during the study of the scenarios of change and the nature of the relationships between each feature and the response variable.

#### Study area

The study area is located in southeastern Spain, straddling parts of Almería and Granada provinces (Figure 4.3).

It covers around 500,000 Ha and lies in the Baetic System foothills with an irregular relief from high mountains peaks (more than 2,000 meters a.s.l.) to the sea level. This

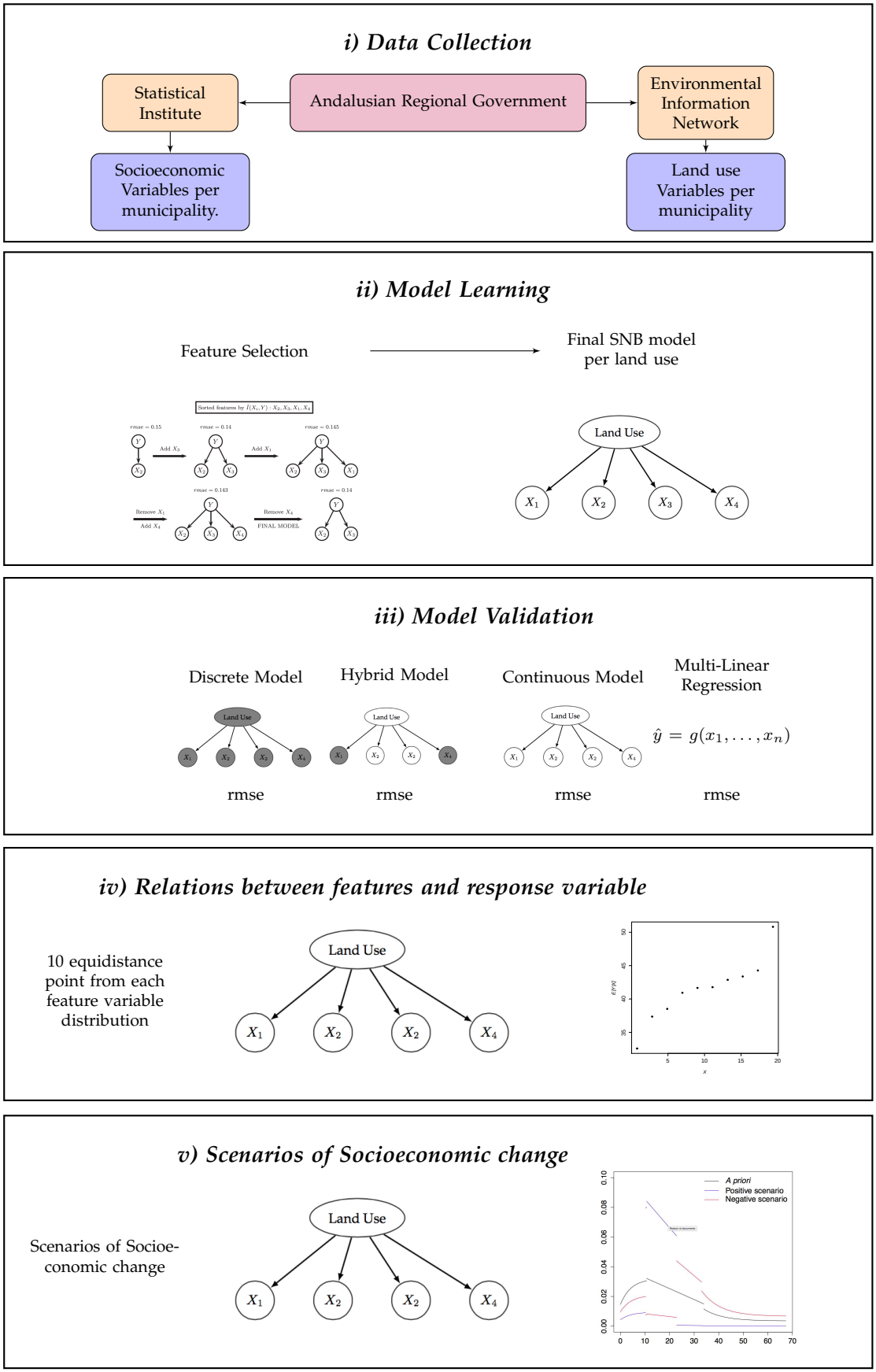



FIGURE 4.2: Outline of the methodology.

relief has determine a spatially and temporally irregular rain pattern. Spatially, rainfall ranges from 300 mm in the South, to 700 mm in the highland area, increasing to 850 mm in wet years. This rain patterns, added to the irregular relief, have configured a particular cultural landscape (García-Latorre & Sánchez-Picón, 2001).

The lowland part of the study area, named “Campo de Dalías" (marked in red color in Figure 4.3) has an extension of more than 18,000 ha covered by greenhouses. Intensive agriculture support an important economic activity that have an impact on the rest of the province, since both primary and secondary productive sector are, in certain way, related with it.

In contrast, the middle to high altitude landscape is configured as an heterogeneous territorial pattern. In Figure 4.3 different agricultural land uses are marked in purple and pink tones. They are mainly located around the river bed through all the study area. This agricultural landscape is determined by a picture a patchwork of olive and almond groves, grapevines, subsistence croplands, mix with scrubs and small patches of woodlands.

Natural landscape is mainly defined by the presence of scrubs both dense and disperse, pointed in dark and light green respectively. These land uses came from the abandonment of traditional agriculture areas, and from the historical deforestation process during the 19th century (Sánchez-Picón *et al.*, 2011) when most of forest patches of the study area were eliminated by the mining activities. Finally, in the upper areas and those with a difficult access due to the irregular relief, original woodland with oaks remains (marked in dark brown) and patches of reforested areas with conifers (in light brown).

From the socioeconomic point of view, the study area contains 90 municipalities, but only the most important ones are pointed in the map (Figure 4.3). Municipalities from “Campo de Dalías" area are quite densely populated with a high degree of migration. Socioeconomic activities are linked to primary (intensive agriculture with greenhouses), secondary (adjacent industry that manufactures the product) and tertiary sector, which is related to the development of intensive agriculture (e.g. large numbers of banks and shopping centers).

By contrast, there is a tendency of becoming less populous and more prone to depopulation through emigration as we are climbing in altitude. In that area, there is less primary sector activity and, in some cases, rural tourism is becoming and emergent economic activity, more pronounced than in the lowland area.

**Data collection**

Table 4.1 shows the selected socioeconomic variables which are representative of the socioeconomic structure of the territory (Schmitz *et al.*, 2005; Aranzabal *et al.*, 2008). Data were obtained per municipality in 2007 from the Andalusian Multiterritorial Information System [2]. Some variables needs to be defined:

Using the landscape typologies described by Schmitz *et al.*, 2005, and taking into account the characterisitic of the study area (described above) three types of landscape were selected: scrubland (dense and sparse scrubland), agricultural Mediterranean

[2] http://www.juntadeandalucia.es/institutodeestadisticaycartografia/sima/index2-en.htm

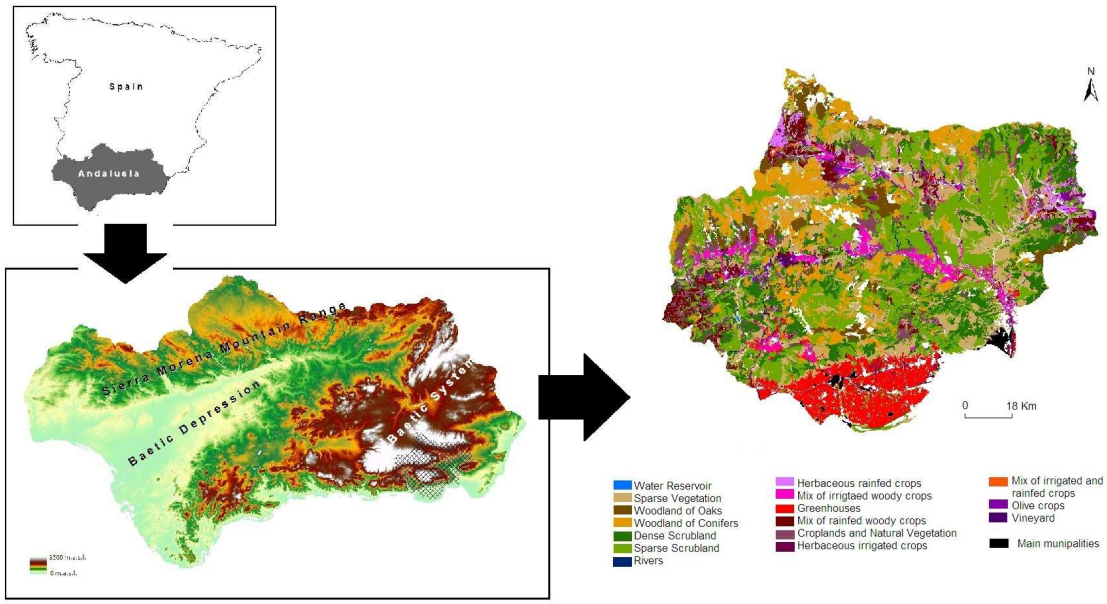



FIGURE 4.3: Study area.

TABLE 4.1: Socioeconomic variables express per municipality.

| **Socioeconomic variables** | **Unit** |
|---|---|
| Total population | No. people |
| Ageing * | % of people |
| Natural increase** | Value of Natural increase |
| Male index *** | No. males / No. females |
| Primary sector<br>Secondary sector<br>Tertiary sector | No. employees |
| Unemployed | No. unemployees |
| National emigration****<br>Foreign emigration****<br>National immigration ****<br>Foreign immigration**** | % of people |
| Illiterate<br>Primary studies<br>Secondary studies<br>Higher studies | % of people |
| * Percentage of the population older than 65<br><br>** The difference between the number of births and deaths | *** It is included since in the last decades in Spain, rural areas presented more male population than female (Camarero *et al.*, 2009).<br>**** National refers to people who emigrate/immigrate to/from other places in Spain, while foreign refers to emigrants/immigrants to/from other countries |

landscape (heterogeneous traditional croplands with olive trees and grapevine), and native forest (oak trees). Corresponding landscape data (percentages per municipalities) were obtained from the Andalusian Regional Environmental Information Network using the Land Use and Land Cover shape file 2007 using ArcGis v.9.3.1 (ESRI, 2006).

The final matrix has a total of 19 variables (16 socioeconomic and 3 land uses variables) over 90 observations (one per municipality).

**Model learning**

A constrained NB structure is selected with a priori feature selection. Three models were learnt, one per each landscape variable: Agricultural Mediterranean landscape (AML), Scrubland and Native Forest. The problem of selecting the features to be included in the MTE model was addressed by Morales *et al.,* 2007 following a *filter-wrapper* approach. The accuracy of the model is measured using the root mean squared error (rmse) (Witten & Frank, 2005) between the actual values of the response variable, $y_1, \ldots, y_n$, and those predicted by the model, $\hat{y}_1, \ldots, \hat{y}_n$, for the records in a test database (the original dataset is randomly divided into two sets, one for learning the model, and the other for testing it). Thus, the rmse is obtained as

$$rmse = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}. \tag{4.4}$$

where *n* is the sample size, $y_i$ is the real value whilst $\hat{y}_i$ is the predicted value.

The mutual information between two random variables $X$ and $Y$ is defined as

$$I(X,Y) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f_{XY}(x,y)\log_2\frac{f_{XY}(x,y)}{f_X(x)f_Y(y)}dxdy, \tag{4.5}$$

where $f_{XY}$ is the joint density for $X$ and $Y$, $f_X$ is the marginal density for $X$ and $f_Y$ is the marginal for $Y$.

The mutual information has been successfully applied as a filter measure in classification problems with continuous features (Pérez *et al.*, 2006). For MTEs, the computation of Equation (4.5) cannot be obtained in closed form. We will therefore use the estimation procedure proposed by Morales *et al.*, 2007, which is based on the estimator

$$\hat{I}(X,Y) = \frac{1}{m}\sum_{i=1}^{m}\left(\log_2 f_{X|Y}(X_i \mid Y_i) - \log_2 f_X(X_i)\right), \tag{4.6}$$

for a sample of size $m$, $(X_1, Y_1), \ldots, (X_m, Y_m)$, drawn from $f_{XY}$.

The steps for NB construction using a filter-wrapper feature selection is graphically shown with an example in Figure 4.4. The main idea is to start with a model containing the class variable and one feature variable, which is the node with the highest mutual information with respect to the response variable ($Y$ and $X_2$). Afterwards, the remaining variables are included in the model in sequence, according to their mutual

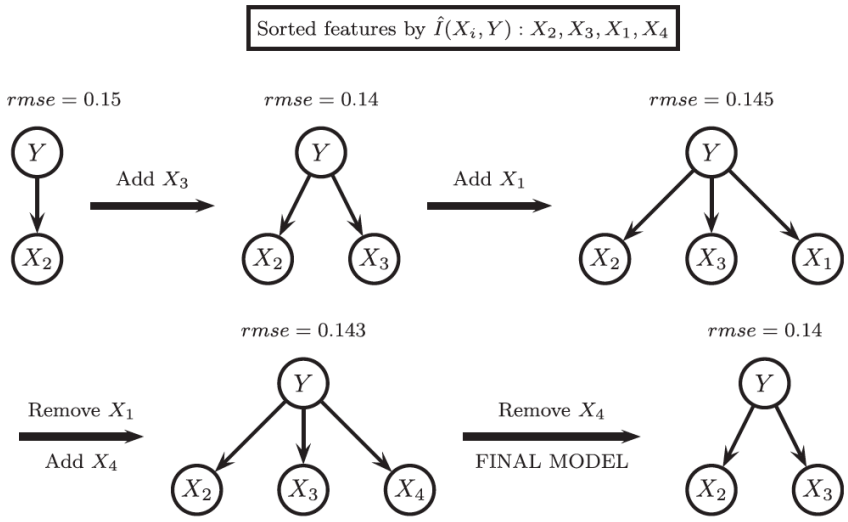



FIGURE 4.4: Example of feature selection in a NB regression model. First, features are sorted in a decreasing order using its mutual information with respect to $Y$. Then, their inclusion is checked step by step. Note that only the inclusion of $X_2$ and $X_3$ reduces the error. Finally, the procedure selects two out of four variables to be part of the final model.

information with respect to $Y$. In each step, if the included variable reduces the error defined in Equation (4.4), it is kept. Otherwise, it is discarded.

However, performing feature selection is influenced by two issues. Firstly, mutual information cannot be analytically computed, but it must be estimated from a simulated sample instead. If this sample size small, the selected features can vary between different executions (Fernández *et al.*, 2007). Secondly, the scarcity of data (only 90 instances) implies that the selected features strongly depends on the random test selected from the original dataset.

To solve both problems, the methodology was run twenty times and the variables appearing at least $75\%$ of the time were chosen. Accordingly, three continuous regression models were learned, one for each landscape (Figures 4.5, 4.7 and 4.9).

In order to compare the performance of the continuous model (presented above), against other alternatives, a hybrid and a discrete model were learned with the same set of variables selected for the continuous case. In this way, the comparison is more reliable as the model structure remains fixed, and it makes more sense from an environmental point of view.

In the hybrid approach, half of the variables were discretised (see Table 4.2). Note that, he CG model cannot be applied in this situation, since there are some discrete feature variables with a continuous parent (the response variable). Several discretisation methods (equal frequency, equal width and $k$-means) were tested to obtaining the hybrid and the fully discrete model (including the response variable). Finally, the $k$-means algorithm with three intervals was used as it reported the best results in terms of rmse.

It should be remembered that a fully discrete model is mainly oriented towards classification and not to regression. Consequently, in order to compare this model with

TABLE 4.2: Intervals of socioeconomic and land use variables included in the discrete model. * refers to those variables discretised in the hybrid model. $k$-means method is used to discretise the variables.

| **Socioeconomic variables** | **Intervals** |
|---|---|
| Total population* | [98, 9519) [9519, 47510) [ 47510, 186651] |
| Ageing* | [ 6.76, 19.13) [ 19.13, 29.77 ) [ 29.77, 48.44 ] |
| Natural increase* | [ -29, 46 ) [ 46, 528) [ 528, 982 ] |
| Male index | [ 0.49, 0.72 ) [ 0.72, 0.88 ) [ 0.88, 1.04] |
| Tertiary sector* | [ 0.0, 2095.5 ) [ 2095.5, 8041.5 ) [ 8041.5, 11819.0 ] |
| Unemployed* | [ 2.0, 975.5 ) [ 975.5, 7936.5 ) [ 7036.5, 12645.0 ] |
| National emigration* | [ 0.78, 5.38) [ 5.38, 9.61 ) [ 9.61, 19.35 ] |
| Foreign emigration | [ 0.0, 0.24) [ 0.24, 1.58 ) [ 1.58, 3.87 ] |
| National immigration | [ 0.0, 5.38 ) [ 5.38, 16.63 ) [ 16.63, 28.21 ] |
| Foreign immigration | [ 0.0, 1.18 ) [1.18, 3.14 ) [ 3.14, 6.70 ] |
| Primary studies | [ 4.03, 16.85 ) [ 16.85, 28.96 ) [ 28.96, 43.0 ] |
| Secondary studies* | [ 14.87, 25.04) [ 25.04, 32.21) [ 32.21, 45.07 ] |
| *AML* | [ 4.45, 23.39 ) [ 23.39, 44.49 ) [44.49, 80.93 ] |
| *Scrubland* | [ 0.0, 5.20 ) [ 5.20, 15.94 ) [ 15.94, 29.75 ] |
| *Native forest* | [ 0.0, 18.11 ) [ 18.11, 36.71 ) [ 36.71, 67.35 ] |

the hybrid and continuous cases, the rmse specified in Equation (4.4) needs to be recomputed for the discrete version as:

$$rmse = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - ca(\hat{c}_i))^2}\,, \tag{4.7}$$

where $ca(\hat{c}_i)$ is the class average for the predicted category after propagating the records in the discrete case, and $y_i$ is the actual continuous value for the response variable. Note that, once the data are discretised, the original continuous values are still necessary to compute this version of the rmse.

Direct (marked in red color) and inverse (marked in green color) relationships between each feature and the response variable were analysed. Two variables, $X$ and $Y$, are considered to have a direct relationship if an increase (or decrease) in the value of $X$ implies an increase (or decrease) in the expected value of the posterior distribution of $Y$. In contrast, an inverse relationship means that when the value of $X$ increases (or decreases), the expected value of $Y$ decreases (or increases). In order to check the sign of the relationships, for each feature, 10 equidistant values from its domain (including the minimum and maximum) were used as evidences for carrying out different propagations on the model. Thus, 10 expected values (means) of each posterior distribution for the response variable gave us information about the type of relationship (direct or inverse).

**Validation of the model**

The model was tested using $k$-fold cross-validation (Stone, 1974). It is a widely used technique in AI to validate models. The aim is to check how predictive a model is

when confronted with data that have not been previously used for learning the model. It is based on the holdout method in which the data set is separated into two complementary sets, one for learning ($D_l$) and another for testing ($D_t$). In this way, we can estimate the error of a model built from $D_l$ according to set $D_t$, using the formula in Equation (4.4).

To reduce variability, the data set is initially divided into $k$ subsets, and the holdout method is repeated $k$ times. Each time, one of the $k$ subsets is used as $D_t$ and the other $k-1$ subsets are put together to form $D_l$. Then the average error across all $k$ trials is computed. For the case study presented in this chapter, we set the value of $k$ to 10.

Finally, the validation was conducted by comparing the BN-based proposals (continuous, hybrid, and discrete models), with a MLR implemented in `R` software (R Development Core Team, 2012), since it is the most common regression solution used in environmental sciences.

The MLR model can also be applied in the presence of categorical variables, usually by transforming them into dummy variables. In particular, each categorical variable with $k$ states has to be converted into $k-1$ binary variables, one for each category of the variable. However, the interpretation of the regression coefficients for the categorical variables is different from the continuous ones. Another disadvantage of this hybrid MLR approach is that the manual construction of dummy variables can be laborious and even error prone, especially in the case of many categories. On the other hand, Bayesian networks naturally include categorical and continuous variables in the same model using the MTE distributions without the need of creating new variables.

#### Scenarios of socioeconomic change definition

Two scenarios of socioeconomic change that represent general tendencies in the socioeconomic structure (Aranzabal *et al.*, 2008; Schmitz *et al.*, 2005) were proposed (Table 4.3). The first scenario shows a positive socioeconomic development which involves an increase in the variables related to population, migration movement, study level (mainly in secondary and higher studies), and primary and tertiary economic sector. The second scenario shows a negative socioeconomic change. It involves a decrease in study level and primary and tertiary sector while an increase in emigration rates and unemployment.

As each regression model has a different subgroup of socioeconomic variables selected during the pre-processing step, the evidence is only introduced in those variables included in the corresponding model.

### 4.2.2 Results and Discussion

#### Model validation results

Table 4.4 shows the results in terms of the rmse when comparing the four approaches (continuous, hybrid, discrete BNs and MLR) for the three variables of interest (*AML*, *Scrubland* and *Native Forest*). As expected, the errors for the proposed continuous method are smaller than the other approaches. It is well-known that discretising data implies loss of information as demonstrated in the errors obtained for the hybrid and

TABLE 4.3: Scenarios of socioeconomic change. Minimum and maximum values refer to the minimum and maximum value found in the data set. Percentage changes are taken from Schmitz *et al.*, 2005 and Aranzabal *et al.*, 2008.

| **Scenario** | **Variables involved** | % **Change** |
|---|---|---|
| Positive socioeconomic change | Foreign immigration | Maximum value |
| | National emigration | +50% |
| | Tertiary sector | +60% |
| | Primary sector | +80% |
| | Higher studies | +15% |
| | Secondary studies | +30% |
| | Natural increase | +70% |
| | Ageing | Minimum value |
| Negative socioeconomic change | National emigration | Maximum value |
| | Higher studies | -70% |
| | Natural increase | Minimum value |
| | Primary sector | -20% |
| | Tertiary sector | -80% |
| | Total population | -50% |
| | Ageing | +80% |
| | Secondary studies | -40% |
| | Unemployment | Maximum value |

TABLE 4.4: Root mean squared error for the four BN-based regression models and the MLR. 10 fold-cross-validation is used to reduce variability.

| **Model** | ***Native forest*** | ***AML*** | ***Scrubland*** |
|---|---|---|---|
| Continuous BN | 6.47 | 14.73 | 18.60 |
| Hybrid BN | 6.74 | 14.89 | 19.13 |
| Discrete BN | 6.98 | 16.07 | 26.72 |
| MLR | 8.81 | 19.92 | 29.47 |

discrete approaches. Finally, MLR obtains a significantly larger error than the BN-based approaches.

In any case, the goal of this chapter is not only to compare the models above, but to present different ways to solve a regression problem in environmental modeling that carry fewer limitations, and which depend on the nature of the available data (continuous, hybrid and discrete). For a detailed comparison of BN-based regression models see Morales *et al.*, 2007.

#### Scenario results

The results of the comparison among the three BNs approaches suggests the continuous approach to be the most appropriate one for modeling this problem as it has the lowest rmse. For this reason, only results from the continuous models are presented here.

Results are presented according to three different settings: *a priori* and *a posteriori* (two scenarios). Figures 4.5, 4.7, 4.9 show the qualitative part of the BNs developed and the direct and inverse relationships with the selected socioeconomic variables. Figures 4.6, 4.8 and 4.10 shows the probability distributions *a priori* without introducing any scenario (black line), and the posterior probability distributions (blue and red line) of the variables after introducing the two socioeconomic scenarios according to Table 4.3. Finally, Table 4.5 shows some statistics for each variable in the three different settings specified above.

**Agricultural mediterranean landscape**

*A priori*, *AML* (Figure 4.5) is related to a socioeconomic structure characterized by a sparse total population with a low male index, but a positive natural increase. Educative level is medium with a high unemployment rate. National immigration is low. Variable *F*oreign immigration has a peculiar behaviour, since its middle and low values are related to agriculture workforce (mainly coming from northern Africa), and have a direct relationship with *AML*. On the other hand, its high values are more related to retired population coming mostly from northern Europe, who have a second home in the area, and this has a inverse relationship with *AML*.

A negative scenario means a decrease in total population, natural increase and study level variables (as specified in Table 4.3). In Figure 4.6, red line shows the probability distribution function of the *AML* under this negative scenario. From a value of 20-25 the probability distribution is higher than *a priori* distribution. It means that, due to the rural abandonment in which elder and non-qualified population remains in the area and keep these heterogeneous agricultural systems in a large extension (*AML* is expressed as the surface in the municipality). Also, it involves an increase in the mean of the posterior probability distribution of *AML* (Table 4.5).

On the other hand, a positive scenario supposes larger values in natural increase, foreign immigration and study level (as specified in Table 4.3). In Figure 4.6 it is represented by the blue line. In general, under this positive scenario the mean of the posterior probability distribution of *AML* decrease, but the probability distribution shows that, even when it is under the *a priori* distribution, in the range between 10 to 22, the distribution of this scenario is higher. Even when the economic development implies an interest in other economic sectors rather than the agriculture, some small patches are still remain. However, there is not enough to maintain the traditional cultural landscape.

**Scrubland**

*A priori*, *Scrubland* (Figure 4.7) is related to a socioeconomic structure characterized by an ageing population with low study levels. Unemployment is considerable and national emigration prevails over international. In this context, tertiary sector is the main economic activity.

A negative scenario means an increase in ageing, national emigration and unemployment; and a decrease in secondary studies and tertiary sector (Table 4.3). This situation provokes a rural abandonment which brings the increase of the scrubland surface. As Figure 4.8 shows, the probability distribution under this negative scenario (red line) is

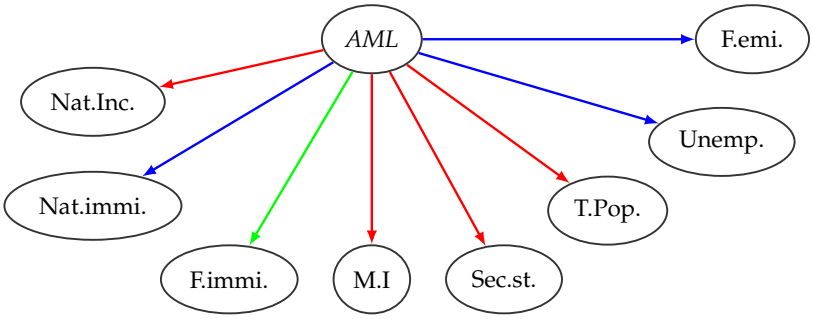



FIGURE 4.5: Regression model for variable *Agricultural Mediterranean Landscape* (*AML*) after the feature selection. Direct and inverse relationships between each feature and the variable *AML* are labelled with red and blue color, respectively, on the corresponding arc. *F*oreign immigration is labelled with a green color as it has a peculiar behaviour, low and middle values in its domain present a direct relationship with *AML*, however high values have a inverse relationship. M.I., Male Index; Sec.st., Secondary Studies; T.Pop., Total Population; Unemp., Unemployment; F.emi.; Foreing Emigration; F.immi, Foreign Immigration; Nat.immi. National Immigration; Nat.Inc., Natural Increase.

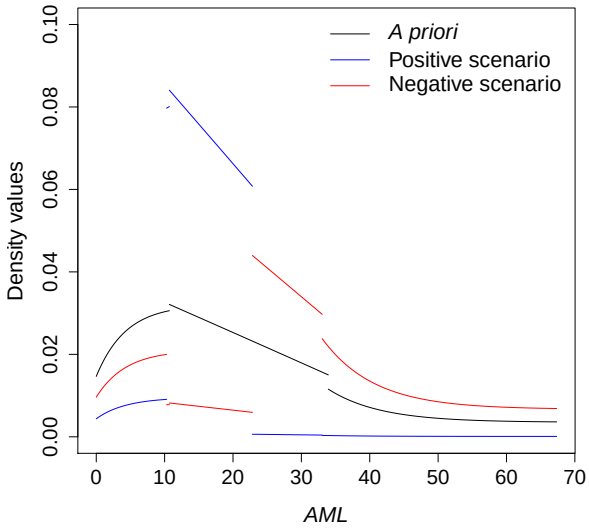



FIGURE 4.6: Probability distributions of the AML. Three density functions are displayed: *a priori* (without introducing any scenario to the model), and *a posteriori* (introducing both positive and negative scenarios according to Table 4.3). Note that density functions are defined as a piecewise function using MTEs.

higher than the *a priori* for the higher values of the scrubland variable. It means that municipalities landscape presents a high percentage of surface occupied by scrubs. It is also emphasized by the mean of the posterior probability distribution (Table 4.5).

On the other hand, a positive scenario entails a decrease in variable *A*geing to its minimum and an increase in secondary studies, national emigration and tertiary sector (as specified in Table 4.3). Population growth and a increase in secondary study levels cause a higher interest in other economic activities replacing scrubland with other land uses. It, therefore, involves a decrease in the mean of the posterior probability distribution of *Scrubland*. It is also shown in the probability distribution (Figure 4.8) in which the probability of lowest values os *Scrubland* are higher than *a priori*.

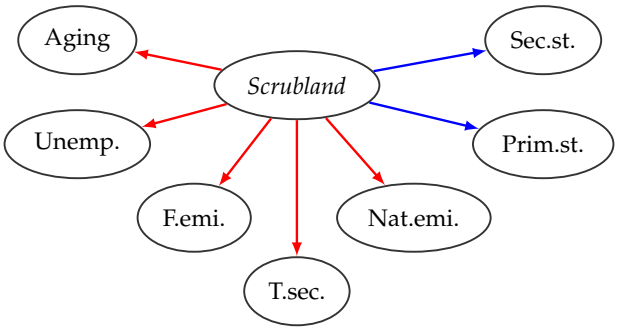



FIGURE 4.7: Regression model for variable *Scrubland* after the feature selection. Red and blue color have the same meaning as in Figure 4.5. T.sec., Tertiary Sector; Nat.emi., National Emigration; Prim.st., Primary Studies; Sec.st., Secondary Studies; F.emi., Foreing Emigration; Unemp., Unemployment.

**Native forest**

*A priori*, *Native forest* (with oak trees) (Figure 4.9) is related to a socioeconomic structure characterized by low population with primary studies, but a positive value in natural increase variable. Moreover, the tertiary sector is well-developed and there is a national migration (immigration and emigration).

A negative scenario entails a drop in the value of natural increase variable, total population and tertiary sector (rural tourism), whilst there is an increase in national emigration. Rural abandonment and low tourism levels entail less interest in maintaining native forest. It involves a slight decrease in the mean of the posterior probability distribution, whilst an increase in the probability of lower values of *Native forest* (Figure 4.10)

On the other hand, a positive scenario means an increase in the tertiary sector, national emigration and natural increase variables. Population growth and greater touristic activities lead to an improvement in infrastructure not only for tourism, but also for residents. It entails the replacement of native forest with land uses related to those improvements. It involves a decrease in the mean of the posterior probability distribution of *Native forest* and an increase in the lowest values of the probability distribution function (Figure 4.8).

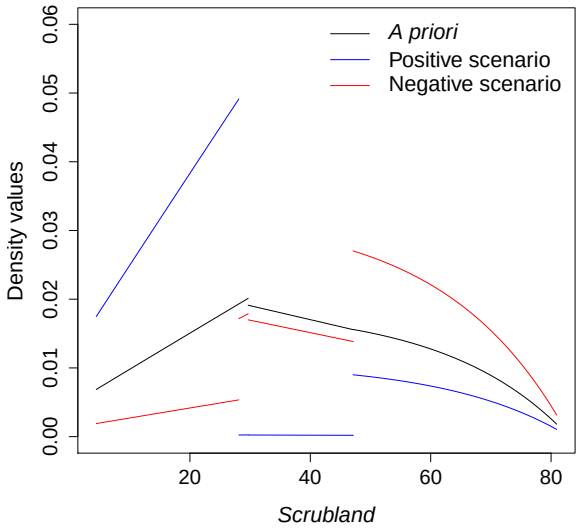



FIGURE 4.8: Probability distributions of the Scrubland. The same explication as in Figure 4.6.

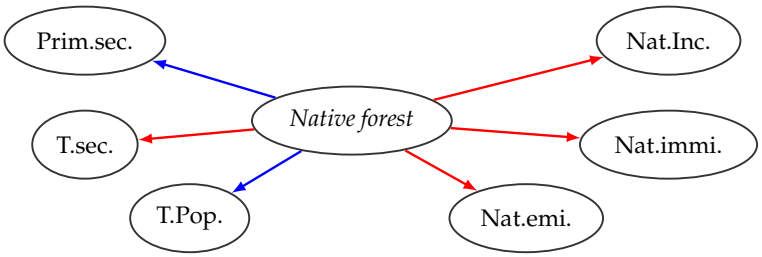



FIGURE 4.9: Regression model for variable *Native forest* after the feature selection. Red and blue color have the same meaning as in Figure 4.5. Nat.emi., National Emigration; Nat.immi., National Immigration; Nat.Inc., Natural Increase; T.Pop., Total Population; T.sec., Tertiary Sector; Prim.sec., Primary Sector.

TABLE 4.5: Mean and standard deviation values *a priori* and in each scenario of socioeconomic changes. *AML* refers to *Agricultural Mediterranean Landscape*. Sc., Scenario

| | ***A priori*** | | **Negative Sc.** | | **Positive Sc.** | |
|---|---|---|---|---|---|---|
| | Mean | Sd | Mean | Sd | Mean | Sd |
| *AML* | 21.50 | 15.16 | 29.65 | 16.18 | 15.62 | 5.17 |
| *Native forest* | 7.66 | 6.71 | 6.32 | 5.79 | 3.81 | 3.48 |
| *Scrubland* | 39.62 | 18.73 | 49.98 | 16.22 | 26.96 | 18.52 |

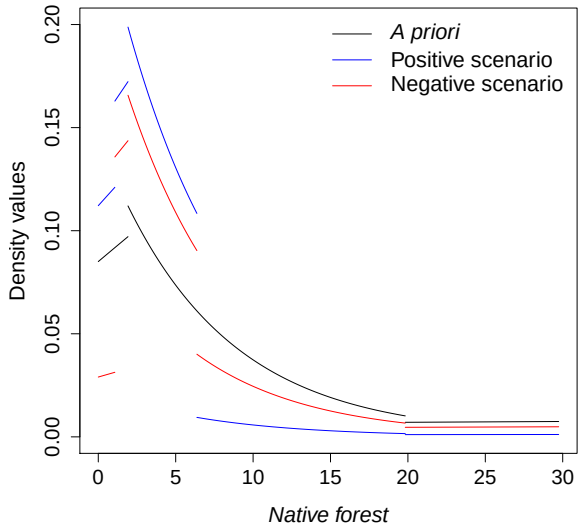



FIGURE 4.10: Probability distributions of the Native Forest. The same explication as in Figure 4.6.

## 4.3 Conclusion

This chapter presents MTE-based BNs as a tool for solving regression problems in SES modeling, using the modeling of landscape - socioeconomy relationship in southern Spain as our study case. Two gradual socioeconomic changes impact over the three main landscapes are studied. A global understanding of the change in the SES can be obtained.

Both socioeconomic changes represent a general tendency of mediterranean landscape to a richer and improved socioeconomy structure, and to a rural abandonment process. In the case of the positive scenario, defined as a development in the socioeconomic structure, traditional landscape (*AML*) is reduced. This tendency would imply an increase in the scrub surface in order to substitute these areas, however, both *Scrubland* and *Native forest* are also reduced in a great extension. As some previous studies reveals (Schmitz *et al.*, 2005), this gradual change to a richer socioeconomic structure involves the promotion of other landscapes typologies more related with intensive agriculture or tourist interests. In this case, the SES tends to change into a more intensive and homogeneous system with less presence of traditional cultural landscapes and native forest.

On the other hand, a negative scenario describes a similar situation to a rural abandonment. In such conditions, both *AML* and *Scrubland* tend to increase, while *Native forest* undergoes a slight reduction. Agricultural Mediterranean landscape, as a heterogeneous landscape, only in rural areas is kept, where elderly people cultivate small patches of traditional croplands (Schmitz *et al.*, 2003). A lower level of education and fewer job opportunities mean a restriction in the number of economic activities, so that several patches are abandoned promoting the increase of *Scrubland* (Geri *et al.*, 2010; Camarero *et al.*, 2009). In that situation, traditional activities related to the maintenance

of native forest are somehow forgotten, so the surface area of native forest is slightly reduced (Jiménez-Herrero *et al.,* 2011). Under this second scenario, it is demonstrated the impact and the close relationships between social and natural systems, where an absence of population that keeps traditional agriculture and forest maintenance, provoke that natural heterogeneous system (measure as the number of patches from different land uses) changes to a landscape more homogeneous (land use change from a mixture of different patches to a bigger patch of scrubs); scrublands tend to substitute both native forest and agricultural systems.

# Chapter 5

# Classification through Bayesian networks: Socio-Ecological Cartography

Territorial planning and management requires that the spatial structure of the socio-ecological sectors is adequately understood. In this chapter, a hierarchical hybrid BN classifier is applied to identify the different socio-ecological sectors in Andalusia, a region in southern Spain. Besides, a Global Environmental Change scenario is included into the model. Results show that a priori, the socio-ecological structure is highly heterogeneous, with an altitude gradient from the river basin to the mountain peaks. However, under a scenario of global environmental change this heterogeneity is lost, making the territory more vulnerable to any alteration or disturbance. The methodology applied allows dealing with complex problems, containing a large number of variables, by splitting them into several sub-problems that can be easily solved. In the case of territorial planning, each component of the territory is modeled independently before combining them into a general classifier model.

## 5.1 Introduction: Bayesian networks for classification

In the previous Chapter, the aim of the model was to predict the behavior of a continuos variable as accurate as possible solving a *regression* problem. But, if this variable is discrete it is called *class* variable and we are facing with a *Classification*. In environmental modeling is common that no information about this class variable is given, so the problem becomes an *unsupervised* classification or clustering problem.

**Clustering or unsupervised classification:** *It is understood as a partition of a data set into groups in such a way that the individuals in one group are similar to each other but as different as possible from the individuals in other groups*.

Hybrid BNs based on MTE models have been successfully applied in *supervised classification* problems in environmental sciences (Aguilera *et al.,* 2010), for that reason this Chapter is focused on *unsupervised classification* problems instead.

BNs can be used to solve both *supervised* and *unsupervised classification* tasks (Aguilera *et al.*, 2013; Anderberg, 1973; Fernández *et al.*, 2014; Gieder *et al.*, 2014) if they contain a set of feature variables $X_1, \ldots, X_n$, and a class variable (in the case of supervised classification), where an individual with observed features $x_1$, ..., $x_n$ will be classified as belonging to a class c. Unsupervised classification is performed taking into account that no information about class variable C is given. Therefore, a hidden variable H whose values are initially missing is included in the dataset to represent the membership of each case to the different clusters.

As BNs express the results by means of probability distribution functions, each observation in the dataset is classified according to the class label with the highest probability value. In that case, the interest is on what state of the class variable is the most probable one, rather than the probability of it. Again, the behavior of the system can be modeled under a scenario of change using probabilistic propagation (Aguilera *et al.*, 2011; Liedloff & Smith, 2010).

In this Chapter the methodology applied is based on the probabilistic clustering methodology using hBNs proposed by Fernández *et al.*, 2014, and implemented in the Elvira software (Elvira-Consortium, 2002). Figure 5.1 shows an outline of this methodology which is divided into two steps:

1. *Estimation of the optimal number of states*. Initially, no information about the class variable is given, so we consider it as a hidden variable $H$, whose values are missing (Figure 5.1 *i)*). Firstly, we consider only two states for variable $H$ that are uniformly distributed (Figure 5.1 *ii)*). Now, the model is estimated based on the *data augmentation* method (Tanner & Wong, 1987), an iterative procedure similar to the Expectation Maximization algorithm (Lauritzen, 1995) as follows: *a)* the values of $H$ are simulated for each data sample according to the probability distribution of $H$, updated specifically for the corresponding data sample, and *b)* the parameters of the probability distribution are re-estimated according to the new simulated data. In each iteration, the BIC score of the model is computed, and the process is repeated until there is no improvement. In this way, the optimal parameters of the probability distribution function of the model with two states and its likelihood value are obtained (Figure 5.1 *iii)*). The following step consists of a new iterative process in which a new state is included in variable $H$ by splitting one of the existing states (Figure 5.1 *iv)*). The model is again re-estimated (by repeating the *data augmentation* method) and the BIC score is compared with the previous run. The process is repeated until there is no improvement in the BIC score, so achieving the final model containing the optimal number of states (Figure 5.1 *v)*).

2. *Computation of the probability of each observation belonging to each state*. Once we have obtained the final model (with the optimal number of class variable states), the next step consists of *inference* process. In this step, all the available information for each data sample is introduced into the model as an *evidence*, and propagated through the network, updating the probability distribution of the class variable. Finally, from this new distribution the most probable state of the variable $H$ for each data sample is achieved.

BNs for *classification* is the second most applied aim in environmental modeling (Figure 2.7) with more than 20% of the paper reviewed. As it was pointed out in Figure 1

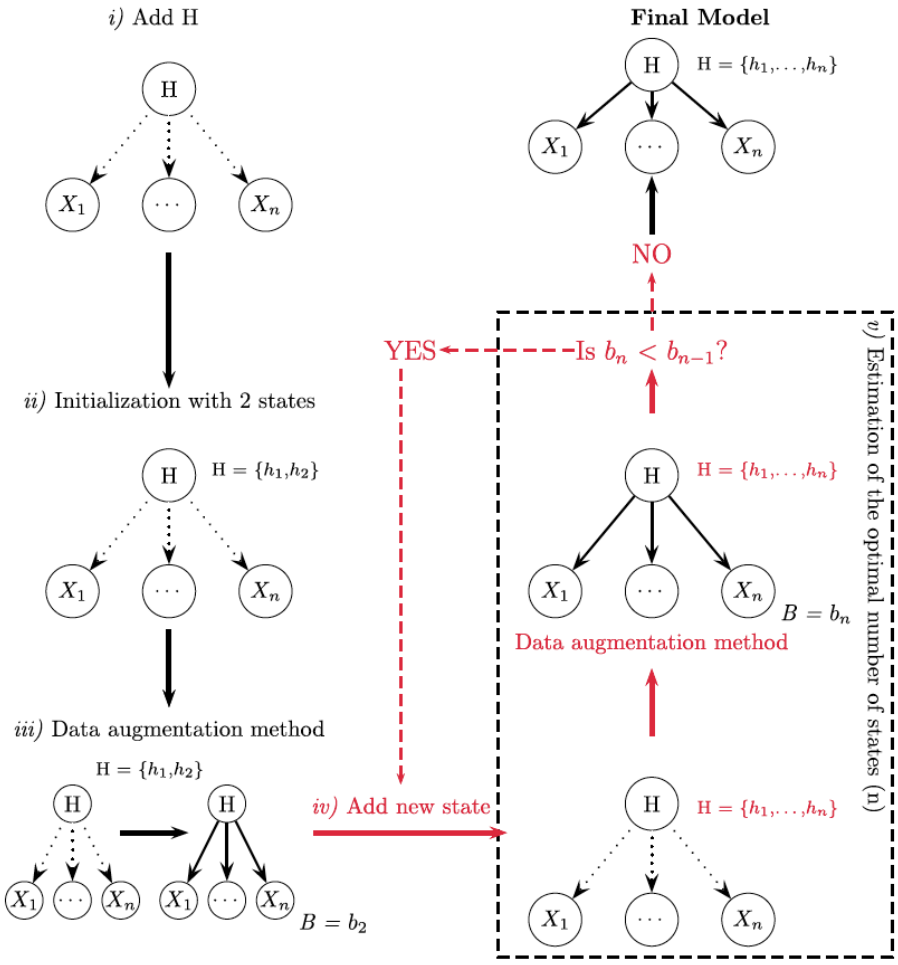



FIGURE 5.1: Outline of the HBNs probabilistic clustering methodology to construct both sub-models and the classifier. Dotted lines represent the relationships between the variables when the parameters of the probability distribution functions have not been yet estimated. B, BIC score.

BNs can be learnt using an optimal or fixed structures. For both *regression* and *classification* problems, fixed and constrained structures are the most usually applied and more recommended. However, in environmental and ecological modeling just a 12.5% of the paper reviewer that applied BNs for *Classification* used them (Aguilera *et al.,* 2010; Bressan *et al.,* 2009; Park & Stenstrom, 2008; Park & Stenstrom, 2006; Porwal *et al.,* 2006), whilst the remainder 87.5% learns an optimal structure even when their focus is on only one discrete variable (Grech & Coles, 2010; Walton & Meidinger, 2006; Stow *et al.,* 2003; Raphael *et al.,* 2001).

With respect to the data, mostly of the paper discretized the features and deal with a discrete classification model (Figure 5.2), following the general tendency in environmental and ecological modeling (Boets *et al.,* 2015; Fletcher *et al.,* 2014). What is interesting is that model learning based on the data (automatic leaning) is found in the 35% of the papers (Keshavarz & Karami, 2013; Palmsten *et al.,* 2013), the same percentage of papers used both data and experts (semi-automatic approach) (Figure 5.3) (Boets *et al.,* 2015). By contrast, the percentage of papers that use only experts is 22% (Fletcher *et al.,* 2014).

Figure 5.4 shows the distribution of papers in each research topic. *Ecology* and areas that are not so representative (Figure 2.7), *Biodiversity & Conservation*, *Marine & Freshwater Biology* and *Forestry*; are the most usually applied areas of BNs for *classification*.

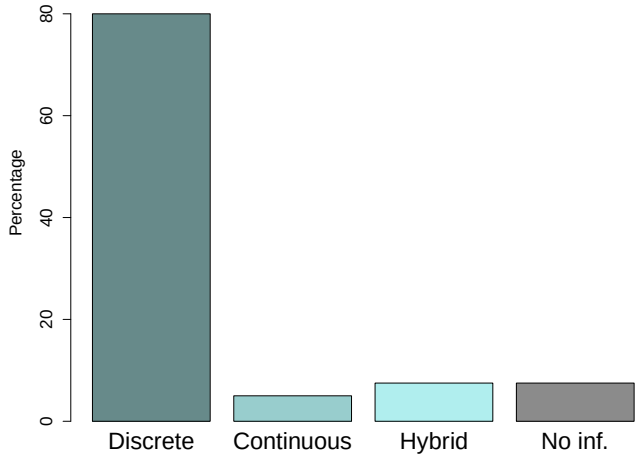



FIGURE 5.2: Percentage of papers that applied BNs for *Classification* for each type of data.

## 5.2 Analysis of the socio-ecological structure and dynamics of the territory

Under the SES framework, the process of planning and management requires that the spatial structure of the territory is adequately understood, particularly given the current context of Global Environmental Change (GEC) (Basurto *et al.,* 2013; Clark & Dickson, 2003; Hufnagl-Eichiner *et al.,* 2011; Kotova *et al.,* 2000; Turner *et al.,* 2003). Spatial

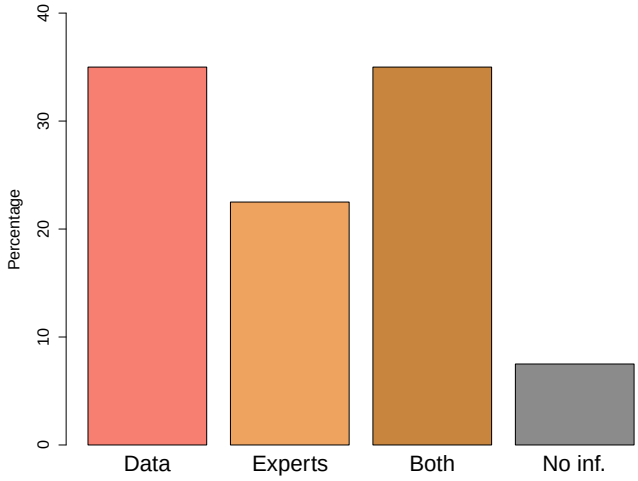



FIGURE 5.3: Percentage of papers that applied BNs for *Classification* for each learning process

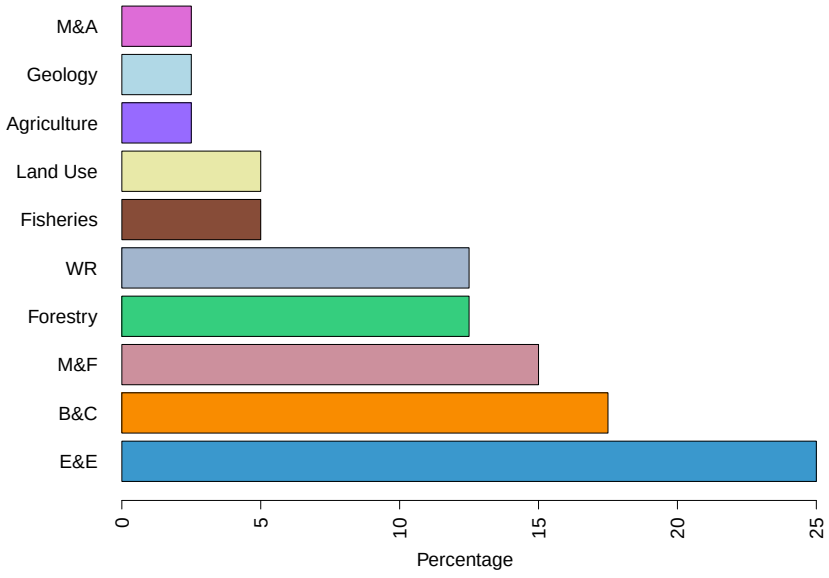



FIGURE 5.4: Percentage of papers that applied BNs for *Classification* for each topic. M&A, Meteorological and Atmospheric sciences; M&F, Marine and Freshwater Biology; E&E, Ecology and Environmental sciences; B&C, Biodiversity and Conservation; WR, Water research

analysis allows the territory to be divided into a number of different units (Schmitz *et al.*, 2005), which can reflect the spatial patterns caused by interactions between social and ecological systems and between the elements of the territory.

Despite that human's role in nature is being recognized, most of papers are focused on determining the ecological sectors and their interactions excluding the social component (Jackson *et al.*, 2012). To obtain these ecological sectors, a variety of methodologies have been applied including both subjective methods - based on expert knowledge- and objective ones, based on the data available (Chuman & Romportl, 2010; Schmitz *et al.*, 2005; Trincsi *et al.*, 2014; Vezeanu *et al.*, 2010). One of the most important methodologies is classification, with recent advances promoted by the development of new technologies, such as GIS techniques and software. The most common classification methodologies are based on spatial overlapping of thematic maps and other GIS techniques (Villamagna *et al.*, 2014), the study of satellite images (Rapinel *et al.*, 2014) and various statistical methods, such as hard-clustering or geospatial analysis (Giménez-Casalduero *et al.*, 2011; Liu *et al.*, 2014; Ruiz-Labourdette *et al.*, 2011; Trincsi *et al.*, 2014; Vezeanu *et al.*, 2010) to perform data analysis and ecological mapping (Lahr & Kooistra, 2010).

Even though the methodologies mentioned above provide robust and appropriate results, they have certain limitations, which basically relate to the amount of information the models can cope with and the rigidity of the boundaries between the different sectors identified (Niederscheider *et al.*, 2014; Smith & Brennan, 2012). Moreover, new tools are required that can include socioeconomic components in the same way as other components of natural systems, under the current SES framework (Challies *et al.*, 2014; Dearing *et al.*, 2014; Strand, 2011).

In this Chapter, the objective is to develop a new methodological approach based on a hBN hierarchical classifier and apply it to characterize the socio-ecological structure of a territory, and study its dynamic under different drivers of GEC, in the Spanish region of Andalusia. This mathematical approach is considered hierarchical, since the model is divided into two levels of classification; in the first, both natural and socioeconomic components are modeled using independent hBN sub-models, with the aim of classifying the territory into several groups. In the second, the sub-models are joined into a classifier model that divides the territory into several socio-ecological sectors. Using the final classifier model obtained, we can predict how the socio-ecological structure of the territory might change as a consequence of various GEC drivers through the inference or probability propagation process.

### 5.2.1 Methodology

Figures 5.5 and 5.6 outline the methodology followed in this chapter. Firstly, Figure 5.5 shows the methodology for model learning divided in three main steps: *i)* Data collection, *ii)* Submodels learning, and *iii)* Classifier learning. Figure 5.6 shows the inference process followed to predict the behavior of the SES modeled under a Global Change scenario. Elvira software was used for both learning and inference processes.

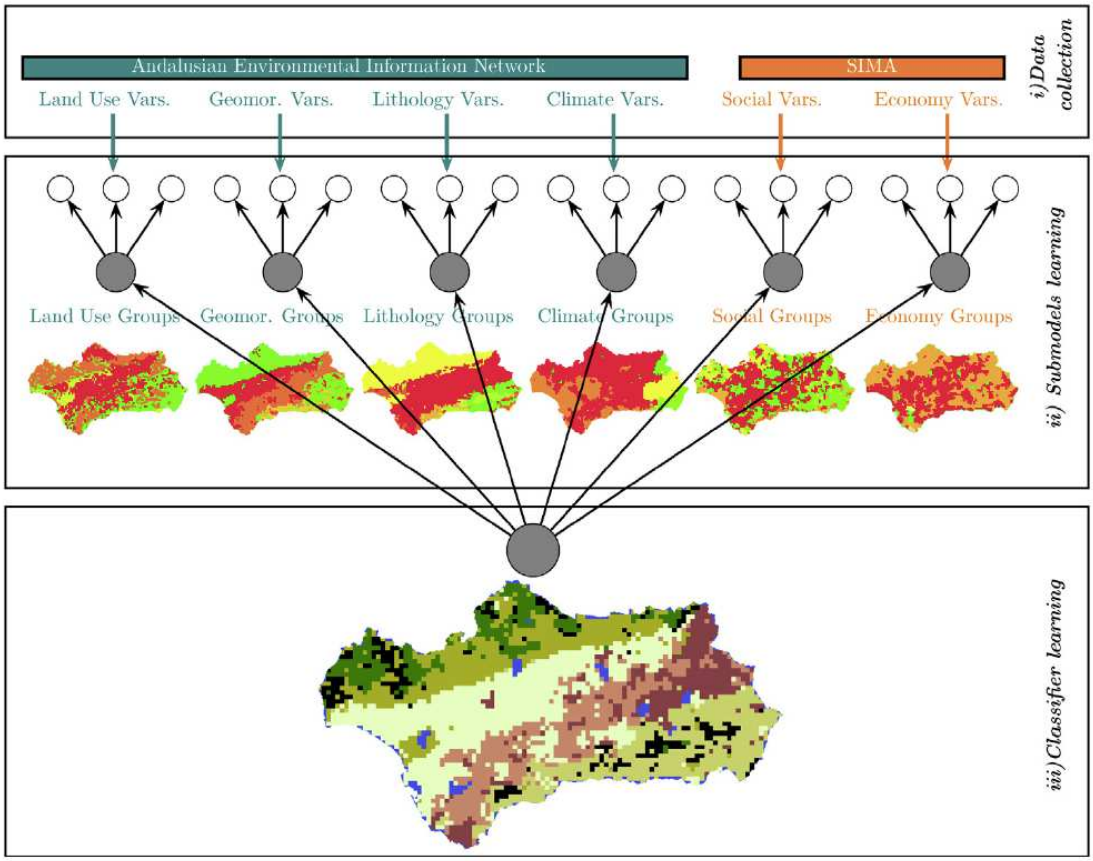



FIGURE 5.5: Methodological diagram of the hierarchical classifier model divided into three steps: i) Data collection, ii) Submodels learning and iii) Meta-classifier learning. White nodes refer to original variables (either discrete or continuous), grey nodes refer to artificial discrete class variables, which represent the membership of each observation to submodels groups (*i.e.* Land uses groups) and classifier sectors respectively. SIMA, Andalusian Multiterritorial Information System; Vars., Variables; Geomor., Geomorphology

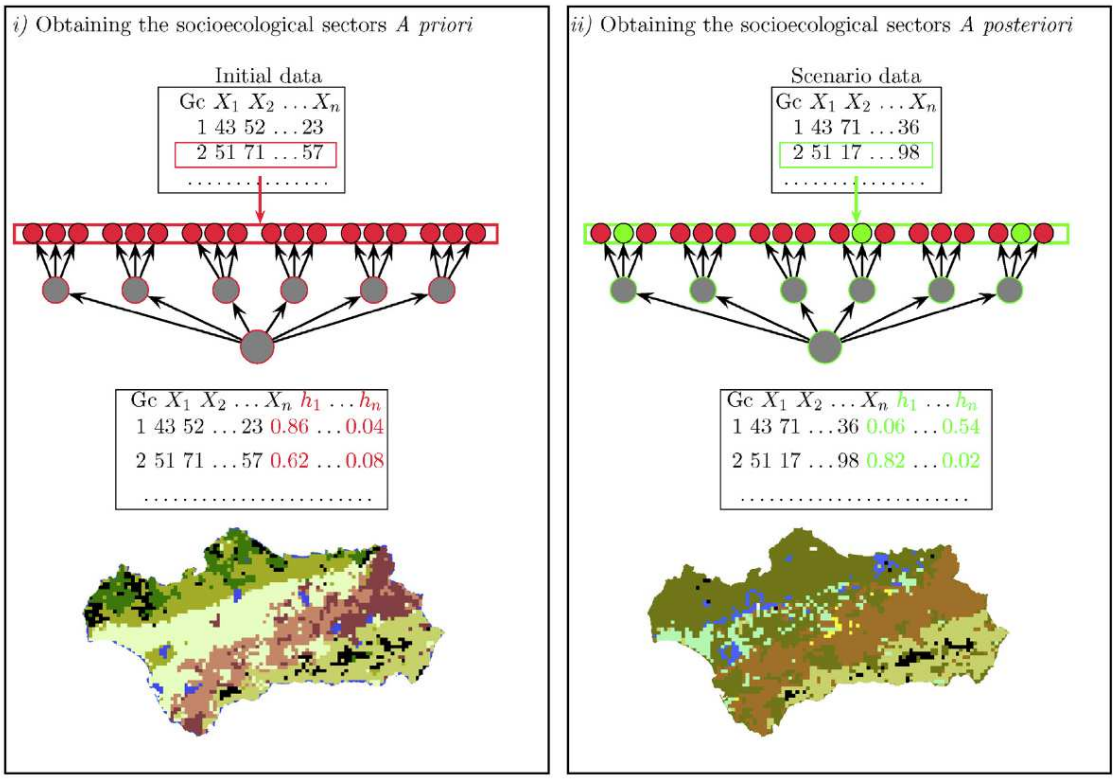



FIGURE 5.6: Methodological diagram of the Inference process. *A priori* the information about the current situation is introduced into the model and propagated to obtain the probability of each grid cell (Gc) belonging to socio-ecological sectors. *A posteriori*, information about drivers of GEC is collected and included - as new values or evidences - into several variables of the classifier model, and the probability values are updated.

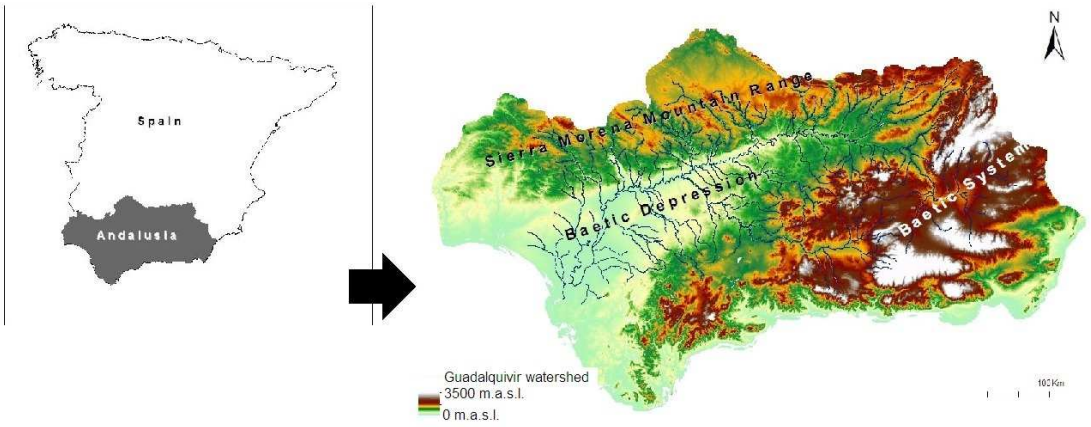



FIGURE 5.7: Study area.

### Study area

Andalusia (Figure 5.7) is located in the South of Spain and configures the second largest Autonomous Region, and the most-densely populated. It covers a surface area[1] of 87.600 km$^2$, which represents 17.3% of the national territory. Bounded by the Mediterranean Sea (at the East) and Atlantic Ocean (at the West), Andalusia lies on the frontier between Europe and Africa and contains a mixture of landscapes and cultural heritage from both continents.

Andalusian terrain covers a wide range of altitude, from the *Baetic Depression* to the mountainous ranges of the *Sierra Morena* and *Baetic System*, which boast the highest peaks in Spain, lying above 3000 m. a.s.l. The landscape is quite heterogeneous, with huge differences between the densely populated and irrigated rich croplands areas of the river basin and coastlands, to the sparsely populated forested areas of the uplands. Besides, in the last century it suffers a process of rural depopulation that provokes an increase in the scrub landscape in the higher relief.

Its climate is similarly heterogeneous. Even though Andalusia is included in the Mediterranean climate zone, there are stark differences between different areas. The climate in the southeast part is semiarid, with less than 200 mm of annual rainfall in several areas, whilst the middle and northern parts are under a continental climate influence, with more than 4000 mm rainfall.

All natural and social conditions make Andalusia a heterogeneous region with deep differences in terms of social structures, cultural heritage, and territorial structure.

### Data collection

In accordance with the environmental and socioeconomic characteristics of the territory, six groups of variables were selected for the hBN hierarchical classifier model.

Environmental information was collected from Andalusian Regional Environmental Information Network[2] (Figure 5.5 *i)*) and divided into four different sub-models: land

[1]Data from the Spanish Statistical Institute
[2]http://www.juntadeandalucia.es/medioambiente/site/rediam

use, geomorphology, lithology and climate. ArcGis v10.0 (ESRI, 2006) was used to retrieve the data, using a grid of 5x5 km. Land use, geomorphology and lithology variables are expressed as the percentage of the surface area of each grid cell, whilst climatic variables are expressed as an absolute value per grid cell (see Appendix A for a detailed explanation).

The Andalusian Multiterritorial Information System [3] was searched to obtain social and economic information for each municipality to feed to the corresponding sub-models (Figure 5.5 *i)*). In order to obtain information that related to uniform spatial units, ArcGis v10.0 (ESRI, 2006) was used to transform the data into a 5x5 km grid by overlapping it onto the municipal information shape file. In this way two cases were found: *i)* grid cells containing only one municipality, where the information was collected; *ii)* grid cells that overlap two or more municipalities; in these cases variables were obtained as a weighted mean of each municipal values. Variables are expressed in different ways, such as rates, percentage of the municipal population, percentage surface area of the territory (see Appendix A for a detailed explanation).

Variables were selected by experts and from literature review; they were preprocessed with the aim of avoiding repeated information. The preprocessing steps included the elimination of variables providing equivalent information by means of the analysis of a correlation matrix, and the selection of the appropriate level of detail in the shape file information in the ArcGIS. In addition, environmental variables comprising more than 70% of data equal to zero were discretized using the equal frequency method into three different states (0- no presence; 1- low presence; 2- high presence. Thresholds of each variable are shown in A).

The final data set contained 3630 grid cells and 151 variables, both discrete and continuous.

### Sub-model learning

This section describes the steps for constructing the first level of the classifier (Figure 5.5 *ii)*). Data collected were organized into six different groups Land use, Geomorphology, Lithology, Climate, Social, and Economy. Taking independently the variables for each group, six different sub-models were learnt following the method explained in the Introduction of this Chapter based on the proposal of Fernández *et al.*, 2014.

For each one, the structure used is a fixed NB in which the features are the variables collected, whilst the *class* variable is the hidden variable that represent the membership of each grid cell to a cluster with common characteristics (Land use, Geomorphology, Lithology, Climate, Social, and Economy group respectively). Table 5.1 shows the number of both discrete and continuos variables in each sub-model.

### Classifier learning

Once the various sub-models are learned, the next step consists of joining them in the second level of classification in the classifier model (Figure 5.5 *iii)*). A new virtual data set is created where the feature variables are the results of the previous six sub-models

[3] http://www.juntadeandalucia.es/institutodeestadisticaycartografia/sima/index2-en.htm

TABLE 5.1: Sub-models characteristics. No., number; Vars., variables.

| Sub-model | No. Vars. | Discrete Vars. | Continuous Vars. |
|---|---|---|---|
| Land Use | 10 | 0 | 10 |
| Geomorphology | 50 | 48 | 2 |
| Lithology | 41 | 39 | 2 |
| Climate | 7 | 0 | 7 |
| Social | 18 | 4 | 14 |
| Economy | 25 | 15 | 10 |
| **Total** | **151** | **106** | **45** |

(*i.e.*, the most probable land use, geomorphology, lithology, climate, social and economic cluster for each grid cell), whilst the hidden class variable expresses the membership of each grid cell to the socio-ecological sectors.

Note that, in this level, both feature and class variables are discrete, but the flexibility of the methodology proposed allows this kind of data to be dealt with in exactly the same way as in the previous step. The process is repeated, as explained in the Introduction and Figure 5.1, to obtain the final model with the optimal number of socio-ecological sectors. Once we know the parameters of the model, the *inference* process is carried out and the probability that a particular grid cell belongs to a particular sector is calculated; then the most probable one is represented.

#### Global environmental change scenario

Taking the information provided by the Intergovernmental Panel on Climate Change, both national and regional governments have developed climate change scenarios for their particular territory. A number of reports and studies have been written about the impact of these scenarios on the economy, on society, and on land use and land cover (Gasca, 2014; Méndez-Jiménez, 2012; Nieto & Linares, 2011). In Andalusia, two scenarios are considered: A2 and B2 (Méndez-Jiménez, 2012).

The A2 scenario describes a heterogeneous world, where self reliance and preservation of local identity are key. Population increases continuously and economic development is based on national decisions (regionally oriented), whilst per capita economic growth and technological change are fragmented and slow (Gasca, 2014; Solomon *et al.*, 2007). By contrast, the B2 scenario describes a situation in which economic development is not important and the environmental and socioeconomic problems are solved at local level. This scenario implies a slow population increase (Gasca, 2014; Solomon *et al.*, 2007).

In this application the focus is on the A2 scenario - the 2040 horizon scenario for Andalusia, since it closer to the current trend of socio-ecological change. Information to describe the impact of several GEC drivers on different sectors of the natural and social-economic environments in Andalusia were collected from various sources: the Assessment of the International Panel on Climate Change (Stocker *et al.*, 2013), national and regional reports (Gasca, 2014; Méndez-Jiménez, 2012; Nieto & Linares, 2011), and from the Andalusian Environmental Information Network.

One advantage of BNs is that it is not necessary to include information for all feature variables in order to be able to make the prediction. Rather, only new information is included as evidences in those variables in which we have knowledge about their change. In our case, evidences are included for the variables of climate, land use and economic sub-models (Table 5.2). Lithology and Geomorphology are consider stable. Whilst no reliable information about social changes is available, no evidences have been introduced into these variables.

1. **Climate change**. Climate change is one of the most important and commonly studied natural drivers modeled under different perspectives and methodologies (Keenan *et al.,* 2011; Rubidge *et al.,* 2011; Quisthoudt *et al.,* 2013). Its interactions with land use provoke changes in the structure of both natural and socioeconomic components through different agents (Anderson-Teixeira *et al.,* 2013; Claesson & Nycander, 2013). In Andalusia, the A2 scenario implies an increase in temperature (of up to 4 degrees in some locations), and changes in rainfall distribution. These data were included as evidences in the Climate sub-model variables; Annual average rainfall and Annual average temperature (Table 5.2).

2. **Land uses change**. The pattern of land uses supports ecosystems and societies due to the fact that any alteration of land use leads to changes in biodiversity, primary production, alterations in soil productivity and the capacity to provide ecosystem services to societies (Lambin *et al.,* 2001). In Spain, several reports based on information from the International Panel on Climate Change have been written to describe the expected change in land uses. Our study used information from the 2040 scenario of land use change (Nieto & Linares, 2011; Méndez-Jiménez, 2012). The expected changes include several that relate to the distribution of vegetation, both crops and forest species. These new values were included into the model as evidences in the following Land use sub-model variables: Dense woodland, Irrigated cropland, and Rainfed cropland (Table 5.2).

3. **Economic change**. Due to the alteration of natural conditions, several changes are expected in the economic and social component of the SES. No reliable information was found about changes in social variables, but economic changes were identified. Two economic sectors are important in Andalusia. The first is the primary sector (livestock and agriculture). Modifications in this sector are reflected in the Land use sub-model (as changes to the extent of Rainfed crops and Irrigated crops variables). The second is the Tourism sector, which could be affected in the future if climate and weather conditions change. Information was collected from regional reports (Méndez-Jiménez, 2012) and introduced as evidences in the following variables: Business activities tax in primary, secondary and tertiary sectors, tertiary sector employment, number of rural hotels, winter and summer water consumption, and farming units cattle and pigs (Table 5.2).

Once the evidences are identified, they are introduced into the corresponding sub-model and propagated from the sub-models to the classifier, updating the distribution of the socio-ecological sectors in Andalusia (Figure 5.6 *ii)*).

TABLE 5.2: Variables in which new evidences are introduced under the scenario of GEC.

| Sub-model | Variables |
|---|---|
| Climate | Annual average rainfall; Annual average temperature |
| Land Use | Dense woodland; Irrigated cropland; Rainfed cropland |
| Economy | Business Activities Tax in primary sector; Business Activities Tax in secondary sector; Business Activities Tax in tertiary sector; Tertiary sector employment; Number of rural hotels; Winter water consumption; Summer water consumption; Farming units cattles; Farming units pigs |

### 5.2.2 Results and Discussion

***A priori* results**

Figure 5.9(a) shows the socio-ecological structure of Andalusia in the current situation, which identifies eight different sectors. Several non-parametric hypothesis test (Chi-square for discrete variables and Kruskall-Wallis for continuous variables) were carried out to check if significant differences exist between these sectors. Using a significance level of 0.05, the tests showed that the differences between sectors are significant.

The sectors are aligned geographically with a southwest to northeast orientation, following a gradient of increasing altitude from the *Guadalquivir* river basin to the peaks of *Sierra Morena* and *Sistema Bético* mountain ranges *Mountain peaks* sector. Figure 5.8 shows the box plot of certain variables, as an example of how this gradient is revealed (*i.e.* rainfed crops surface increase from the mountain peak to the *Guadalquivir* river.)

The first sector, called *Guadalquivir river* covers the river basin area, with its gentle geomorphology of rich sedimentary plains, whose climate enables an important rainfed agriculture to be practiced. This sector is the one most-affected by human activities, containing few natural areas and supporting a wealthy population with a high level of education.

In the foothills of the mountains to the north and south, there are two transitional bands of mixed cropland with forestland, subject to cooler, wetter weather. From the socioeconomic point of view, both areas have significant agricultural activity, but their wealth and structure are different: there are fewer urban areas, lower level of education, lower income per capita, and a change from agricultural areas to one with a high proportion of natural areas (Figure 5.8).

The northern transitional band can be differentiated into two sectors:

- *Northern transition, medium socioeconomic sector*. Located along the edge of the river basin plain, it is dedicated to agricultural activity with a slightly less wealthy population who are educated to a lower level than the *Guadalquivir* sector. This area still contains some areas of significant agricultural investment.

- *Northern transition, low socioeconomic sector*. Located on the hillslopes of the *Sierra Morena*, its landscape is woodland with some patches of rainfed crops. The main difference with the other northern transitional sector is its socioeconomic structure, which corresponds to a sparse population of poorer ageing people.

The differences between these two sectors and the river basin area are slight and gradual. By contrast, to the south, the transition band - also represented by two sectors- shows greater contrast and clearer differences to the river plain:

- *Southern transition, contrast sector*. This is characterized by a steep, eroded relief, containing contrasting areas and an important livestock activity. Close to the river *Guadalquivir*, its socioeconomy comprises a wealthier population with a high agricultural investment. At higher elevations in this sector, the population is characterized by higher migration rates and the economic variables are more depressed than in the previous one.
- *Southern transition, heterogeneous sector*. Located in the highlands of the *Sistema Bético*, this sector presents a heterogeneous landscape with significant forest cover, as well as areas with degraded natural vegetation. Croplands are fewer common than in the lower foothills and the population is characterized by ageing and abandonment areas.

Dotted around within these four zones of the northern and southern transition bands are seven patches, which belong to the *Irrigated cropland* sector. These patches have similar characteristics to the sector within they lie, but they are principally dedicated to irrigated croplands and reveal industrial, rather than agricultural, investment. They also contain a significant proportion of urban landscape. Despite this, these patches have the lowest income per capita and the lowest level of education.

At the top of the mountains are several local patches, which make up the *Mountain peaks sector*. In the *Sierra Morena* this sector appears over 400 m.a.s.l. whilst in the *Sistema Bético*, it lies above 500 m.a.s.l., so the weather is colder and rained in the last one. However, both zones contain more natural landscape (forest and scrubland) with some olive groves in the northern part. The geography of these areas comprises an elevated, steep relief, whilst its sparse and ageing population is mainly dedicated to subsistence agriculture.

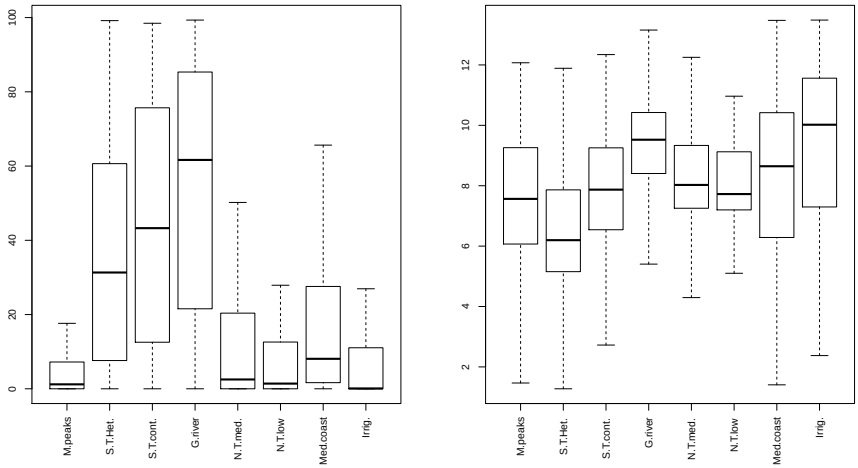
Finally, the *Mediterranean coast sector* lies on the South face of the *Sistema Bético* foothills, over a mixture of sedimentary, metamorphic, volcanic and even karst materials. Its eroded relief is composed of hills, mountains and coastal plains. It is a warm sector, the driest one of Andalusia, and its heterogeneous landscape includes a high proportion of scrubland and sparse vegetation. From the socioeconomic point of view, this sector is mainly dedicated to the primary sector, though contrasts exist between medium income per capita and medium educational level to poorly developed areas. It also has an important tourism sector.

### *A posteriori* results

Figure 5.9(b) shows the socio-ecological structure of Andalusia under the GEC scenario. The number of sectors have decreased to seven. As in the *a priori* situation, Chi-square and Kruskall-Wallis tests were carried out. There are significant differences between the sectors *a posteriori*.

a) Rainfed crops

b) Annual average temperature




c) Forest

d) Income per capita

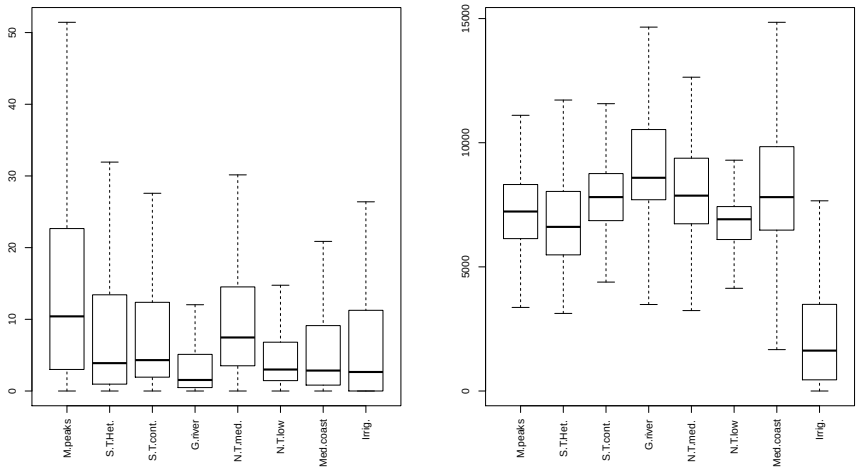



FIGURE 5.8: Extension of some land use (Rainfed crops and Forest expressed in percentage of the grid cell), climate (Annual average temperature express in Celsius) and economic (Income per capita express as a rate) variables in *a priori* sectors. M.peaks, Mountain peaks; S.T.Het, Southern transition, heterogeneous; S.T.cont., Southern transition, contrast; G.river, *Guadalquivir* river basin; N.T.med., Northern transition, medium; N.T.low, Northern transition, low; Med.coast, Mediterranean coast; Irrig., Irrigated cropland.

Under this scenario of change, the socio-ecological structure of the territory is different to *a priori*, which makes some of the new sectors have a different name. This new structure indicates three main sectors, oriented southwest - northeast. These three sectors contain patches of the four sectors dotted within them (Figure 5.9(b)). The gradient corresponding to altitude from the river to the mountain peaks is no longer observed.

The sector called *Woodland in the Sierra Morena foothills* now covers the *Sierra Morena* and part of the *Guadalquivir* river basin, as well as several patches in southern Andalusia. It is characterized by woodland and rainfed landscape on the eroded slopes of dry areas. From the socioeconomic point of view, it is a varied sector with an ageing population and a low level of education.

The next sector is called *Woodland in the Sistema Bético foothills*. It is a continuous area that runs from southwest to northeast through Andalusia, comprising woodland with patches of rainfed crops. It corresponds to areas that are depressed socioeconomically, similar to the previous sector.

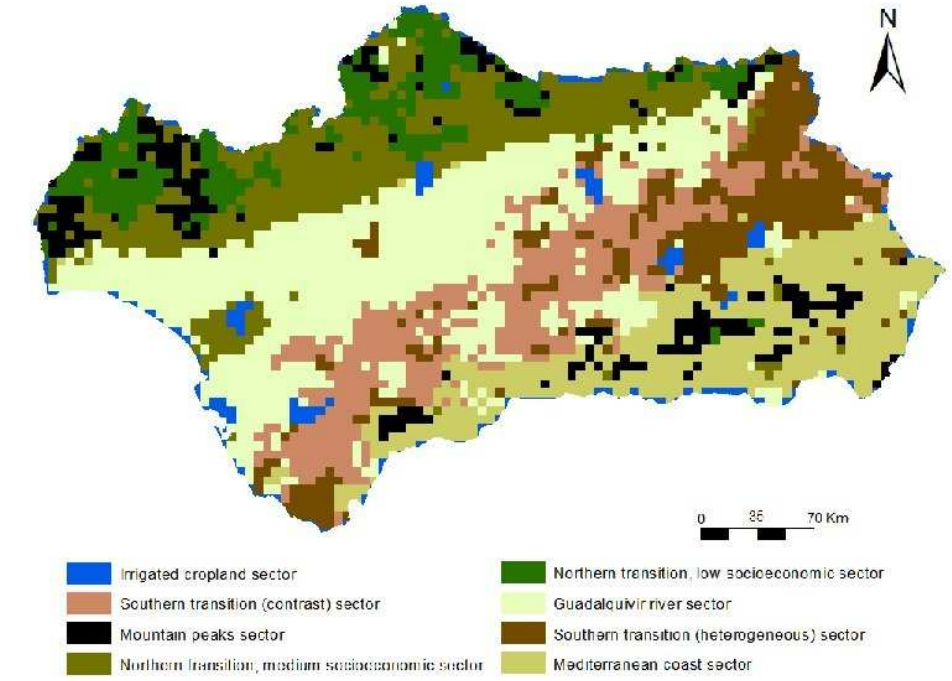
Among them, some agricultural relic areas are found. They support an agricultural society with a high level of education, a positive natural increase and tourist activity. There is now the *Rainfed cropland* sector, comprising several patches within the river basin and the *Sistema Bético* foothills of rainfed agriculture that contains no natural landscapes. In a similar way, *Woodland-croplands* sector is composed of a number of small patches, mostly located in the river basin area, containing both natural and crop landscapes. The *Irrigated croplands* sector is composed of several patches dedicated to irrigated crops.

Lastly, two sectors are found with similar characteristics (and also the same name) as *a priori*, namely the *Mediterranean coast* and the *Mountain peaks* sectors. The *Mediterranean coast* sector covers the same area as before and supports a quite similar socio-ecological structure. In the same way, the landscapes belonging to the*Mountain peaks* sector are still located at the top of the mountain ranges, but they occur only in the *Sistema Bético* whilst this sector has almost disappeared in the case of *Sierra Morena* (Figure 5.9(b)).

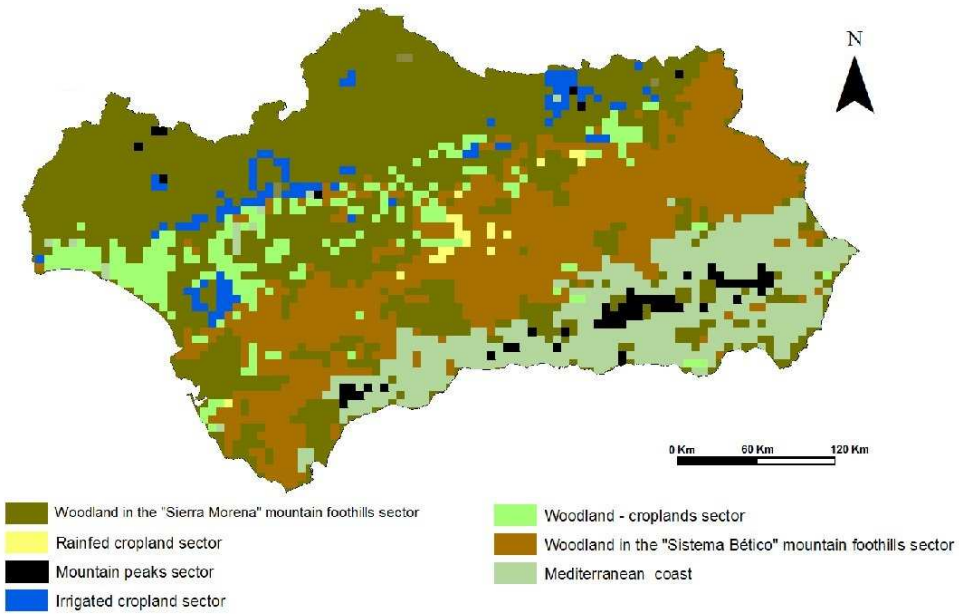
**Comparisons between *A priori* and *A posteriori* situations**

In order to study the dynamics of the structure of the territory, a confusion matrix was drawn up to highlight the differences between the *a priori* and *a posteriori* situation (Table 5.3). This matrix represents the percentage of each sector in the *a priori* situation that is included in each of the *a posteriori* sectors. Also, Figure 5.9 shows both situations, *a priori* and *a posteriori*. From studying this table, it becomes clear that parts of both the northern and southern transitional areas have been incorporated into the *Woodland in the Sierra Morena foothills* and *Woodland in the Sistema Bético mountain foothills* sectors (Table 5.3), with corresponding change in landscape to scrubland and degraded vegetation. From the socioeconomic point of view, the diversity and heterogeneity of the transition band between the river basin and the mountain peaks has been minimized and the variables have become more homogeneous.

Whilst, in the *a priori* situation, agricultural activity extended over the river basin and both mountain foothill areas, under this scenario agricultural activity has been reduced to a number of small patches. Both *Rainfed cropland* and *Woodland-croplands* sectors replace part of the previous *Guadalquivir river* sector. However, the *Irrigated crops* sector




(a) *A priori*




(b) *A posteriori*

FIGURE 5.9: socio-ecological sectors of Andalusia, results *a priori* and *a posteriori*.

TABLE 5.3: Confusion matrix showing the percentage of grid cells in common between each *a priori* and *a posteriori* sectors.

| | | *A posteriori* | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | | **Woodland in *Sierra Morena*** | **Rainfed crops** | **Mountainous peaks** | **Irrigated crops** | **Woodland & crops** | **Woodland in *Sistema Bético*** | **Mediterranean coast** |
| *A priori* | **Irrigated crops** | 14.6 | 0 | 1.5 | 0.9 | 0 | 0.2 | 2.05 |
| | **Southern transition (contrast)** | 3.4 | 42.8 | 0 | 0 | 0.5 | 43.4 | 2.4 |
| | **Mountain peaks** | 8.57 | 0 | 87.7 | 0 | 0 | 0 | 4.8 |
| | **Northern transition, medium** | 26.1 | 0 | 1.5 | 88.2 | 0 | 0.2 | 0.7 |
| | **Northern transition, low** | 16.6 | 0 | 0 | 0 | 0 | 0 | 0 |
| | **Guadalquivir river** | 25.6 | 57.1 | 1.5 | 10.7 | 99.5 | 21 | 4.3 |
| | **Southern transition (heterogeneity)** | 1.3 | 0 | 0 | 0 | 0 | 34.9 | 0 |
| | **Mediterranean coast** | 3.6 | 0 | 7.7 | 0 | 0 | 0.1 | 85.6 |
| | **Total** | 100 | 100 | 100 | 100 | 100 | 100 | 100 |

is no longer located in the same areas as *a priori*; now these occur at higher altitude - within the *Northern transition, medium socioeconomy* (Table 5.3).

The *Mediterranean coast* sector, is a heterogeneous area quite similar to the *a priori* one. From the socioeconomic point of view, they have similar characteristics, but the climate under this A2 scenario is warmer and drier.

Lastly, the *Mountain peaks sector* covers the same geographical area as *a priori*, but the extent of these areas has decreased. Under the A2 scenario of change, the mountain peaks show greater presence of forest and scrublands. The fall in both temperature and rainfall occurs because this sector now occurs at higher altitude (in both areas, this sector is found above 600 m.a.s.l. in the *a posteriori*, whilst in *a priori* corresponded to land above 400-500 m.a.s.l.).

## 5.3 Conclusion

This chapter presents a new methodological proposal based on hBNs hierarchical classifier and applied to identify the socio-ecological structure of a territory. The dynamics of the territory under a scenario of GEC was studied.

Andalusia is a heterogeneous Mediterranean region, where extensive beaches lie only a short distance from high and wild mountain peaks, and where large extensions of homogeneous monocrops lie a short distance from heterogeneous subsistence crops. However, there is a clear difference between the Mediterranean coast and inland Andalusia (which are separated by the *Baetic System*).

Under the current situation, in inland Andalusia there is a clear separation between socio-ecological sectors. There is a transition from the lowland river basin to the mountain peaks, which is reflected by a gradual change from an agriculturally rich society to forestland and rural structure, with high emigration rates, illiteracy and abandonment areas. This heterogeneity implies a wide variety of ecosystems which, in turn,

supports great biodiversity - Andalusia, being a Mediterranean region, is a global biodiversity hotspot (Myers *et al.,* 2000). Inland Andalusia supports a strong economic sector, with opportunities for a huge range of economic activities (tourism, agriculture, and industry between others). However, its socioeconomy is mainly based on extensive (homogeneous) single crop farms, on which a large percentage of the population depend for their livelihood. Under the scenario of GEC, this structure is lost and the diversity and richness of the socioeconomic structure will tend to decrease.

In comparison to the *a priori* situation, changes in the environmental conditions will cause a shift in the optimal growing areas for several crop species (including olive, wheat and barley) (Méndez-Jiménez, 2012). For that reason, the agricultural diversity would be reduced to a number of relict areas and would provoke the irrigated crops to shift to a higher altitudes in the *Guadalquivir* river basin area. In turn, this would provoke changes in the socio-ecological structure of the territory. The loss of socio-ecological heterogeneity would provoke a decrease in the resilience of Andalusian ecosystems (Virah-Sawmy *et al.,* 2009), making them vulnerable to any disturbance from either natural disaster or socioeconomic and political decisions.

In contrast, in the case of the *Mediterranean coast* sector, even though the GEC scenario implies a decrease in the extent of agricultural activities, the socioeconomic characteristics would be hardly affected. This area supports an important tourist industry, apart from agriculture. Due to both increases in temperature and a longer warm season, tourism might benefit under GEC. Coastal areas would see an increase in the tertiary sector (Méndez-Jiménez, 2012). Under the A2 scenario of change, the socioeconomic heterogeneity would help to mitigate the impact on the socio-ecological structure of the territory and the effects of GEC would be less profound than in inland Andalusia.

As far as the *Mountain peaks* sector is concerned, our results show an increase in the surface area of forest, but further work is needed to study these areas, since climate change could provoke the extinction of the species unable to climb in altitude in the search for colder conditions (Méndez-Jiménez, 2012). On the other hand, the warmer conditions would allow an increase in population, including tourism, which might provide an opportunity in these areas to develop a sustainable touristic activity (Méndez-Jiménez, 2012).

Under an A2 scenario of GEC, it is demonstrated how Andalusia would tend to suffer a loss in its inherent territorial heterogeneity. This might involve important losses in the socio-ecological diversity, as well as a decrease in resilience that would leave the territory more vulnerable to impacts arising from political and economic decisions or natural disasters.

# Chapter 6

# Dynamic Bayesian network: Water Reservoir Management

Including time as a component of the models is still a challenge in data mining and decision support systems. Even when several methodologies are applied in environmental sciences to do so, they are mainly developed for specific topics, and their strong mathematical context make them hard to be extended to other fields. In this Chapter, the extension of BNs, the so-called DBNs, are defined and explained in two parts related with a regression problem: *i)* DBNs and BNs are compared in terms of error rate, and *ii)* both DBNs learning and inference processes are applied to a water reservoir dataset and compared in terms of error, structure and complexity. As a summary, both 1 step and 2 steps learning approach provide similar results, so the election of one of them would be based more on the software and algorithm available. For the inference process, even when both approaches give similar results, *Window* approach seems to be more appropriate.

## 6.1 Introduction: Dynamic Bayesian networks

Nowadays, it is widely recognized that including time as a component of models is an important challenge in the field of data mining, reasoning and decision support systems (Russel & Norvig, 2002; Mihajlovic & Petkovic, 2001). In environmental sciences, time series analysis has a wide range of applications, and some models have been successfully applied such as autorregresive models (Davidson *et al.*, 2016; Parmar & Bhardwaj, 2015), hidden Markov models (Lagona *et al.*, 2015; Spezia *et al.*, 2010), order series method (Arya & Zhang, 2015), multi-temporal analysis (Lobo *et al.*, 2015), autocorrelation functions (Farah *et al.*, 2014), functional depth for outliers (Raña *et al.*, 2014), and state space models (Bojarova & Sundberg, 2010). However, temporal models are usually based on specific software or mathematical notation that experts from other areas are not often familiar with. This makes them hard to apply and also, often specific literature is difficult to find (von Asmuth *et al.*, 2012).

BNs can be used to a obtain prediction about the change of the system under some scenarios, but the conclusions obtained cannot be extrapolated to a particular time, nor time series can be handled. For these reasons, the extension of BNs, the so-called Dynamic Bayesian networks (DBNs), has begun to be applied to face this new challenge

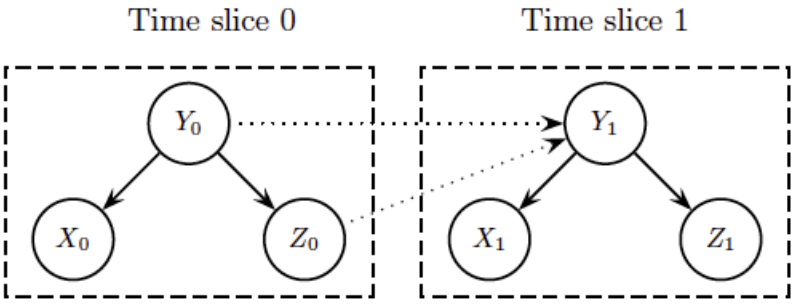



FIGURE 6.1: Example of a Dynamic Bayesian network following the *first-order Markov assumption* with a fixed naïve Bayes structure with two features, $X$ and $Z$, and a class variable, $Y$ composed of 2 time slices. Solid links represent intra-slice arcs, whilst dotted lines represent inter-slice arcs.

(Hill, 2013; Molina *et al.*, 2013). The first attempt to deal with time using BNs appeared in Provan, 1993, which proposed their use for modeling a generic system in each time step, joining the BNs with links which represent the transition from one time to the next. They were defined as (Nicholson & Flores, 2011):

**Dynamic Bayesian networks:** *A long-established extension of BNs that can represent the evolution of variables over time.*

The term dynamic means the system is changing over time, not that the network and the relations between variables change (Murphy, 2002). For simplicity, it is assumed that a DBN is a time-invariant model composed by a sequence of identical BNs representing the system in each time step, and a set of temporal links between variables in the different time steps representing a temporal probabilistic dependence between them (Pérez-Ramiréz & Bouwer-Utne, 2015). Thus, the components of a DBN are (Korb & Nicholson, 2011) (Figure 6.1):

- **Time slice**: the state of the system at a particular time $t$, represented by a static BN identical in each time step.
- **Intra-slice arcs**: the relationships between variables in a time-slice (*i.e.* in Figure 6.1 links between $X_0$ and $Y_0$). They remain constant regardless of the particular time.
- **Inter-slice arcs**: also called temporal arcs, they represent the relationships between variables at successive, or not successive, time slices both *(i)* the same variable over time (*i.e.* in Figure 6.1 links between $Y_0$ and $Y_1$) or *(ii)* between different variables over time (*i.e.* in Figure 6.1 links between $Z_0$ and $Y_1$).

In order to reduce the potential number of temporal parents in the network, and also the computational cost, the *Markov assumption* is followed (Murphy, 2002). That is that *the state of the world at a particular time depends on only a finite history of previous states*. In the simplest case, the current state of the system depends only on the previous state, called a *first-order Markov process* (Figure 6.1). Given these restrictions, a DBN can be represented with only two consecutive time slices (time 0 and time 1) and the relationship between both (Figure 6.1). Only if it is necessary, the DBN can be rolled out and more than two time slices would be represented.

Nowadays two main approaches to learn DBN are considered (Black *et al.,* 2014):

- **In one step**: In this case, the temporal structure (both the structure of the time slice and inter-slice links) is learnt from the data following the time-invariant property, using a specific software, such as Causal Discovery via Minimum Message Length (Korb & Nicholson, 2011; O´ Donnell, 2000).
- **In two steps**: As it was originally proposed, first, the structure of a static model is learnt using all the information available. In a second step, this structure is repeated and connected through (temporal) links. Parameters can be obtained from the data, or elicited by expert knowledge. In this way DBNs can be represented and solved as a kind of "static" model divided into different sub-models (model for time 0, model for time 1, and so on), which allows the available algorithms developed for static BNs to be used.

Once the dynamic model is learnt, inference process can be carried out. Several algorithms have been proposed for both exact inference - *Forward-Backward algorithm* (Baum *et al.*, 1970) and *interface algorithm* (Murphy, 2002) - and approximate inference - *BK algorithm* (Boyen & Koller, 1998) and *FF algorithm* (Murphy & Weiss, 2001) - in DBN. However, there is no BN software that implements these algorithms in such a way that experts from other fields can easily apply them. For that reason, in this Thesis a framework based on the available algorithms and software is proposed for ecologist to use DBN in a easy way. Since DBNs are represented as a set of identical static BNs connected trough (temporal) links, they can be treated as a kind of static model in which each time slice is considered a part of the complete model, as an example of an Object Oriented BN.

By this way, both continuous and discrete data can be included since several models have been developed to represent this type of data within the BN framework. In literature, there are several examples of DBN with hybrid data based on CG models (Wu *et al.*, 2014; Zhang & Dong, 2014). Although, they provide accurate results, the limitation they impose restrict their expansion to other areas and applications. In this Thesis we applied MTE models to DBN.

Even when they are still under development, some real applications of DBNs in environmental sciences can be found in literature. In the works of Hill *et al.,* 2009 and Hill, 2013, hybrid DBNs are applied to the control of streaming climatic data, in an attempt to detect anomalies and errors in the data. Zhang *et al.,* 2012 uses discrete DBNs to integrate data from different times series into a model to accurately estimate the Leaf Area Index in a region of China. In both cases, the application of DBNs is focused on the pre-processing step, trying to correctly collect the data, or merge different data sets. In the paper of Molina *et al.,* 2013, discrete DBNs are learnt as a Decision Support System to predict, for the 2070-2100 period, the effects of Climate Change scenarios in a groundwater systems in Spain.

In this Chapter, the aim is to explore the applicability of DBNs in environmental sciences following the framework mentioned above. To achieve this goal, firstly a comparison between static and dynamic BNs is studied. Secondly, DBN learning (based both on 1-step and 2-steps approaches) and inference methodologies are explained. In both cases, data from the Andalusian Water Reservoir Systems is used. This is the first time that hybrid domains have been included in a DBN based on MTE models for environmental modeling.

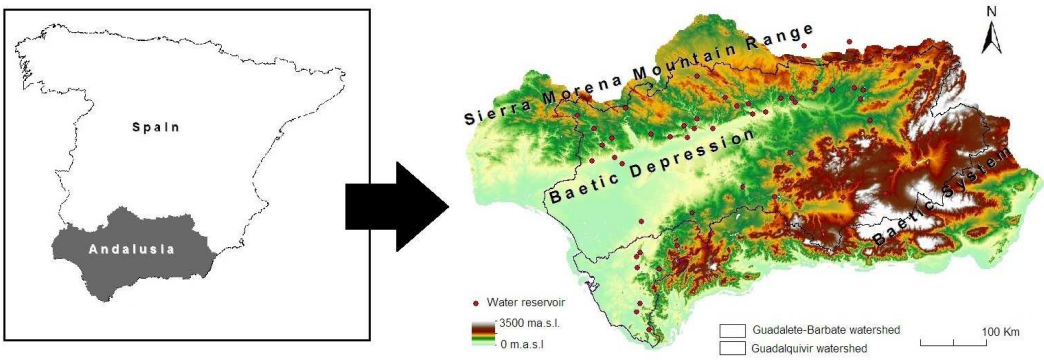



FIGURE 6.2: Study area and the reservoirs selected.

## 6.2 Andalusian water reservoir system

In the SES modeling, freshwater is considered the bloodstream of the biosphere and determines the sustainability of living systems as an indispensable resource for socioeconomic development (Ripl, 2003). A correct governance of water requirements without compromising the future needs is one of the major challenges in environmental sciences (Gordon *et al.*, 2005). Catchment management should be oriented to sustainability and based on ethical principles of human rights, sustaining crucial ecosystem services, and protected ecosystems resilience (Falkenmark & Folke, 2002).

The main characteristic of the annual water cycle in Spain, mainly in Andalusia, is the irregularity. Rainfall spatial and temporal patterns move from extremely strong storms to large drought periods. For that reason, historically, dam construction has been the main solution to this water scarcity and irregularity with more than 1200 dams currently working in Spain. Apart from water and agriculture consumption, the current system of dams has been designed to control and avoid the danger and loss from flood. Also, they provide natural bed with a minimum water flow during drought periods that allow biodiversity to be kept.

Figure 6.2 shows the location of the dams used in this chapter. Andalusia can be divided in two different areas: Inland Andalusia and Mediterraneam coastal area separated by the *Baetic Systems* mountain ranges. Inland Andalusia is composed by the *Baetic Depression* with a similar behavior in terms of rainfall and temperature patterns, whilst coastal area is a really dry area. Information about Andalusia relief, socioeconomy and landscape was provided in previous chapters.

From the hydrological point of view, the relief determines the division in 6 different watersheds with a total of 111 dams. For this Chapter, only those located in the Guadalquivir and Guadalete-Barbate watershed are selected (Figure 6.2), because they shared similar rainfall and temperature patterns, quite different from the rest, resulting in a total of 61.

## 6.3 Water reservoir model: Static *vs.* dynamic models

In the first part of this Chapter, a comparison between static and dynamic BNs is done. The goal is to accurately estimate the amount of water currently stored in the reservoir system, and its evolution over time. Both static and dynamic models based on two constrained structures were used following the 2-steps approach. By this way, algorithms for static BNs can be used. As it was explained in Chapter 4, a NB is a fixed structure consisting of a BN with a single root node and a set of feature variables having only the root node as a parent, in which all the feature variables are independet given the class. A step beyond this is to allow each feature to have one more parent besides the target variable, configuring a *Tree Augmented Naive Bayes* (TAN) structure (Friedman *et al.*, 1997). To learn this structure, the first step is to learnt a directed tree structure with the features variables, using the mutual information with respect to the target variable. In the second step, the relationships between the target variable and each feature are included (Chow & Liu, 1968). These relationships between features are not based on an ecological interpretation but on the amount of information they share with the target variable.

After model learning, a simulated scenario of change is included and some metrics are calculated for a better understanding of the results.

The methodology followed is divided into four steps: *i)* Data collection and pre-processing, *ii)* Models learning, *iii)* Models validation and *iv)* Scenario of change. Elvira software (Elvira-Consortium, 2002) was used for both models learning and validation, and scenario of change propagation.

### 6.3.1 Data collection and pre-processing

Data were collected from the Water Quality Dataset from the Andalusian Regional Environmental Information Network[1] (Andalusian Regional Government) for the 61 reservoirs selected. They consist of 6 continuous and 1 discrete variables collected per month from October 1999 to September 2008.

*Temperature* in °C (T) and *Rainfall* in $m^3/m^2$ (R) represent the climatic conditions in the vicinity of the reservoir. *Percentage Evaporation* (E) is the percentage of the reservoir capacity that evaporates. *Water level* (WL) indicates the height of the water column in m.a.s.l., whilst *Percent Fulness* (PF) expresses the percentage of the reservoir capacity that is currently used, from 0 to more than 100% (following a storm event, the reservoir can exceed the dam capacity). Finally, reservoir management is represented by Amount Discharge and Amount Transfer in. *Amount Discharge* in $m^3$ (AD) refers to the amount of water that is released for ecological, water consumption or regulation purposes. By contrast, *Amount Transfer in* (AT, expressed as a discrete variable with three states: No transfer, less than $0.5m^3$ and more than $0.5m^3$) is the amount of water deliberately added to the reservoir, e.g., pumped in from another reservoir.

With this information two different datasets were created (Figure **??** *ii)*):

- Static dataset. Once the data are collected, values of variables in different times are put together to create a new unique variable, in which time is excluded (*e.g.*

[1] http://www.juntadeandalucia.es/medioambiente/site/rediam

| Dam | $T$ | $R$ | ... |
|---|---|---|---|
| 1 | $T_{oct1999}$ | $R_{oct1999}$ | ... |
| 2 | $T_{oct1999}$ | $R_{oct1999}$ | ... |
| ... | $T_{oct1999}$ | $R_{oct1999}$ | ... |
| 1 | $T_{nov1999}$ | $R_{nov1999}$ | ... |
| 2 | $T_{nov1999}$ | $R_{nov1999}$ | ... |
| ... | $T_{nov1999}$ | $R_{nov1999}$ | ... |
| 1 | $T_{dec1999}$ | $R_{dec1999}$ | ... |
| 2 | $T_{dec1999}$ | $R_{dec1999}$ | ... |
| ... | $T_{dec1999}$ | $R_{dec1999}$ | ... |

(a) Dataset for Static models

| Dam | $T_0$ | $R_0$ | ... | $T_1$ | $R_1$ | ... |
|---|---|---|---|---|---|---|
| 1 | $T_{oct1999}$ | $R_{oct1999}$ | ... | $T_{nov1999}$ | $R_{nov1999}$ | ... |
| 1 | $T_{nov1999}$ | $R_{nov1999}$ | ... | $T_{dec1999}$ | $R_{dec1999}$ | ... |
| ... | ... | ... | ... | ... | ... | ... |
| 2 | $T_{oct1999}$ | $R_{oct1999}$ | ... | $T_{nov1999}$ | $R_{nov1999}$ | ... |
| 2 | $T_{nov1999}$ | $R_{nov1999}$ | ... | $T_{dec1999}$ | $R_{dec1999}$ | ... |
| ... | ... | ... | ... | ... | ... | ... |

(b) Dataset for Dynamic models

FIGURE 6.3: Example of both datasets (for the static (a) and dynamic (b) models) for the *Temperature* (T) and *Rainfall* (R) variables.

in Figure 6.3(a), the variable *Temperature* is configured by taking the temperature data for october 1999, november 1999 and so on). This static dataset has 7 variables and 6588 observations and it was used for static BNs learning and validation.

- Dynamic dataset. For each dam, data are organized into two-time slices, comprising every consecutive pair of months (Figure 6.3(b)). This temporal dataset has 14 variables (temperature at time 0, temperature at time 1, rainfall at time 0, rainfall at time 1, and so on) and 6527 observations [2]. This dataset was used for dynamic BNs learning and validation.

### 6.3.2 BN and DBN learning and validation

The specific goal of these models is to predict, as accurately as possible, the behavior of the continuous variable *Percent Fulness*, which represents a *regression* task. Static BNs consist of a single NB and TAN in which *Percent Fulness* variable is the root node, and the features are the rest of the variables. In the case of the DBN, these structures are repeated and connected through a temporal link between *Percent Fulness* at time 0 and *Percent Fulness* at time 1. Elvira software was used for both structure learning and parameter estimation based on MTE models. Temporal links were learnt from the dataset.

*10-fold Cross Validation* was carried out to compute the *rmse* and validated the model.

### 6.3.3 Scenario of change

DBNs allow the evolution of variables to be studied. As an example, a simulated scenario is proposed: under the current climatic change framework, the model is used for predicting the behavior of *Percent Fulness* variable, assuming that the temperature will rise by 10% and rainfall decrease by 15% in each time step (these values are quite drastic in order to see significant differences in the density function in only 2 months). To carry out the prediction, these new values are included as evidences in variables

[2]Note that the difference in the sample size in both dataset is due to the different organization of the data.

TABLE 6.1: Values for the *rmse* calculated by means of a *10-fold Cross Validation* for each method. NB, Bayesian networks based on naïve Bayes structure; TAN, Bayesian networks based on TAN structure.

| Model | Static models | Dynamic models |
|---|---|---|
| NB | 35.68 | 25.82 |
| TAN | 34.62 | 33.93 |

*Temperature* and *Rainfall* both at time 0 and 1 at the NB dynamic model. Note that the rest of the feature values do not need to be evidenced.

From the water management point of view, it is often interesting to compute the probability that a reservoir reaches a certain level of Percent Fulness, both in the lowest and highest values. As an example, the probability of values below 25% (left tail) and over 80% (right tail) of *Percent Fulness* (for a detailed explanation of how to compute the probability of a range of values, see Chapter 3) were computed.

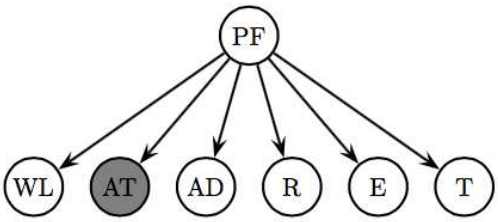
### 6.3.4 Results and Discussion

Figure 6.4 and 6.5 show the structure of both static BN and DBN based on NB and TAN structures. Table 6.1 shows the average *rmse* value of each model, obtained from the *10-fold Cross Validation*. Note that for the static models, *rmse* values are similar, but not in the case of the dynamic ones. Friedman's Test was performed for both static and dynamic models (Figure 6.6) to detect significant differences, returning that dynamic NB outperforms the rest of the models. Furthermore, results show that for the static models no significant differences are found. Comparing static and dynamic TAN even if the *rmse* is slightly lower for the dynamic model, the difference is not significant.

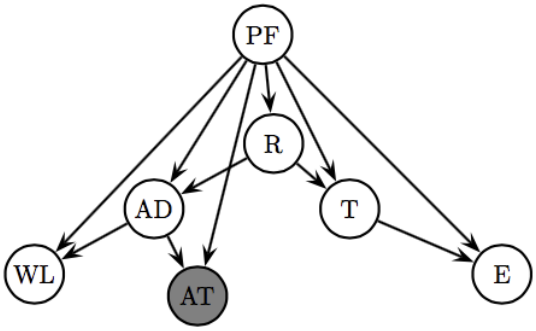
Even when static models seems to provide accurate results, dynamic models add an important advantage related with the inference process. Both static BNs and DBNs allow results to be deeply studied and compared between the situation *a priori* and under the scenario proposed (*a posteriori*), but only DBNs allow their evolution over time to be studied. Figure 6.7 and Table 6.2 show the density function and the metrics obtained from *Percent Fulness* variable at time 0 and 1, both in the current situation (*a priori*), and under this scenario (*a posteriori*).

*A priori*, both PF0 and PF1 variables show a similar behavior, with a probability of both extreme values over 0.5 (in PF0, $0.25$ and $0.33$; in PF1, $0.06$ and $0.50$). However, when the proposed scenario is included, the probability of highest values (right tail) at time 0, increases from $0.33$ to $0.43$, and also the mean (from $59.74$ to $69.77$). By contrast, at time 1 the values tend to be more probable in the middle of the function, with a decrease in the probability of both right and left tails. This information is also confirmed by the behavior of the rest of the metrics in which standard deviation is reduced and the values are more concentrated around the mean.

From the environmental point of view, in the case of a rise in temperature and fall in rainfall, (which can be interpreted as a drought situation), the reservoir will be initially distributed from the smaller and secondaries dams to those that can collect a high amount of water reservoir and satisfied the water demand. Accordingly, at time 0, the values over 80% of Percent Fulness are more probable. If the scenario proposed persists, this would provoke a fall in the amount of water stored in the reservoir of the

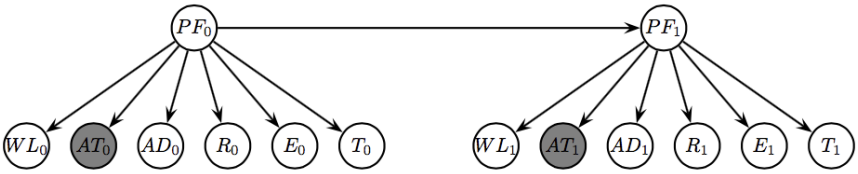



(a) Static NB

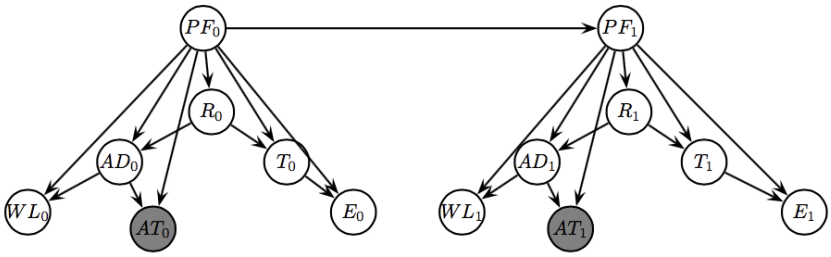



(b) Static TAN

FIGURE 6.4: Static naïve Bayes (a) and TAN (b) structures for the reservoir example. Discrete variable is filled in gray. PF, Percent Fulness; T, Temperature; R, Rainfall; E, Percentage Evaporation; AD, Amount Discharge; AT, Amount Transfer in; WL, Water Level.




(a) Dynamic NB




(b) Dynamic TAN

FIGURE 6.5: Dynamic naïve Bayes (a) and TAN (b) structures for the reservoir example. Discrete variables are filled in gray. PF, Percent Fulness; T, Temperature; R, Rainfall; E, Percentage Evaporation; AD, Amount Discharge; AT, Amount Transfer in; WL, Water Level.

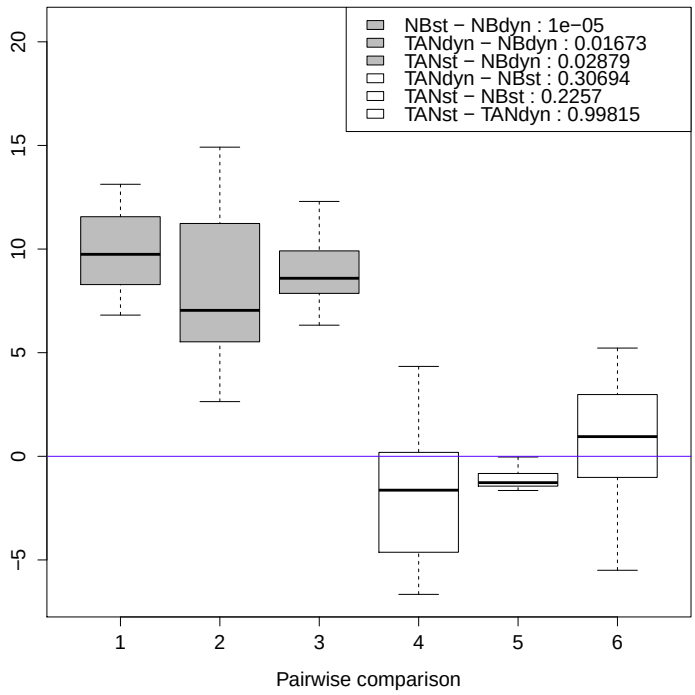



FIGURE 6.6: Box-plot summarizing the results of the pairwise comparison between static (a) and dynamic (b) regression models, p-values are shown in the legend. The gray-shaded boxes indicate significant differences between the corresponding models.

TABLE 6.2: Metrics calculated from the density functions of variables *Percent Fulness* at time 0 (PF0) and 1 (PF1) in both *a priori* and *a posteriori* situations. Var., Variable; SD, Standard Deviation.

| | **A priori** | | | | **A posteriori** | | | |
|---|---|---|---|---|---|---|---|---|
| **Var.** | **Mean** | **SD** | $P(x \leq 25)$ | $P(x \geq 80)$ | **Mean** | **SD** | $P(x \leq 25)$ | $P(x \geq 80)$ |
| PF0 | 59.74 | 40.18 | 0.25 | 0.33 | 69.77 | 39.04 | 0.16 | 0.43 |
| PF1 | 61.11 | 64.70 | 0.06 | 0.50 | 89.45 | 58.45 | 0.02 | 0.41 |

system being modeled. As it was said above, this values are quite extreme with the aim of check the ability of DBN to study the evolution of variables, and it is not a real scenario.

## 6.4 Water reservoir dynamic model: learning and inference

In this second part of this Chapter, several DBNs learning and inference methodologies are studied. Learning process is shown in Figure 6.8 divided into: *i)* Data Collection, *ii)* Structural Learning and *iii)* Parameter Estimation and Model Validation. Since there is no algorithm for directly DBN learning implemented in Elvira. Even though an optimal structure is the best solution, a direct exploration of the causal structure is useful. The structure of the model needs then to be learnt in three steps: *i) Omnigram Explorer* [3] (OE) software is used for an interactive exploration of the data from the Water Reservoir System to detect important relationships between variables. *ii)* this knowledge is included as an input information in CaMML software, in which the causal structure of both 1step and 2-steps DBNs are learnt and, *iv)* both structures were included in Elvira software to estimate the parameters of the models.

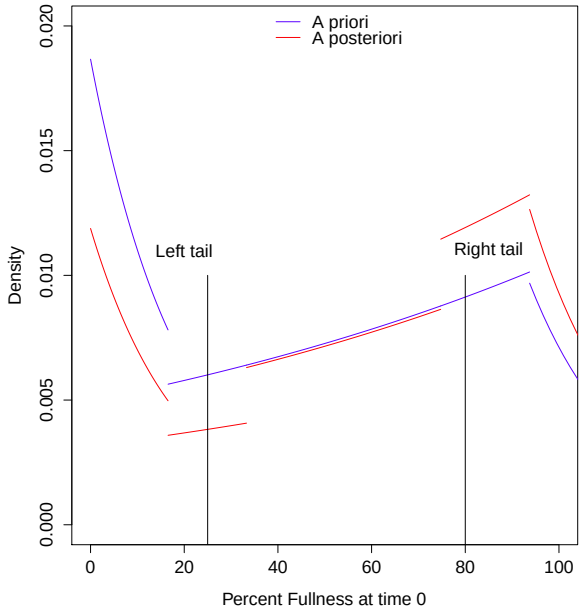
A comparison between both approaches is carried on and one of them is selected to perform the inference process (Figure 6.12). DBN inference can be done by two main methodologies: *Windows* and by a *Roll-out* approach.

### 6.4.1 Data collection and preprocessing

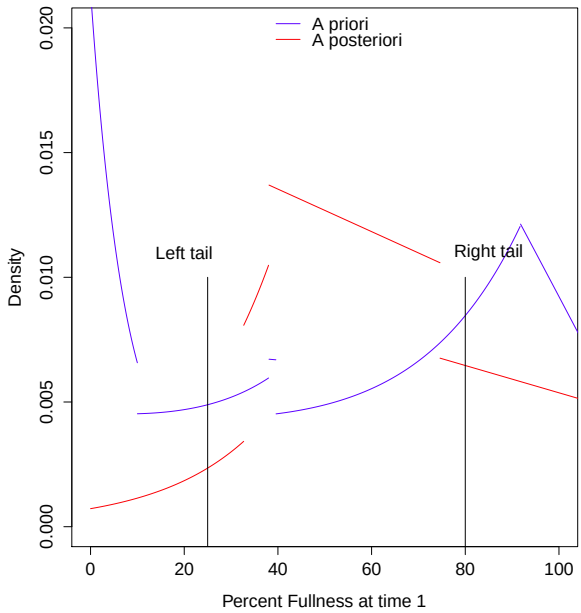
The data from the Water Quality Dataset is again used. For the 61 reservoirs selected a total of 9 variables from october 1999 to september 2007 (for the DBN learning) and from october 2007 to september 2008 (for DBN inference) were collected per month. These new data set includes the variables mentioned in the previous section (now all continuous: *Temperature, Rainfall, Evaporation, Amount Transfer, Amount Discharge, Water Level, Percentage Fullness*) and two added discrete variables: *Reservoir Use* (RU) which represents the main use/s of each reservoir classified by the regional Government of Andalusia (Hydroelectric; General regulation; Irrigation; Human consumption; Industry; No information; Ecological; Irrigation and other; Irrigation and consumption; Consumption and others); and *Time* (Ti) which represents the month.

Data were organized in two dataset in the same way that in the previous section:

[3]For more detail information about the data requirements see the link: http://www.tim-taylor.com/omnigram/




(a) Percent Fulness at time 0




(b) Percent Fulness at time 1

FIGURE 6.7: Probability distribution functions of *Percent Fulness* at time 0 (PF0) and 1 (PF1) variables in dynamic naïve Bayes (NB). Note that probability functions are defined as a piecewise function using MTEs.

- Static dataset: for the data exploration with OE software, and the static structure learning in the 2-steps approach.
- Dynamic dataset: for the dynamic structure learning with CaMML during the 1-step approach, and later on for both 1-step and 2-steps DBN models parameters estimation with Elvira software.

### 6.4.2 DBN learning approaches

The aim is to learn a DBN for modeling the behavior and evolution of the water storage in the reservoir system, representing by the variable *Percentage Fullness*. But also, relationships with the rest of variables need to be studied, so fixed and constrained structures are not suitable for this purpose. Using the software OE, variables and the relationships between them are explored.

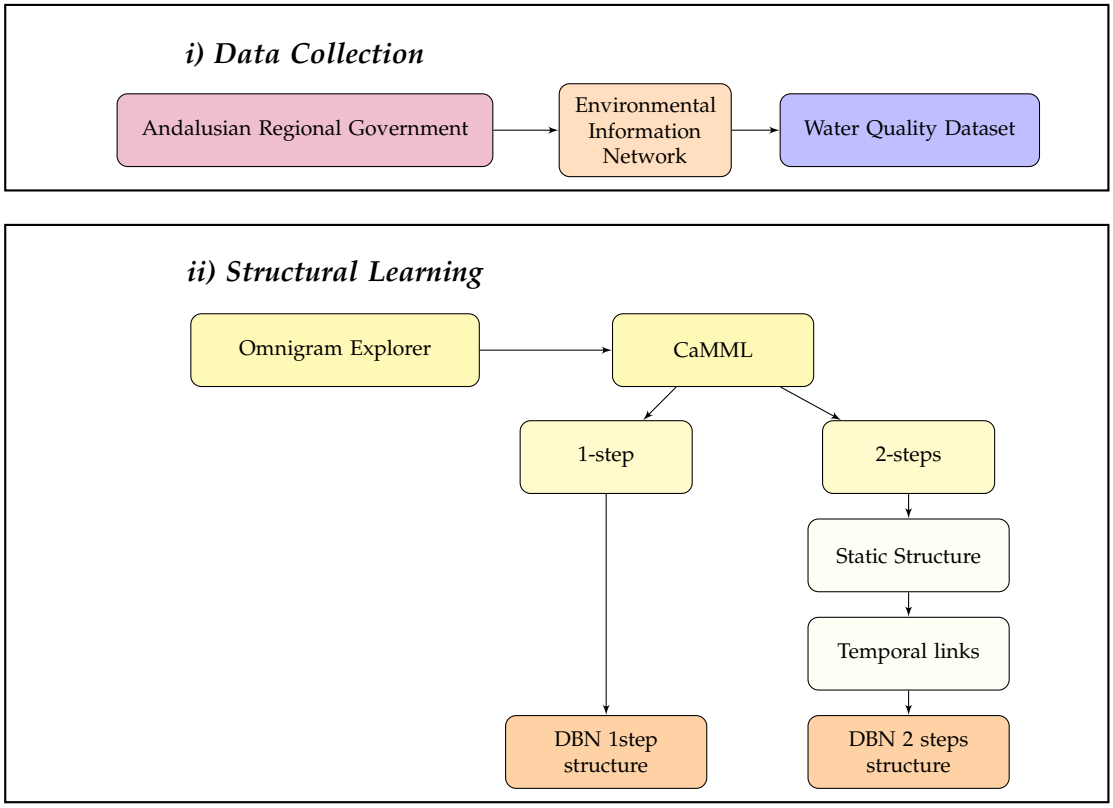
#### *Omnigram Explorer* data exploration prior to modeling

*OE* was designed as a tool for interactive exploration of relations between variables in an agent-based simulation (Taylor *et al.,* 2015). It draws upon ideas for visualization in the *Attribute Explorer* (Spence & Tweedie, 1998), where data is presented in a set of histograms, one per variable.

To begin, a data file containing a joint data sample are loaded and presented by OE in a graphical form (Figure 6.9(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 Figure 6.9(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 format, but a subset of them can be highlighted in a *linking and brushing* process (in dark red color).

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.

- Single Node Brushing (Figure 6.9(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 Figure 6.9(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 (Figure 6.9(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 to match one of the selected ranges are shown in light green; those which match all but two of the ranges



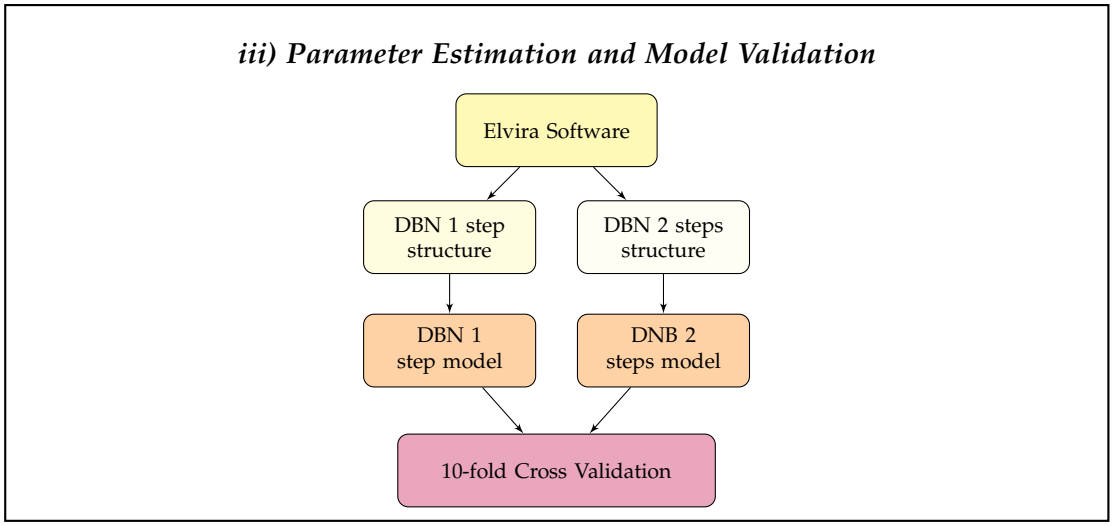




FIGURE 6.8: Outline of the DBN learning methodology.

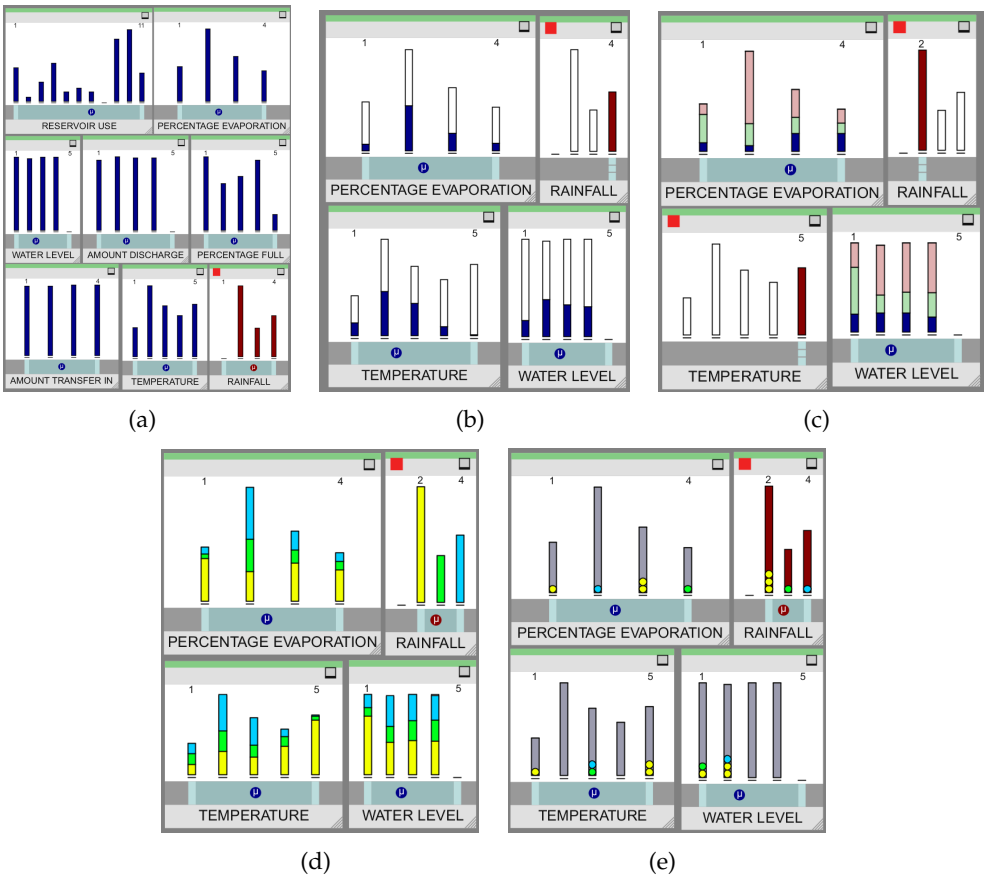



FIGURE 6.9: Initial histograms for the reservoir example with the focus in *Rainfall* variable (a) and modes of interaction in OE for a subset of variables: Single node (b), Multi node (c), Omnibrushing (d) and Sample view (e).

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 (Figure 6.9(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 (Figure 6.9(e)) again uses a single node, and the bins are represented by different colours. The difference is the way data is visualized. Rather than representing 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.

The static dataset is explored by *OE* and some initial understanding of how the variables are related, but also some idea the system's causal structure is gained.

Firstly, the behavior of the system is explored when *Rainfall* is altered. 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*.

If the lowest *Rainfall* value is selected and moved through to the highest value in the Single Node Brushing, a negative relation between *Rainfall* and *Temperature* and a clear positive relation with *Percentage Full*, *Water Level* and *Amount Transfer in* are discovered. 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.

Another variable of prime interest is *Temperature*. A initial view of the variable shows 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.

Lastly, the same procedure is followed with *Percentage Full*. Both *Amount Transfer in* and *Amount Discharge* behave in the same way with respect to *Percentage Full* and that the relation between all three is positive. Pearson correlation between *Amount Transfer in* and *Amount Discharge* conditioned on *Water Reservoir* was computed, 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.

As a summary, *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 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.

#### *2-steps* approach

Causal discovery program CaMMLallows the structure of a BN to be learnt from the available data. It 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 (for more information see Korb & Nicholson, 2011).

It also 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 expert opinion, or, in this case, with the information that data exploration provides. Inspired by *OE*, the following Tiers and Priors were included:

- *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: in a first level *Rainfall* and *Temperature* as parent of *Percentage Evaporation*, *Amount Discharge* and *Amount Transfer in* that are positioned in a second level; and, finally, *Percentage Full* and *Water Level*.

Once the static structure is learnt, it is repeated and included into CaMML, to obtain the temporal links between time slices. By this way, the 2-steps approach is carried out and the final DBN structure is shown in Figure 6.10. Note that a direct link from *Rainfall* and *Percentage Fullness* is not included, but there is a relation between both variables through the variable *Evaporation*.

#### *1-step* approach

One advantage of CaMML is that it allows 1-step DBN learning to be performed. However, both prior and tiers are not allowed, and the causal structure of the DBN is learnt directly from the data. Figure 6.11 shows the structure obtained.

#### Parameter estimation and model validation

Once the dynamic structure of both models are learnt, Elvira software is used to estimate the parameters of the relationships from dynamic dataset. Now both models can be considered as a kind of complex static model divided in two parts: one for time 0, and the other for time 1. Both 1-step and 2-steps DBNs structures were included and

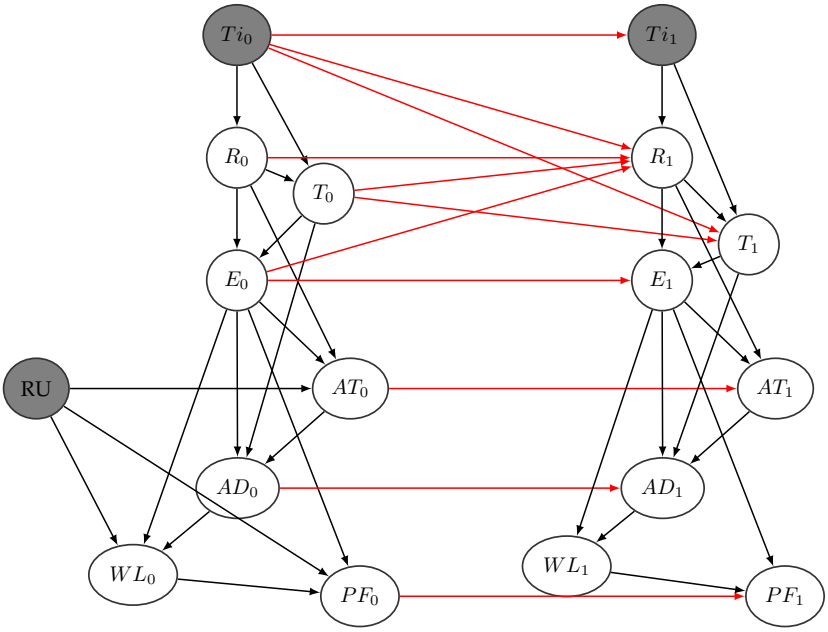



FIGURE 6.10: DBNs model learnt with 2 steps approach. Gray nodes represent discrete variables. Black lines represent the intra-slices links; red lines represent the inter-slices links. RU, Reservoir Use; Ti, Time; T, Temperature; R, Rainfall; WL, Water Level; AT, Amount Transfer; AD, Amount Discharge; E, Evaporation; PF, Percentage Fullness.

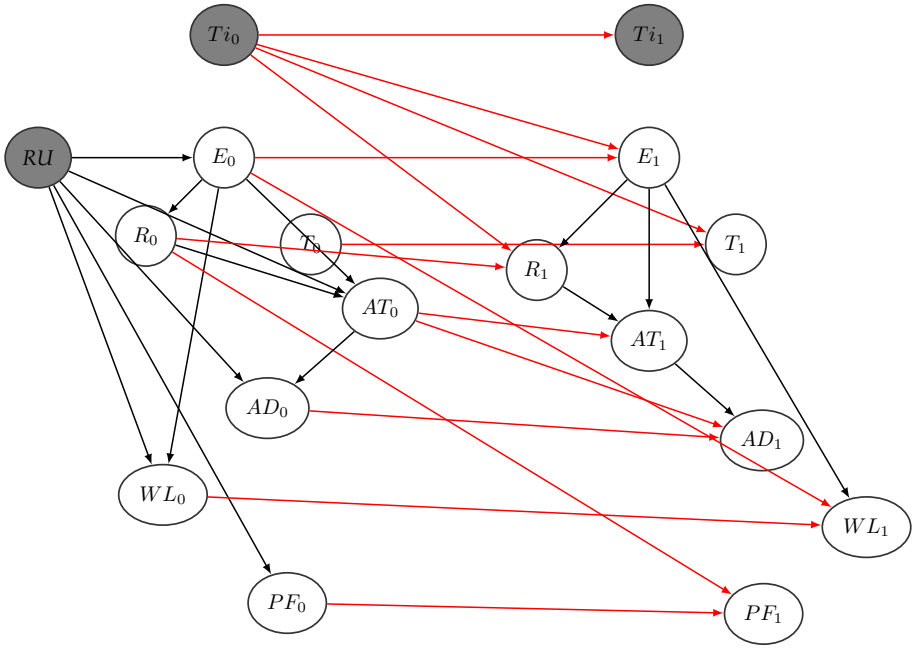



FIGURE 6.11: DBNs model learnt with 1 step approach. Gray nodes represent discrete variables. Black lines represent the intra-slices links; red lines represent the inter-slices links.RU, Reservoir Use; Ti, Time; T, Temperature; R, Rainfall; WL, Water Level; AT, Amount Transfer; AD, Amount Discharge; E, Evaporation; PF, Percentage Fullness.

their parameters estimated based on MTE models. For validation purpose, a *10-fold Cross Validation* was carried out and *rmse* for the variable $PF_1$ was computed.

### 6.4.3 Results

Figures 6.10 and 6.11 show 2-steps and 1-step DBN models, respectively. Table 6.3 shows the *rmse* values for $PF_1$ variable and the number of intra and inter slices links. Both approaches provides similar model structures even when 1-step does not allow prior knowledge to be included. However, a common pattern is found with a sequence of levels in which *Time* is the root node. In a second level climatic variables are related between them, but differences between both approaches are clearly visible. In 2-step DBN model, due to the expert knowledge, *Rainfall* and *Temperature* are both parents of *Evaporation*, whilst in 1-step DBN model is just the contrary. Following down in the network, both variables of reservoir management are found linked between them (*Amount Transfer in* and *Amount Discharge*). Finally, the bottom of the network is represented by *Water Level* and *Percentage Fulness*. Relationships between these last variables and the rest are different in both approaches.

Besides, the number of relationships are different and 1-step approach provides a simpler network structure (Table 6.3). Despite the differences, both models provide similar values of *rmse*. Wilcoxon test was carried out and there is no significative difference between them.

TABLE 6.3: Values of *rmse* in *PF* variable and the number of intra and inter-slices links in both 1-step and 2-steps DBN models.

| Model | *PF rmse* | Intra-slices links | Inter-slices links | Total links |
|---|---|---|---|---|
| 1-step | 36.66 | 15 | 14 | 29 |
| 2-steps | 34.81 | 33 | 11 | 44 |

### 6.4.4 Inference in DBN

Once the models are learnt, 1-step DBN is selected for the inference process. It has less number of links and provides not a high error value. Data from october 2007 to august 2008 were collected per month for the inference process. These data were not included in the previous model learning. The goal is to check the predictive accuracy of DBN and two main methodologies are applied: *Window* and *Roll out*. In each time step, information about both *Rainfall* and *Temperature* is included as *evidences* and the Penniless algorithm (Cano *et al.*, 2002; Cano *et al.*, 2000) is carried out . Finally, the *rmse* of *PF* variable at each time step are obtained and its evolution over time is studied.

#### *Window* approach

Figure 6.12 a) shows the *Window* approach following an example: a DBN with five variables in each time step, and one temporal link. We have information about variables $X_1$ and $X_2$, and want to check the temporal behavior of variable $X_4$.

- Firstly, *evidences* are included into the model (variables $X_1$ and $X_2$ in time 0, marked in red color) and propagated. Mean values for the variable $X_4$ in next

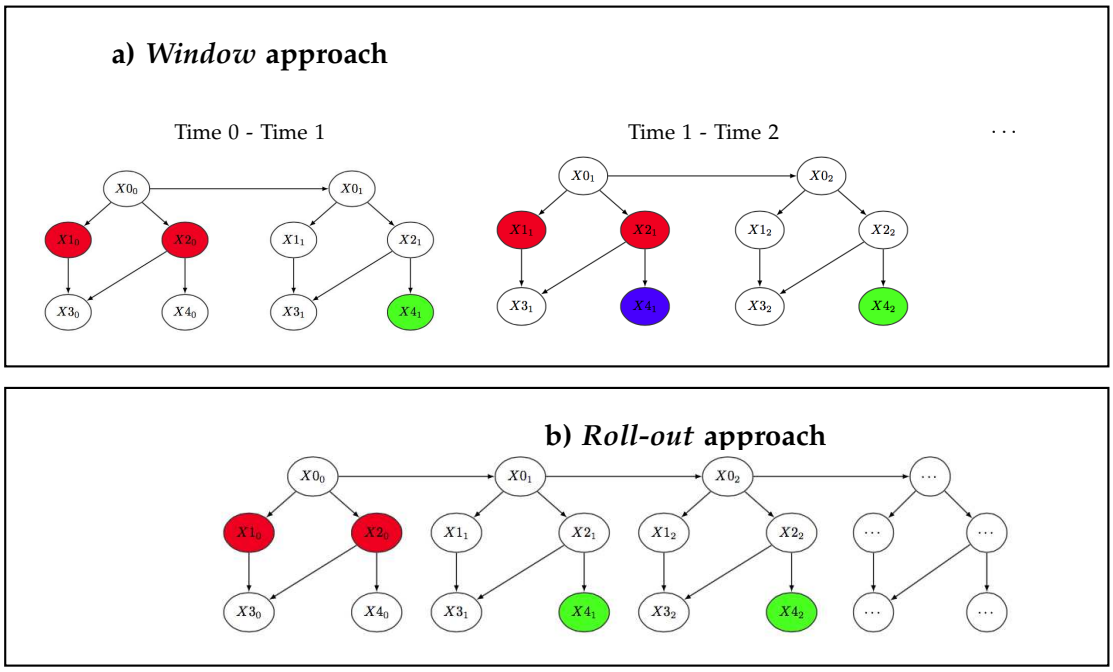



FIGURE 6.12: Outline of the inference in DBN. Red nodes indicate evidenced variables; green nodes indicate the goal variable, blue node indicates a evidenced node obtained from the previous time slice.

time step are obtained and used as an input for the next step (marked in green color).

- If only two time steps are required, the process is stopped. If not, we need to "move the window" in that way that now, we can see time 1 and 2. In this step, evidences are obtained from the prior propagation (values of variables in time 1, marked in blue color), and propagated to the next time step (time 2). The process continues as far as we need.

The idea is to use the DBN as simple as possible maintaining only two time-slices. In the water reservoir model, this process is repeated from time slice 0 to time slice 10, it means, from october 2007 to august 2008, using the information of both *Rainfall* and *Temperature* variables in each time as evidences.

### *Roll-out* approach

Figure 6.12 b) shows the *Roll out* approach. In this case, the network is repeated in the total number of time slices we need. By this way, the new evidences are included simultaneously in all variables rather than in consecutive steps as the previous approach.

In our case, the behavior of *Percentage Fullness* variable want to be studied from october 2007 to august 2008, so the network is *rolled out* to show the eleven time slices. Information about *Rainfall* and *Temperature* is introduced in each time step.

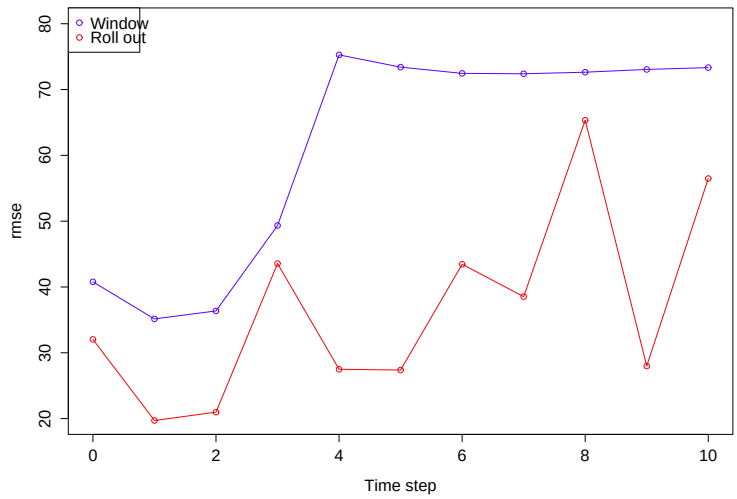



FIGURE 6.13: Evolution of *rmse* value in each time step in both *Window* and *Roll out* inference.

### 6.4.5 Results

Figure 6.13 shows the evolution of the *rmse* of *Percentage Fullness* variable in each time step for both *Window* and *Roll out* approaches. Wilcoxon test was carried out to check if there are differences between both approaches, and there is no significative difference between them. In general, *Roll out* provides less error values than *Windows* approach, mainly due to that the inference process is done in just one step, rather than repeating it. However, *Window* approach reach an estable value of error after 5 time slices, whilst *Roll out* error hardly depends on the time step.

## 6.5 Discussion and Conclusion

In this Chapter, DBN applicability in environmental science is studied using the Water Reservoir Systems of Andalusia. Firstly, a comparison between both static and dynamic BNs was done. DBNs outperform static BNs in terms of error when a temporal problem is modeled. In literature is possible to find several examples in which scenarios of future change are included into static BNs models with the aim of predict future behavior of the system (Dyer *et al.*, 2014; Lowe *et al.*, 2014; Keshtkar *et al.*, 2013). However, DBNs are a more realistic approach to deal with this kind of problems.

One of the main advantages of BNs is that they provide not only a numeric prediction of the class variable but also its probability distribution, which allows several metrics to be calculated (*i.e.* mean, median, probability of a certain range of values). This advantage is extended to the DBNs. As Figure 6.7 shows, the target variable *Percent Fulness* can be studied in detail, its probability distribution, mean, standard deviation, or even the probability of extreme (tail) values. This is quite interesting from the management point of view since it allows, for example, computing the probability of having a low level of water in the reservoir, or by contrast, an amount exceeding its capacity.

Once the advantages of DBNs over static ones have been demonstrate, learning approaches were presented. A comparison between them demonstrated that no significant differences are found in terms of error neither in the structure. The only difference appears in terms of complexity of the network, measured as the number of links. 1 step approach learns the structure, as its name says, in one step and the links between variables are fewer. By contrast, 2 step approach consists of repeating a static structure, and the number of intra-slices tends to be higher. Depending of the goal of the model, both learning methods can be applied and results obtained would be similar.

Lastly, inference process can be also carry out following two methodologies: *Window* and *Roll out*. In that case, even when results in terms of error are not significatively different, a *Window* approach seems to be more recomendable for three reasons:

- Moving the window approach allows maintaining not a complex model, and in each time step only two slices are presented.
- It is not as computational costly as *Roll out* approach.
- The results are not so influenced by the time slice.

However, if finally *Roll out* method is applied, its main advantage is that it allows to see all the evidence propagation just in one step rather than checking the behavior of the system in several windows.

Nowadays, algorithms for DBNs learning and inference are still under development. For a successful application in environmental sciences a further effort is needed to encourage ecologists to apply them.

# Part III

# Concluding remarks

# Chapter 7

# Conclusions

BNs were defined at the beginning of the nineties for solving problems in which a reasoning process was involved. In ecology and environmental sciences their application is still scarce and partially focused on some types of data and problems, for example, discretized data for characterization purpose. All the algorithms and methodological frameworks presented have been previously developed and published in journals of Mathematics and Computer Science areas, and this dissertation just demonstrates how to apply them to real life problems.

In Section 2.5, literature review shows that most applied papers are focused on water research, ecology and environmental, and biodiversity and conservation areas. Trough out this dissertation, several case studies have been included about SES modeling. However, BNs can be applied to any problem and area in environmental science. Also, just a few sets of all algorithms and software designed for BNs modeling are currently used in ecology and environmental modeling. In this Thesis, Elvira software and MTE models were applied. This does not mean that they are the only solution or even the most suitable one for environmental modeling. MTEs were used since they allow both continuous and discrete variables to be included in the same model with no restriction in the structure. Elvira software includes algorithms for characterization, regression and classification models based on MTE learning.

The main contribution of this Thesis is that it presents a complete explanation of what hBNs is used for and how they can be applied in ecology and environmental modeling. To the best of our knowledge, in literature any paper deals with that item. Throughout this manuscript four main problems have been solved by hBNs.

In Chapter 3, hBNs for Characterization purpose was proposed. According to Figure 1, this model goal is recommended when the problem requires the study of the relations between almost all variables included in the model. Due to BN's qualitative part these relationships are easily interpreted, and modifications in the interactions between variables can be assessed through the application of the d-separation concept. Results of this Chapter highlight that BNs are powerful tools for representing complexity and are able to deal with some of the challenges of SES modeling. Firstly, important interactions among components are not omitted, and a balance between model complexity and computational time is achieved. Furthermore, using hBNs mean that the model learning stage is carried out with all the statistical information contained in the data. Thus, the loss of information implied in the discretization process is avoided.

BNs are able to deal with probability propagation, since new information can be introduced into one or more components of the natural or social subsystems and the effects over the rest of the SES can be inferred. Therefore, the current situation and the

new system state can be easily compared because the model results can be displayed together in a single graph showing changes in probability distribution, which allows systemic change to be evaluated. Besides, several statistics can be calculated from these results as mean values, the probability of tails and goodness of fit tests. Taken together, these provide experts with a wide range of tools to aid the decision-making process, regarding the uncertainty in the modeling of systemic change under the SES framework.

Thus, expert knowledge and machine learning techniques can be combined in different ways as an important part in SES modeling. Modeling with the participation of experts and stakeholders has several advantages from a social, instrumental and methodological point of view. Management decisions are usually more effective if all the social groups take part in the management process, sharing information and opinions, each being aware of their responsibility and roles.

Otherwise, if our problem needs one continuous variable to be accurately predicted, a regression problem is faced. In Chapter 4, hBN regression models were compared to a traditional regression technique (MLR). In the previous chapter, hBNs demonstrated their ability to provide accurate results, but in this Chapter we go a step further and compare continuous, discrete, and hybrid approaches with MLR obtaining some advantages for BNs.

Firstly, BNs are able to deal with different types of data, totally continuous, totally discrete, and hybrid, avoiding the loss of information from the discretization. Validation reports better results in terms of error for the BN-based solutions vs. MLR, thus, the continuous model obtain the lowest error. This is explained since MTEs split the probability densities into pieces to better fit the real density determined from data, whilst other traditional techniques use only one function. However, the number of parameters to be determined from data is higher for MTEs. Thus, although more complexity in learning and inference is assumed, the results in terms of error are better. A further advantage of using hBN for regression is that several statistics of interest can be computed from the probability distribution, rather than obtaining just a value as in the MLR.

In addition, not all features must be instantiated to obtain a prediction, *i.e.*, information about the response variable can be obtained even if only partial information about the features is available. It allows scenarios of change to be designed and the behavior of the response variable to be checked. Also, probabilistic information can be extracted from other non-evidenced variables which cannot be done with traditional regression techniques. This means BNs provide a more flexible model, with fewer initial assumptions.

If the variable of interest instead of the continuous one is discrete, we are facing a Classification, not a regression problem. Chapter 5 deals with a complex environmental problem, in which the heterogeneity inherent in the social-natural systems needs to be identified. For solving this problem, several traditional clustering techniques have been applied but some challenges were identified in literature. In this Thesis, we proposed a methodology based on a hierarchical hBNs model.

Traditional clustering usually has a limit on the number of variables that can be included in the model. In contrast, the methodology proposed in this chapter highlights the ability of hBNs to manage datasets containing a large number of variables and observations providing robust and easy-to-interpret results due to the proposed structure. Since it is based on a hierarchical classifier - in which the problem is split into

sub-problems - the model is able to deal with this really complex task, simplifying the problem in the manner of divide and conquer.

The majority of the distances used in traditional *unsupervised* classification methodologies can not deal with both continuous and discrete variables in the same hybrid model. Trough out this dissertation the ability of BNs for dealing with both discrete and continuous data have been demonstrated.

Finally, when data are of different magnitudes, (for example, land use variables are expressed as percentage, whilst some social variables such as age are expressed as a rate or number) some variables could have more impact on the model than the rest, and need to be standardized. Since BNs are based on probability distribution functions, they can cope with those differences without data transformation beforehand.

In these three chapters, hBNs were applied to data in which no temporal behavior was observed. But, problems in ecology and environmental sciences, often present time series data. Chapter 6 copes with DBNs. In other areas of knowledge, such as Health and Life Sciences, DBNs are widely applied and several learning and inference algorithms are used. In environmental sciences, their application is still scarce and further efforts are required to encourage researchers in that way. Most of the applications of static BNs in ecology and environmental sciences follow the same pattern: this tool is used as a model approach for solving a real life problem and a software package is often treated as a black box, so a high percentage of papers are based on discrete or discretized data treated with the same methodology, algorithms and software. Besides, a deep study of these papers reveals that a small percentage of research groups are composed of an interdisciplinary team, with experts from both mathematics, statistics and ecology areas. For that reason, DBNs are still an unknown tool.

Even when some specific algorithms have been proposed in literature, in order to encourage ecologists to use DBNs, a framework in which available static algorithm can be applied was proposed. Both 1 step and 2-step approaches were applied and two different dynamic structures obtained. A comparison between them demonstrated that no significant differences are found in terms of error nor in the structure. The only difference appears in terms of complexity of the network, measured as the number of links. The 1 step approach learns the structure, as its name implies, in one step and the links between variables are fewer. In contrast, the 2 step approach consists of repeating a static structure, and the number of intra-slices tends to be higher. Depending on the goal of the model, both learning methods can be applied and results obtained would be similar.

Lastly, inference process can also be carried out following two methodologies: *Window* and *Roll out*. In that case, even when results in terms of error are not significative different, a *Window* approach seems to be more recomendable for three reasons: *i)* moving the window approach allows maintaining not a complex model, and in each time step only two slices are presented; *ii)* it is not as computationally costly as the *Roll out* approach, and *iii)* the results are not as influenced by the time slice. However, if finally the *Roll out* method is applied, its main advantage is that it allows us to see all the evidence propagation just in one step rather than checking the behavior of the system in several windows.

As a summary, we can conclude this Thesis achieves its initial objectives:

- It has been demonstrated that hybrid Bayesian networks are an appropriate tool in ecology and environmental modeling. The main advantage is their ability to include both discrete and continuous variables in the same model without any change in the structure.
- For *Characterization* problems, hBNs qualitative part allows a visual and easy representation of the model complexity, including a combination between expert knowledge and machine learning techniques. The concept of *d-separation* helps to asses changes in the interactions between variables in an intuitive way.
- *Regression* models based on hBNs provide better results than traditional techniques (MLR) due to MTEs flexibility. Besides, not all features need to be instantiated to get an accurate prediction.
- For *Unsupervised Classification* problems, hBNs have demonstrated their ability to manage large datasets, since the hierarchical model proposed simplify the problem in a divide and conquer way. This involves data to not need to be previously transform and results obtained are easy to interpret.
- Even when a further effort is needed to apply *Dynamic hBNs*, results show that this methodology provides more flexible and visual results.

Some future work can be identified from this dissertation:

- A further development of dynamic BNs that allow an easier application of this methodology to real life problems.
- Missing values is a reality in environmental datasets, thus their treatment with BNs should be considered.
- In this Thesis, spatial relations between observations has not been taken into account. This item should be further researched.
- The study of the strength of the relationships between variables is an important challenge in ecology and environmental sciences. Even when there are methodologies able to deal with this problem, BNs could be used to study it in an intuitive way because of their qualitative interpretation.

# Part IV

# Appendix

# Appendix A

# Variables included in the Classifier model

In this appendix variables including in each Sub-Model of the Classifier model learnt in Chapter 5 are shown.

TABLE A.1: Variables included the **Social** Sub-Model. P/A, Presence / Absence

| Variable | Type of Variable | Units | Thresholds |
|---|---|---|---|
| Rate of school attendance between 14 and17 years old | Continuous | Rate | - |
| Population average age | Discrete | Year | 37.9; 40.9 |
| Number of libraries | Discrete | Number per population in each municipality | P/A |
| Number of Cinemas | Discrete | Number per population in each municipality | P/A |
| Number of private schools | Continuous | Number per population in each municipality | - |
| Number of public schools | Continuous | Number per population in each municipality | - |
| Health care centres | Continuous | Number per population in each municipality | - |
| Number of pharmacies | Continuous | Number per population in each municipality | - |
| Rate of iliteracy | Continuous | Percentage of the municipal population | - |
| Primary studies | Continuous | Percentage of the municipal population | - |
| Secondary studies | Continuous | Percentage of the municipal population | - |
| Tertiary studies | Continuous | Percentage of the municipal population | - |
| National Emigration | Continuous | Percentage of the municipal population | - |
| Foreign Emigration | Continuous | Percentage of the municipal population | - |
| National Immigration | Continuous | Percentage of the municipal population | - |
| Foreign Immigration | Continuous | Percentage of the municipal population | - |
| Natural increase | Continuous | Rate | - |
| Total population | Discrete | Population per 25 $Km^2$ | 474.1; 1320.4 |

TABLE A.2: Variables included the **Economic** Sub-Model.

| Variable | Type of Variable | Units | Thresholds |
|---|---|---|---|
| Employed population | Discrete | Rate | 39.9; 44.3 |
| Internet facilities | Discrete | Number per head of population in each municipality | 9.2; 12.8 |
| Number of bank branches | Discrete | Number per head of population in each municipality | 0.07; 0.09 |
| Unemployment rate | Continuous | Percentage of the municipal population | - |
| Business Activities Tax in primary sector | Continuous | Rate | - |
| Business Activities Tax in secondary sector | Discrete | Rate | 20.0; 24.4 |
| Business Activities Tax in tertiary sector | Discrete | Rate | 71.8; 78.4 |
| Primary sector employment | Discrete | Percentage of the employed population | 16.9; 27.3 |
| Secondary sector employment | Continuous | Percentage of the employed population | - |
| Tertiary sector employment | Continuous | Percentage of the employed population | - |
| Number of agricultural cooperatives | Continuous | Percentage per municipal territory | - |
| Number of home owners | Discrete | Percentage of the total flats in the municipality | 80.6; 86.5 |
| Number of rented homes | Continuous | Percentage of the total flats in the municipality | - |
| Agricultural investment | Discrete | Percentage per municipal territory | 0.44; 22.9 |
| Industrial investment | Discrete | Percentage per municipal territory | 1.6; 38.5 |
| Investment in tertiary sector activities | Discrete | Percentage per municipal territory | 0.01; 8.4 |
| Income per capita | Continuous | Rate | - |
| Number of hotels | Discrete | Percentage per municipal territory | 0.6; 2.1 |
| Number of campsites | Discrete | Percentage per municipal territory | 0.001; 0.08 |
| Number of rural hotels | Discrete | Percentage per municipal territory | 0.027; 0.23 |
| Winter water consumption | Continuous | Percentage per municipal territory | - |
| Summer water consumption | Continuous | Percentage per municipal territory | - |
| Farming units bovines | Continuous | Percentage per municipal territory | - |
| Farming units ovines | Continuous | Percentage per municipal territory | - |
| Farming units goats | Continuous | Percentage per municipal territory | - |
| Farming units equines | Discrete | Percentage per municipal territory | 6.9; 19.47 |
| Farming units pigs | Discrete | Percentage per municipal territory | 23.7; 320.5 |

TABLE A.3: Variables included the **Climate** Sub-Model.

| Variable | Type of Variable | Unit |
|---|---|---|
| Evapotranspiration rate | Continuous | mm per year |
| Annual average temperature | Continuous | Celsius |
| Annual average rainfall | Continuous | mm |
| Spring number of rainfall days | Continuous | days |
| Winter number of rainfall days | Continuous | days |
| Summer average rainfall | Continuous | mm |
| Winter average rainfall | Continuous | mm |

TABLE A.4: Variables included the **Land Use** Sub-Model, expressed as the percentage of the cell surface area.

| Variable | Type of Variable |
|---|---|
| Heterogeneous cropland | Continuous |
| Landscape with scarce vegetation | Continuous |
| Dense Woodland | Continuous |
| Scrubland | Continuous |
| Woodland with scrub | Continuous |
| Woodland with herbaceous vegetation | Continuous |
| Human infrastructure | Continuous |
| Irrigated cropland | Continuous |
| Rainfed cropland | Continuous |
| Water surface | Continuous |

TABLE A.5: Variables included in the **Lithology** Sub-Model, expressed as the percentage of the cell surface area

| Variable | Type of Variable | Thresholds | Variable | Type of Variable | Thresholds |
|---|---|---|---|---|---|
| Amphibolite | Discrete | 0.001; 0.078 | Basic volcanic complex | Discrete | 0.001; 0.069 |
| Clay with red sand | Discrete | 0.001; 0.23 | Clay with marl | Discrete | 0.001; 0.25 |
| Clay with limestone | Discrete | 0.001; 0.09 | Clay with dolomite | Discrete | 0.002; 0.17 |
| Sand | Discrete | 0.001; 0.42 | Sand and marl | Discrete | 0.001; 0.16 |
| Sand and silt | Continuous | - | Silicaceous sandstone | Discrete | 0.001; 0.41 |
| Sandstone with marl | Discrete | 0.001; 0.16 | Calcarenite | Continuous | - |
| Metamorphosized limestone | Discrete | 0.001; 0.14 | Limestone with dolomite | Discrete | 0.001; 0.22 |
| Greywacke | Discrete | 0.001; 0.07 | Volcanic complex | Discrete | 0.001; 0.30 |
| Conglomerates in sand | Discrete | 0.001; 0.22 | Conglomerate in lutite | Discrete | 0.001; 0.10 |
| Quartzite | Discrete | 0.001; 0.12 | Schist and quartzite | Discrete | 0.001; 0.12 |
| Schists with gneiss | Discrete | 0.001; 0.24 | Phyllite | Discrete | 0.001; 0.21 |
| Grabo | Discrete | 0.001; 0.07 | Gneiss | Discrete | 0.001; 0.13 |
| Granite | Discrete | 0.001; 0.18 | Granodiorite | Discrete | 0.001; 0.37 |
| Silt with clay | Discrete | 0.001; 0.48 | Breccia in marl | Discrete | 0.001; 0.13 |
| Marl with limestone | Discrete | 0.001; 0.20 | Marl and gypsum | Discrete | 0.001; 0.19 |
| Marl with sandstone | Discrete | 0.001; 0.16 | Marly limestone | Discrete | 0.001; 0.10 |
| Metabasite | Discrete | 0.011; 0.023 | Mica schist | Discrete | 0.001; 0.28 |
| Marble | Discrete | 0.001; 0.12 | Peridotite | Discrete | 0.001; 0.18 |
| Calcoschist slate | Discrete | 0.001; 0.19 | Quartzite slate | Discrete | 0.001; 0.37 |
| Schisty slate | Discrete | 0.001; 0.36 | Greywacke slate | Discrete | 0.001; 0.49 |
| Volcanic complex of Cabo de Gata | Discrete | 0.001; 0.69 | | | |

TABLE A.6: Variables included in the **Geomorphology** Sub-Model, expressed as the percentage of the cell surface area.

| Variable | Type of Variable | Thresholds | Variable | Type of Variable | Thresholds |
|---|---|---|---|---|---|
| Badland | Discrete | 0.001; 0.17 | Gully | Discrete | 0.001; 0.09 |
| Scree | Discrete | 0.001; 0.022 | Structural outlier | Discrete | 0.001; 0.061 |
| Marl outlier | Discrete | 0.001; 0.087 | Metamorphosized outlier | Discrete | 0.001; 0.077 |
| Gypsum outlier | Discrete | 0.001; 0.12 | Crested hill | Discrete | 0.001; 0.19 |
| Eroded hills | Discrete | 0.001; 0.14 | Peripheral depression | Discrete | 0.0012; 0.23 |
| Piedmont hills | Discrete | 0.001; 0.096 | Structural hill | Discrete | 0.001; 0.15 |
| Conglomerate hills | Discrete | 0.001; 0.067 | Volcanic hill | Discrete | 0.001; 0.083 |
| Hill of intrusive rock | Discrete | 0.001; 0.15 | Gypsum hill | Discrete | 0.001; 0.12 |
| Dissected knoll (outlier) | Continuous | - | Alluvial fan | Discrete | 0.001; 0.036 |
| Crest | Discrete | 0.001; 0.044 | Cuvette | Discrete | 0.001; 0.035 |
| Conserved glacis | Discrete | 0.001; 0.061 | Dissected glacis | Discrete | 0.001; 0.085 |
| River bed | Discrete | 0.001; 0.045 | Colluvia | Discrete | 0.001; 0.037 |
| Floodplain | Discrete | 0.001; 0.11 | Floodplain | Discrete | 0.001; 0.10 |
| Former mudflat | Discrete | 0.001; 0.33 | Glacis | Discrete | 0.001; 0.13 |
| Peneplain | Discrete | 0.0011; 0.37 | Piedmont | Discrete | 0.001; 0.045 |
| Karstified shelf | Discrete | 0.001; 0.16 | Granite pluton | Discrete | 0.001; 0.51 |
| Shallow erosion surface | Discrete | 0.001; 0.11 | Seasonal watercourse | Discrete | 0.001; 0.037 |
| Laminated relief | Discrete | 0.001; 0.38 | Tabletop relief | Discrete | 0.001; 0.059 |
| Appalachian mountain chain | Discrete | 0.001; 0.48 | Intrusive mountain chain | DIscrete | 0.001; 0.068 |
| Metamorphic mountain chain | Discrete | 0.001; 0.077 | Conglomerate mountain chain | Discrete | 0.001; 0.14 |
| Marly mountain chain | Discrete | 0.001; 0.10 | Slate mountain chain | Continuous | - |
| Volcanic mountain chain | Discrete | 0.001; 0.13 | Scarcely dissected erosion relief | Discrete | 0.001; 0.18 |
| Moderately dissected erosion surface | Discrete | 0.001; 0.21 | Highly dissected erosion relief | Discrete | 0.001; 0.20 |
| Peneplanization | Discrete | 0.002; 0.28 | Low terrace | Discrete | 0.001; 0.072 |
| Terrace | Discrete | 0.001; 0.029 | Medium terrace | Discrete | 0.001; 0.091 |

# Bibliography


1. Aguilera, P. A., Fernández, A., Fernández, R., Rumí, R. & Salmerón, A. Bayesian networks in environmental modelling. *Environmental Modelling & Software* **26,** 1376–1388 (2011).
2. Aguilera, P. A., Fernández, A., Ropero, R. F. & Molina, L. Groundwater quality assessment using data clustering based on hybrid Bayesian networks. *Stochastic Environmental Research & Risk Assessment* **27,** 435–447 (2013).
3. Aguilera, P. A., Fernández, A., Reche, F. & Rumí, R. Hybrid Bayesian network classifiers: Application to species distribution models. *Environmental Modelling & Software* **25,** 1630–1639 (2010).
4. Aitkenhead, M. J. & Aalders, I. H. Predicting land cover using GIS, Bayesian and evolutionary algorithm methods. *Journal of Environmental Management* **90,** 236–250 (2009).
5. Ajzen, L. The theory of planned behaviour. *Organ.Behav.Hum.Decis,Process* **50,** 179–211 (1991).
6. An, L. Modeling human decisions in coupled human and natural systems: review of agent-based models. *Ecological Modelling* **229,** 25–36 (2012).
7. Anderberg, M. R. *Cluster Analysis for Applications* (Academic Press, 1973).
8. Anderies, J., Janssen, M. & Ostom, E. A framework to analyze the robustness of socio-ecological systems from an institutional perspective. *Ecology & Society* **9(1),** 1–18 (2004).
9. Andersen, S. K., Olesen, K. G., Jensen, F. V. & Jensen, F. *HUGIN: a shell for building Bayesian belief universes for expert systems* in *Readings in Uncertain Reasoning* (eds Shafer, G. & Pearl, J.) (Kaufmann, San Mateo, 1990), 332–337.
10. Anderson, S. H., Udawatta, R., t. Seobi & Garrett, H. Soil water content and infiltration in agroforestry buffer strips. *Agroforest System* **75,** 5–16 (2009).
11. Anderson-Teixeira, K. J. *et al.* Altered dynamics of forest recovery under a changing climate. *Global Change Biology* **19,** 2001–2021 (2013).
12. Aranzabal, I. D., Schmitz, M. F., Aguilera, P. A. & Pineda, F. D. Modelling of landscape changes derived from the dynamics of socio-ecological systems. A case in a semiarid Mediterraneam landscape. *Ecological Indicators* **8,** 672–685 (2008).
13. Arya, F. K. & Zhang, L. Time series analysis of water quality parameters at Stillaguamish river using order series method. *Stochastic Environmental Research & Risk Assessment* **29,** 227–239 (2015).
14. Atuo, F., Timothy, J. & Peter, U. An assessment of socio-ecological drivers of avian body parts trade in West African rainforests. *Biological Conservation* **191,** 614–622 (2015).
15. Barnard, S. & Elliott, M. The 10-tenets of adaptive management and sustainability: An holistic framework for understanding and managing the socio-ecological system. *Environmental Science /& Policy* **51,** 181–191 (2015).
16. Basurto, X., Gelcich, S. & Ostrom, E. The social-ecological system framework as a knowledge classificatory system for benthic small-scale fisheries. *Global Environmental Change* **23,** 1366–1380 (2013).

17. Baum, L., Peterie, T., Souled, G. & Weiss, N. A maximization technique ocurring tin the statistical analusis of probabilistics functions of Markov chains. *The Annals of Mathematical Statistics* **40(1),** 164–171 (1970).
18. Baur, B. & Bozdag, S. A Canonical correlation Analysis-Based Dynamic Bayesian Network Prior to Infer Gene Regulatory Networks from Multiple Types of Biological Data. *Journal of Computational Biology* **22,** 289–299 (2015).
19. Bellassen, V. *et al.* Multi-criteria spatialization of soil organic carbon sequestration potential from agricultural intensificaation in Senegal. *Climatic Change* **98,** 213–243 (2010).
20. Ben-Bassat, M. *Use of distance measures, information measures and error bounds in feature evaluation* (eds Krishnaiah, P. & Kanal, L.) 773–791 (North Holland Publishing Company, 1982).
21. Berkes, F. & Folke, C. *Linking Social and Ecological Systems: Management Practices and Social Mechanisms for Building Resilience* (ed Cambridge, U.) (Cambridge University Press, 1998).
22. Beverly, J., Flannigan, M., Stocks, B. & Bothwell, P. The association between Northern Hemisphere climate patterns and interannual varaibility in Canadian wildfier activity. *Canadian Journal of Forest Research* **41,** 2193–2201 (2011).
23. Bicik, I., Jelecek, L. & Stepanek, V. Land use changes and their social driving forces in Czechia in the 19th and 20th centuries. *Land Use Policy* **18(1),** 65–73 (2001).
24. Black, A., Korb, K. B. & Nicholson, A. E. *Intrinsic Learning of Dynamic Bayesian Networks* in *PRICAI 2014, LNAI 8862* (eds Pham, N. & Park, S.) (2014), 256–269.
25. Boets, P., Landuyt, D., Everaert, G., Broekx, S. & Goethals, P. L. M. Evaluation and comparison of data-driven and knowledge-supported Bayesian belief networks to assess the habitat suitability for alien macroinvertebrates. *Environmental Modelling & Software* **74,** 92–103 (2015).
26. Bojarova, J. & Sundberg, R. Non-Gaussian state space models in decomposition of ice core time series in long and short time-scales. *Environmetrics* **21,** 562–587 (2010).
27. Borges, J., Lansink, A., Marques Ribeiro, C. & Lutke, V. Understanding farmers' intention to adopt improved natural grassland using the theory of plaaned behavior. *Livest.Sci.* **169,** 163–174 (2014).
28. Boudali, H. & Dugan, J. A discrete-time Bayesian network reliability modeling and analysis framework. *Reliability Engineering and Systems Safety* **87,** 337–349 (2005).
29. Bousquet, F. & Le Page, C. Multi-agent simulations and ecosystem management: a review. *Ecological Modelling* **176(3-4),** 313–332 (2004).
30. Boyen, X. & Koller, D. *Tractable inference for complex stochastic processes* in *Proceedings of the 14th Conference on Uncertainty in Artificial Intelligence* (1998), 33–42.
31. Bressan, G. M., Oliveira, V. A., Hruschka, E. R. & Nicoletti, M. C. Using Bayesian networks with rule extraction to infer risk of weed infestation in a corn-crop. *Engineering Applications of Artificial Intelligence* **22,** 579–592 (2009).
32. Bromley, J., Jackson, N. A., Clymer, O. J., Giacomello, A. M. & Jensen, F. V. The use of Hugin$^R$ to develop Bayesian networks as an aid to integrated water resource planning. *Environmental Modelling & Software* **20,** 231–242 (2005).
33. Burgi, M., Hersperger, A. M. & Schneeberger, N. Driving forces of landscape change - current and new directions. *Landscape Ecology* **19,** 857–868 (2004).
34. Butz, C., Oliveira, J. & Madsen, A. L. Bayesian network inderence using marginal trees. *International Journal of Approximate Reasoning* **68,** 127–152 (2016).

35. Cadenasso, M., Pickett, S. & Grove, J. Dimensions of ecosystem complexity: Heterogeneity, connectivity, and history. *Ecological Complexity* **3,** 1–12 (2006).
36. Caillault, S. *et al.* Influence of incentive networks on landscape changes: A simple agent-based simulation approach. *Environmental Modelling & Software* **45,** 64–73 (2013).
37. Caley, M. J., O'Leary, R. A., Fisher, R., Low-Choy, S. & Johnson, S. What is an expert? A systems perspectie on expertise. *Ecology and Evolution,* 231–242 (2013).
38. Camarero, L. *et al. La población rural de España. De los desequilibrios a la sostenibilidad social* tech. rep. (Obra Social. Fundación la Caixa, 2009).
39. Cano, A., Moral, S. & Salmerón, A. Lazy evaluation in Penniless propagation over join trees. *Networks* **39,** 175–185 (2002).
40. Cano, A., Moral, S. & Salmerón, A. Penniless propagation in join trees. *International Journal of Intelligent Systems* **15,** 1027–1059 (2000).
41. Carpenter, S. R. & Folke, C. Ecology for transformation. *Trends in Ecology and Evolution* **21,** 309–315 (2006).
42. Castelletti, A. & Soncini-Sessa, R. Bayesian networks and participatory modelling in water resource management. *Environmental Modelling & Software* **22,** 1075–1088 (2007).
43. Castelletti, A. & Soncini-Sessa, R. Coupling real-time control and socio-economic issues in participatory river basin planning. *Environmental Modelling & Software* **22,** 1114–1128 (2007).
44. CCA. *La economía de Andalucía: Diagnóstico estratégico* (Servicio de Estudios La Caixa, Barcelona, Spain. Pp 5-121, (Colección Comunidades Autónomas), 2007).
45. Celio, E., Koellner, T. & Grêt-Regamey, A. Modeling land use decisions with Bayesian networks: Spatially explicit analysis of driving forces on land use change. *Environmental Modelling & Software* **52,** 222–233 (2014).
46. Chai, L. E. *et al.* A review on the computational approaches for gene regulatory network construction. *Computers in Biology and Medicine* **48,** 55–65 (2014).
47. Challies, E., Newig, J. & Lenschow, A. What role for social-ecological systems research in governing global teleconnections?. *Global Environmental Change* **27,** 32–40 (2014).
48. Charniak, E. Bayesian Network without tears. *AI MAGAZINE,* 50–63 (1991).
49. Chen, S. H. & Pollino, C. A. Good practice in Bayesian network modelling. *Environmental Modelling & Software* **37,** 134–145 (2012).
50. Cheng, J., Greiner, R., Kelly, J., Bell, D. & Liu, W. Learning Bayesian networks from data: An information-theory based approach. *Artificial Intelligence* **137,** 43–90 (2002).
51. Chow, C. K. & Liu, C. N. Approximating discrete probability distributions with dependence trees. *IEEE Transactions on Information Theory* **14,** 462–467 (1968).
52. Christofides, A., Tanyi, B., Whobrey, D. & Christofides, N. The optimal discretization of probabilitiy density functions. *Computational Statistics and Data Analysis* **31,** 475–486 (1999).
53. Chuman, T. & Romportl, D. Multivariate classification analysis of cultural landscapes: An example from the Czech Republic. *Landscape and Urban Planning* **98,** 200–209 (2010).
54. Claesson, J. & Nycander, J. Combined effect of global warming and increased CO2-concentration on vegetation growth in water-limited conditions. *Ecological Modelling* **256,** 23–30 (2013).

55. Clark, M. Dealing with uncertainty: adaptive approaches to sustainable river management. *Aquatic Conservation: Marine and Freshwater Ecosystems* **12,** 347–363 (2002).
56. Clark, W. C. & Dickson, N. M. Sustainability science: The emerging research program. *PNAS* **100,** 8059–8061 (2003).
57. Cobb, B. R., Shenoy, P. P. & Rumí, R. Approximating Probability Density Functions with Mixtures of Truncated Exponentials. *Statistics and Computing* **16,** 293–308 (2006).
58. Cooper, G. F. The computational complexity of probabilistic inference using Bayesian belief networks. *Artificial Intelligence* **42,** 393–405 (1990).
59. Cooper, G. F. & Herskovits, E. A Bayesian method for the induction of probabilistic networks from data. *Machine Learning* **9,** 309–347 (1992).
60. Davidson, J. E., Stephenson, D. B. & Turasie, A. A. Time series modeling of paleoclimate data. *Environmetrics* **27,** 55–65 (2016).
61. Dawe, N. K. & Ryan, K. L. The faulty three-legged-stool model of sustainable development. *Conservation Biology* **17(5),** 1458–1460 (2003).
62. De-Lucio-Fernández, J., Atauri-Mezquida, J. A., Sastre-Olmos, P. & Martínez-Alandi, C. in, 29–53 (Consejería de Medio Ambiente. Junta de Andalucía. Sevilla, 2003).
63. Dearing, J. *et al.* Safe and just operating spaces for regional social-ecological systems. *Global Environmental Change* **28,** 227–238 (2014).
64. Deng, J. *et al.* Analysis of the ecological conservation behavior of farmers in payment for ecosystem service programs in eco-environmentally fragile areas using social psychology models. *Science of the Total Environment* **550,** 382–390 (2016).
65. Dougherty, J., Kohavi, R. & Sahami, M. *Supervised and Unsupervised Discretization of Continuous Features* in *Machine Learning: Proceedings of the Twelfth International Conference* (ed y S. Russell, A. P.) (Morgan Kaufmann, San Francisco, 1995), 194 –202.
66. Drees, L. & Liehr, S. Using Bayesian belief networks to analyse social-ecological conditions for migration in the Sahel. *Global Environmental Change* **35,** 323–339 (2015).
67. Duda, R. O., Hart, P. E. & Stork, D. G. *Pattern classification* (Wiley Interscience, 2001).
68. Dyer, F. *et al.* The effects of climate change on ecologically-relevant flow regime and water quality attributes. *Stochastic Environmental Research & Risk Assessment* **28,** 67–82 (2014).
69. Eliott, M. The 10-tenets for integrated, successful and sustainable marine management. *Mar. Pollut. Bull.* **74,** 1–5 (2013).
70. Elvira-Consortium. *Elvira: An Environment for Creating and Using Probabilistic Graphical Models* in *Proceedings of the First European Workshop on Probabilistic Graphical Models* (2002), 222–230. <http://leo.ugr.es/elvira>.
71. ESRI. *ArcMap Version 10.0. Environmental Systems Research Institute (ESRI), Redlands, CA* (2006).
72. Falkenmark, M. Freshwater as shared between society and ecosystems: from divided approaches to integrated challenges. *The Royal Society* **358,** 2037–2049 (2003).
73. Falkenmark, M. Society interaction with the water cycle: a conceptual framework for a more holistic approach. *Hydrological Sciences* **42(4),** 451–466 (1997).

74. Falkenmark, M. & Folke, C. The ethics of socio-ecohydrological catchment management: towards hydrosolidarity. *Hydrology and Earth System Sciences* **6(1),** 1–9 (2002).
75. Farah, W. *et al.* Time series analysis of air pollutants in Beirut, Lebanon. *Environmental Monitoring Assessment* **186,** 8203–8213 (2014).
76. Fernández, A., Gámez, J. A., Rumí, R. & Salmerón, A. Data clustering using hidden variables in hybrid Bayesian networks. *Progress in Artificial Intelligence* **2(2),** 141–152 (2014).
77. Fernández, A., Morales, M. & Salmerón, A. Tree augmented naïve Bayes for regression using mixtures of truncated exponentials: Applications to higher education management. *IDA'07. Lecture Notes in Computer Science* **4723,** 59–69 (2007).
78. Filatova, T. & Polhill, G. *Shocks in coupled socio-ecological systems: what are they and how can we model them?* in *Managing Resources of a Limited Planet* (International Congress on Environmental Modelling & Software, Leipzig, Germany, 2012).
79. Filatova, T., Verburg, P., Parker, D. C. & Stannard, C. A. Spatial agent-based models for socio-ecological systems: Challenges and prospects. *Environmental Modelling & Software* **45,** 1–7 (2013).
80. Fischer-Kowalski, M. Analyzing sustainability transitions as a shift between socio-metabolic regimes. *Environ.Innov.Soc.Transit.* **1,** 152–159 (2011).
81. Fletcher, P., Kelble, C., Nuttle, W. & Kiker, G. Using the intregrated ecosystem assessment framework to build consensus and transfer inforamtion to managers. *Ecological Indicators* **44,** 11–25 (2014).
82. Foley, J. A. *et al.* Global Consequences of Land Use. *Science* **309,** 50–574 (2005).
83. Folke, C. Resilience: The emergence of a perspective for social-ecological systems analyses. *Global Environmental Change* **16,** 253–267 (2006).
84. Friedman, N., Geiger, D. & Goldszmidt, M. Bayesian Network Classifiers. *Machine Learning* **29,** 131–163 (1997).
85. García-Álvarez-Coque, J. M. *La agricultura mediterránea en el siglo XXI* (Caja Rural Intermediterránea, Cajamar, Almería. Spain, pp 7-312, 2002).
86. García-Latorre, J. & Sánchez-Picón, A. Dealing with aridity: socio-economic structures and environmental changes in an arid Mediterranean region. *Land Use Policy* **18,** 53–64 (2001).
87. Gasca, A. M. *Guía de escenarios regionalizados de cambio climático sobre España a partir de los resultados del IPCC-AR4* AEMET, Ministerio de Agricultura, Alimentación y Medio Ambiente (2014).
88. Geri, F., Amici, V. & Rocchine, D. Human acticity impact on the heterogeneity of a Mediterraneam landscape. *Applied Geography* **30,** 370–379 (2010).
89. Getoor, L., Rhee, J. T., Koller, D. & Small, P. Understanding tuberculosis epidemiology using structured statistical models. *Artificial Intelligence in Medicine* **30,** 233–256 (2004).
90. Gieder, K. D. *et al.* A Bayesian network approach to predicting nest presence of the federally-threatened piping plover (Charadrius melodus) using barrier island features. *Ecological Modelling* **276,** 38–50 (2014).
91. Giménez-Casalduero, F., Gomariz-Castillo, F. J. & Calvín, J. C. Hierarchical classification of marine rocky landscape as management tool at southeast Mediterranean coast. *Ocean & Coastal Management* **54,** 497–506 (2011).
92. Gordon, L. J. *et al.* Human modification of global water vapor flows from the land surface. *PNAS* **102,** 7612–7617 (2005).

93. Grech, A. & Coles, G. An ecosystem-scale predictive model of coastal seagrass distribution. *Aquatic Conservation: Marine and Freshwater Ecosystems* **20,** 437–444 (2010).
94. Grêt-Regamey, A. & Straub, D. Spatially explicit avalanche risk assessment linking Bayesian networks to a GIS. *Natural Hazards and Earth System Sciences* **6,** 911 –926 (2006).
95. Harlan, S., Declet-Barreto, J., Stefanov, W. & Pelitti, D. Neighborhood effects on heat deaths: Social and environmental predictors of vulnerability in Maricopa county, Arizona. *Environmental Health Perspectives* **121,** 197–204 (2013).
96. Hastie, T., Tibshirani, R. & Friedman, J. *The elements of Statistical Learning. Data Mining, Inference, and Prediction* (Springer, 2009).
97. Heckerman, D. Bayesian networks for data mining. *Data Mining and Knowledge Discovery* **1,** 79–119 (1997).
98. Heckerman, D., Geiger, D. & Chickering, D. Learning Bayesian Networks - The combination of knowledge and statistical-data. *Machine Learning* **20,** 197–243 (1995).
99. Henriksen, H. J. & Barlebo, H. C. Reflections on the use of Bayesian belief networks for adaptive management. *Journal of Environmental Management* **88,** 1025–1036 (2008).
100. Henriksen, H. J., Rasmussen, P., Brandt, G., von Bülow, D. & Jensen, F. V. Public participation modelling using Bayesian networks in management of groundwater contamination. *Environmental Modelling & Software* **22,** 1101–1113 (2007).
101. Hill, D. J. Automated Bayesian quality control of streaming rain gauge data. *Environmental Modelling & Software* **40,** 289–301 (2013).
102. Hill, D., Minsker, B. S. & Amir, E. Real-time Bayesian anomaly detection in streaming environmental data. *Water Resource Research* **45,** 1–16 (2009).
103. Holling, C. S. Understanding the Complexity of Economic, Ecological, and Social Systems. *Ecosystems* **4,** 390–405 (2001).
104. Hong, J., Merrin, G., Peguero, A., Conzalez-Prendes, A. A. & Lee, N. Exploring the Social-Ecological Determinants of Physical Fightin in U.S. Schools: What about Youth in Inmmigrant Families? *Child and Youth Care Forum* **45,** 279–299 (2016).
105. Hufnagl-Eichiner, S., Wolf, S. A. & Drinkwater, L. E. Assessing social-ecological coupling: Agriculture and hypoxia in the Gulf of Mexico. *Global Environmental Change* **21,** 530–539 (2011).
106. IEC, I. *El Modelo Económico Almería basado en la agricultural intensiva. Un modelo de desarrollo alternativo al modelo urbano - industrial* (Caja Rural Intermediterránea, Cajamar, Almería. Spain. Pp 5-27, 2004).
107. IPCC. <http://www.ipcc.ch> (2014).
108. Jackson, L. E. *et al.* Social-ecological and regional adaptation of agrobiodiversity management across a global set of research regions. *Global Environmental Change* **22,** 623–639 (2012).
109. Jain, A. K., Murty, M. M. & Flynn, P. J. Data clustering: a review. *ACM COmputing Surveys* **31(3),** 264–323 (1999).
110. Jensen, F. & Andersen, S. *Approximations in Bayesian belief universes for knowledge-based systems* in *Proceedings of the 6th Conference on Uncertainty in Artificial Intelligence* (1990), 162–169.
111. Jensen, F. V. & Nielsen, T. D. *Bayesian Networks and Decision Graphs* (Springer, 2007).
112. Jensen, F. V., Olesen, K. G. & Andersen, S. K. An algebra of Bayesian belief universes for knowledge based systems. *Networks* **20,** 637 –659 (1990).

113. Jensen, F. V., Lauritzen, S. L. & Olesen, K. G. Bayesian updating in causal probabilistic networks by local computation. *Computational Statistics Quarterly* **4,** 269–282 (1990).
114. Jiménez-Herrero, L. M., Álvarez-Uría, P. & de la Cruz-Leiva, J. L. *Biodiversidad en España. Base de la Sostenibilidad ante el Cambio Global* (ed Ministerio de Medio Ambiente y Medio Rural y Marino) (Observatorio de la Sostenibilidad en España, 2011).
115. Jorgensen, E. State-of-the-art of ecological modelling with emphasis on development of structural dynamic models. *Ecological Modelling* **120,** 75–96 (1999).
116. Keenan, T., Serra, J., Lloret, F., Ninyerola, M. & Sabate, S. Predicting the future of forests in the Mediterranean under climate change, with niche- and process-based models: CO2 matters! *Global Change Biology* **17,** 565–579 (2011).
117. Keshavarz, M. & Karami, E. Institutional adaptation to drought: The case of Fars Agricultural Organization. *Journal of Environmental Management* **127,** 61–68 (2013).
118. Keshtkar, A. R., Slajegheh, A., Sadoddin, A. & Allan, M. G. Application of Bayesian networks for sustainability assessment in catchment modeling and management (Case study: The Hablehrood river catchment). *Ecological Modelling* **268,** 48–54 (2013).
119. Kinzig, A. *et al.* Resilience and Regime Shifts: Assessing Cascading Effects. *Ecology and Society* **11,** 1–23 (2006).
120. Kjærulff, U. *Reduction of computational complexity in Bayesian networks through removal of weak dependencies* in *Proceedings of the 10th Conference on Uncertainty in Artificial Intelligence* (Morgan Kaufmann, San Francisco, 1994), 374–382.
121. Kocabas, V. & Dragicevic, S. Agent-based model validation using Bayesian networks and vector spactial data. *Environment and Planning B: Planning and Design* **36,** 787–801 (2009).
122. Kohavi, R. & John, G. Wrappers for feature subset selection. *Artificial Intelligence* **97,** 273–324 (1997).
123. Korb, K. B. & Nicholson, A. E. *Bayesian Artificial Intelligence* (CRC Press, 2011).
124. Kotova, T., Miklyaeva, I. M., Ogureeva, G. N., Suslova, E. G. & Shvergunova, L. V. Experience in Mapping the Ecological State of the Plant Cover. *Russian Journal of Ecology* **31,** 318–323 (2000).
125. Kozlov, D. & Koller, D. *Nonuniform dynamic discretization in hybrid networks* in *Proceedings of the 13th Conference on Uncertainty in Artificial Intelligence* (1997), 302–313.
126. Lacitignola, D., Petrosillo, I., Cataldi, M. & Zurlini, G. Modelling socio-ecological tourism-based systems for sustainability. *Ecological Modelling* **206,** 191–204 (2007).
127. Lagona, F., Picone, M. & Maruotti, A. A hidden Mark model for the analysis of cylindrical time series. *Environmetrics* **26,** 534–544 (2015).
128. Lahr, J. & Kooistra, L. Environmental risk mapping of pollutants: State of the art and communication aspects. *Science of the Total Environment* **408,** 3899–3907 (2010).
129. Lambin, E. *et al.* The causes of land-use and land-cover change: moving beyond the myths. *Global Environmental Change* **11,** 261–269 (2001).
130. Langseth, H., Nielsen, T. D., Rumí, R. & Salmerón, A. Mixtures of Truncated Basis Functions. *International Journal of Approximate Reasoning* **53,** 212–227 (2012).
131. Langton, C. *Artificial Life* 47 (Addison-Wesley, Reading, 1988).

132. Lauritzen, S. L. Propagation of probabilities, means and variances in mixed graphical association models. *Journal of the American Statistical Association* **87,** 1098–1108 (1992).
133. Lauritzen, S. L. The EM algorithm for graphical association models with missing data. *Computational Statistics and Data Analysis* **19,** 191 –201 (1995).
134. Laws, R. J. & Kesler, D. C. A Bayesian network approach for selecting translocation sites for endangered island birds. *Biological Conservation* **155,** 178–185 (2012).
135. Lee, S. M. & Abbott, P. A. Bayesian networks for knowledge discovery in large datasets: basics for nurse researchers. *Journal of Biomedical Informatics* **36,** 389–399 (2003).
136. Liedloff, A. & Smith, C. S. Predicting a tree change in Australiaś tropical savannas: Combining different types of models to understand complex ecosystem behaviour. *Ecological Modelling* **221,** 2565–2575 (2010).
137. Lim, S., Lee, S. & Cho, S. A modular approach to landmark detection based on a Bayesian network and categorized context logs. *Information Sciences* **330,** 145–156 (2016).
138. Liu, J. *et al.* Coupled Human and Natural Systems. *Ambio* **36,** 639–649 (2007).
139. Liu, R., Zhang, K., Zhang, Z. & Brothwick, A. G. L. Land-use suitability analysis for urban development in Beijing. *Journal of Environmental Management* **145,** 170–179 (2014).
140. Lobo, F. L., Costa, M. P. & Novo, E. M. Time-series analysis of Landsat-MSS/TM/OLI images over Amazonian waters impacted by gold mining activities. *Remote Sensing of Environment* **157,** 170–184 (2015).
141. Lowe, C. D., Gilbert, A. J. & Mee, L. D. Human-environment interaction in the Baltic Sea. *Marine Policy* **43,** 46–54 (2014).
142. Lucas, P. *Restricted Bayesian network Structure Learning* in *Proceedings of the 1st European Workshop on Probabilistic Graphical Models (PGM'02)* (2002), 217–232.
143. Lucena-Moya, P. *et al.* Discretization of continuous predictor variables in Bayesian networks: An ecological threshold approach. *Environmental Modelling & Software* **66,** 36–45 (2015).
144. Lynam, T., Jong, W., Sheil, D., Kusumanto, T. & Evans, K. A Review of tools for incorporating community knowledge, preferences and values into decision making in natural resources management. *Ecology & Society* **12(1):5** (2007).
145. Madsen, A. L. & Jensen, F. V. Lazy propagation: a junction tree inference algorithm based on lazy evaluation. *Artificial Intelligence* **113,** 203–245 (1999).
146. Maes, W. H., Hueuvelmans, G. & Muys, B. Assessment of Land Use Impact on Water-Related Ecosystem Services Capturing the Integrated Terrestrial - Aquatic System. *Environ Sci Technol* **43,** 7324–7330 (2009).
147. Maldonado, A., Ropero, R. F., Aguilera, P., Rumí, R. & Salmerón, A. Continuous Bayesian networks for the estimation of especies richness. *Progress in Artificial Intelligence* **4,** 49–57 (2015).
148. Maldonado, A., Aguilera, P. & Salmerón, A. Continuous Bayesian networks for probabilistic environmental risk mapping. *Stochastic Environmental Research & Risk Assessment* **30(5),** 1441–1455 (2016).
149. Malekmohammadi, B., Kerachian, R. & Zahraie, B. Developing monthly operating rules for a cascade system of reservoirs: Application of Bayesian networks. *Environmental Modelling & Software* **24,** 1420–1432 (2009).

150. Mantyka-Pringle, C. S., Martin, T. G., Moffatt, D. B., Linke, S. & Rhodes, J. R. Understanding and predicting the combined effects of climate change and land-use change on freshwater macroinvertebrates and fish. *Journal of Applied Ecology* **51,** 572–581 (2014).
151. Marcot, B. Metrics for evaluating performance and uncertainty of Bayesian network models. *Ecological Modelling* **230,** 50–62 (2012).
152. Marcot, B. G., Steventon, J. D., Sutherland, G. D. & McCann, R. K. Guidelines for developing and updating Bayesian belief networks applied to ecological modelling and conservation. *Canadian Journal of Forest Research* **36,** 3063–3074 (2006).
153. Marini, S. *et al.* A Dynamic Bayesian Network model for long-term simulation of clinical complications in type 1 diabetes. *Journal of Biomedical Informatics* **57,** 369.376 (2015).
154. Martín-Lopez, B., Gómez-Baggethun, E. & Montes, C. Un marco conceptual para la gestión de las interacciones naturaleza-sociedad en un mundo cambiante. *Cuides* **03,** 229–258 (2009).
155. Matson, P., Parton, W., Power, A. & Swift, M. Agricultural Intensification and Ecosystem Properties. *Science* **277,** 504–509 (1997).
156. Matthews, R., Polhill, J., Gilbert, N. & Roach, A. *Integrationg agent-based social models and biophysical models* in *MODSIM05 - International Congress on Modelling and Simulation: Advances and Applications for Management and Decision Making, Proceedings* (2005), 1617–1623.
157. M.E.A. <`http://www.millenniumassessment.org/en/index.html`> (2003).
158. Meineri, E., Dahlberg, C. J. & Hylander, K. Using Gaussian Bayesian Network to disentangle direct and indirect associations between landscape physiography, environmental variables and species distribution. *Ecological Modelling* **313,** 127–136 (2015).
159. Mena, C., Walsh, S., Frizzelle, B., Xiaozheng, Y. & Malanson, G. Land use change on household farms in the Ecuadrorian Amazon: Design and implementation of an agent-based model. *Applied Geography* **31,** 210–222 (2011).
160. Méndez-Jiménez, M. *Estudio Básico de Adaptación al Cambio Climático.* (ed Consejería de Medio Ambiente. Junta de Andalucía. Sevilla) (2012).
161. Mihajlovic, V. & Petkovic, M. *Dynamic Bayesian Networks: A State of the Art* tech. rep. (Electrical Engineering, Mathematics and Computer Science (EEMCS), 2001).
162. Miller, R. & Blair, P. *Input-Output Analysis Foundations and Extensions* (Cambridge University Press, 2009).
163. Milne, E., Aspinall, R. J. & Vldkamp, T. A. Integrated modelling of natural and social systems in land change science. *Landscape Ecology* **24,** 1145–1147 (2009).
164. Milns, I., Beale, C. M. & Smith, V. A. Revealing ecological networks using Bayesian network inference algorithms. *Ecology* **91,** 1892–1899 (2010).
165. Minsky, M. Steps towards artificial intelligence. *Computers and Thoughts,* 406 –450 (1963).
166. Molina, J. L., Pulido-Veláquez, D., García-Aróstegui, J. & Pulido-Velázquez, M. Dynamic Bayesian Network as a Decision Support tool for assessing Climate Change impacts on highly stressed groundwater systems. *Journal of Hydrology* **479,** 113–129 (2013).
167. Moral, S., Rumí, R. & Salmerón, A. *Approximating conditional MTE distributions by means of mixed trees* in *ECSQARU'03. Lecture Notes in Artificial Intelligence* **2711** (Springer, 2003), 173–183.

168. Moral, S., Rumí, R. & Salmerón, A. *Mixtures of Truncated Exponentials in Hybrid Bayesian Networks* in *ECSQARU'01. Lecture Notes in Artificial Intelligence* **2143** (Springer, 2001), 156–167.
169. Morales, M., Rodríguez, C. & Salmerón, A. Selective naïve Bayes for regression using mixtures of truncated exponentials. *International Journal of Uncertainty, Fuzziness and Knowledge Based Systems* **15,** 697–716 (2007).
170. Morales, M., Rodríguez, C. & Salmerón, A. *Selective naïve Bayes predictor using mixtures of truncated exponentials* in *Proceedings of the International Conference on Mathematical and Statistical Modelling (ICMSM'06)* (2006).
171. Mumford, J. A. & Ramsey, J. D. Bayesian network for fMRI: A primer. *NeuroImage* **86,** 573–582 (2014).
172. Murphy, K. & Weiss, Y. *The factores frontier algorithm for approximate inference in DBNs* in *Proceedings of the 17th Conference on Uncertainty in Artificial Intelligence* (2001), 378–385.
173. Murphy, K. P. *Dynamic Bayesian Networks: Representation, Inference and Learning* PhD thesis (University of California, Berkeley, 2002).
174. Myers, N., Mittenmeier, R. A., Mittenmeier, C. G., da Fonseca, G. A. B. & Kent, J. Biodiversity hotspots for conservation priorities. *Nature* **403,** 853 –858 (2000).
175. Nair, S. S., Preston, B., King, A. & Mei, R. Using landscape typologies to model socioecological systems: Application to agriculture of the United States Gulf Coast. *Environmental Modelling & Software* **79,** 85–95 (2016).
176. Naranjo-Madrigal, H., van Putten, I. & Norman-López, A. Understainding socioecological drivers of spatial allocation choice in a multi-species artisanal fishery: A Bayesian network modeling approach. *Marine Policy* **62,** 102–115 (2015).
177. Nicholson, A. & Flores, J. Combining state and transition models with dynamic Bayesian networks. *Ecological Modelling* **222,** 555–566 (2011).
178. Niederscheider, M., Kuemmerle, T., Muller, D. & Erb, K. Exploring the effects of drastic institutional and socio-economic changes on land system dynamics in Germany between 1883 and 2007. *Global Environmental Change* **28,** 98–108 (2014).
179. Nieto, J. & Linares, P. *Cambio Global España 2020/50. Energía, Economía y Sociedad* (Centro Complutense de Estudios e Información Medioambiental, 2011).
180. Nikodemus, O., Bell, S., Grine, I. & Liepins, I. The impact of economic, social and politic factors on the landscape structure of the Vidzeme Uplands in Latvia. *Landscape and Urban Planning* **70(1/2),** 57–67 (2005).
181. Norgaard, R. B. Coevolutionary development potential. *Land Economics* **60 (2),** 160–173 (1984).
182. Norgaard, R. B. Sustainable development: A co-evolutionary view. *Futures,* 606–620 (1988).
183. O´ Donnell, R. *Flexible Causal Discovery with MML* PhD thesis (Faculty of Information Technology (Clayton). Monash University, Australia, 3800, 2000).
184. Palmsten, M., Holland, K. T. & Plant, N. G. Velocity estimation using a Bayesian network in a critical-habitat reach of the Kootenai River, Idaho. *Water Resource Research* **49,** 5865–5879 (2013).
185. Park, M. H. & Stenstrom, M. K. Classifying environmentally significant urban land uses with satellite imagery. *Journal of Environmental Management* **86,** 181–192 (2008).
186. Park, M. H. & Stenstrom, M. K. Using satellite imagery for stormwater pollution management with Bayesian networks. *Water Research* **40,** 3429–3438 (2006).

187. Park, Y. S., Kwon, Y., Hwang, S. J. & Park, S. Characterizing effects of landscape and morphometric factors on water quality of reservoirs using a self-organizing map. *Environmental Modelling & Software* **55,** 214–221 (2014).
188. Parmar, K. S. & Bhardwaj, R. Statistical, time series, and fractal analysis of full stretch of river Yamuna (India) for water quality management. *Environmenta Science Pollutant Resource* **22,** 397–414 (2015).
189. Pearl, J. Fusion, propagation and structuring in belief networks. *Artificial Intelligence* **29,** 241 –288 (1986).
190. Pearl, J. *Probabilistic reasoning in intelligent systems* (Morgan-Kaufmann (San Mateo), 1988).
191. Pearl, J. *Probabilistic Reasoning in Intelligent Systems: Network of Plausible Inference* (ed Kaufmann, M.) (San Mateo, California, 1988).
192. Pérez, A., Larrañaga, P. & Inza, I. Supervised classification with conditional Gaussian networks: Increasing the structure complexity from naïve Bayes. *International Journal of Approximate Reasoning* **43,** 1–25 (2006).
193. Pérez-Miñana, E., Krause, P. J. & Thornton, J. Bayesian Network fot the management of greenhouse gas emissions in the British agricultural sector. *Environmental Modelling & Software* **35,** 132–148 (2012).
194. Pérez-Ramiréz, P. A. & Bouwer-Utne, I. Use of dynamic Bayesian networks for life extension assessment of ageing systems. *Reliability Engineering and Systems Safety* **133,** 119–136 (2015).
195. Pineda, F. & Montalvo, J. in (eds Halladay, P. & Gilmour, D.) 107–122 (IUCN, Forest Conservation Programme, Gland, 1995).
196. Pollino, C. A., Woodberry, O., Nicholson, A., Korb, K. & Hart, B. T. Parameterisation and evaluation of a Bayesian network for use in an ecological risk assessment. *Environmental Modelling & Software* **22,** 1140–1152 (2007).
197. Poppenborg, P. & Koellner, T. Do attitudes toward ecosystem services determine agricultural land use practices? An analysis of farmers' decision-making in a South Korean watershed. *Land Use Policy* **31,** 422–429 (2013).
198. Porwal, A., Carranza, E. J. M. & Hale, M. Bayesian network classifiers for mineral potencial mapping. *Computers & Geosciences* **32,** 1–16 (2006).
199. Provan, G. M. *Tradeoffs in Constructing and Evaluating Temporal Influence Diagrams* in *Proceedings of the 9th Conference of the Uncertainty in Artificial Intelligence* (1993), 40–47.
200. Quisthoudt, K. *et al.* Disentangling the effects of global climate and regional land-use change on the current and future distribution of mangroves in South Africa. *Biodiversity and Conservation* **22,** 1369–1390 (2013).
201. R Development Core Team. *R: A Language and Environment for Statistical Computing* ISBN 3-900051-07-0. R Foundation for Statistical Computing (Vienna, Austria, 2012). <http://www.R-project.org>.
202. Raña, P., Aneiros, G. & Vilar, J. M. Detection of outliers in functional time series. *Environmetrics* **26,** 178–191 (2014).
203. Raphael, M. G. *et al.* Status and trends of habitats of terrestrial vertebrates in relation to land management in the interior Columbia river basin. *Forest Ecology and Management* **153,** 63–88 (2001).
204. Rapinel, S., Clément, B., Magnanon, S., Sellin, V. & Hubert-Moy, L. Identification and mapping of natural vegetation on a coastal site using a Worldview-2 satellite image. *Journal of Environmental Management* **144,** 236–246 (2014).

205. Ratnapinda, P. & Druzdzel, M. J. Learning discrete Bayesian network parameters from continuous data streams: What is the best strategy? *Journal of Applied Logic* **13,** 628–642 (2015).
206. Rebaudo, F., Crespo-Pérez, V., Silvain, J. F. & Dangles, O. Agent-based modeling of human-indiced spread of invasive species in agricultural landscapes: Insights from the potato moth in ecuador. *JASSS* **14,** 1–10 (2011).
207. Refsgaard, J. C., van der Sluijis, J. P., Hojberg, A. L. & Vanrolleghem, P. A. Uncertainty in the environmental modelling process - A framework and guidance. *Environmental Modelling & Software* **22,** 1543–1556 (2007).
208. Ricci, P. F., Rice, D., Ziagos, J. & Jr, L. A. C. Precaution, uncertainty and causation in environmental decisions. *Environment International* **29,** 1–19 (2003).
209. Ripl, W. Water: the bloodstream of the biosphere. *The Royal Society* **358,** 1921–1934 (2003).
210. Roberts, J. J., Fausch, K. D., Peterson, D. P. & Hooten, M. B. Fragmentation and thermal risks from climate change interact to affect persistence of native trout in the Colorado river basin. *Global Change Biology* **19,** 1383–1398 (2013).
211. Rockstroem, J. Water Resources Management in Smallholder Farms in Eastern and Southern Africa: An Overview. *Physics and Chemistry of the Earth* **25,** 275–283 (2000).
212. Romero, V., Rumí, R. & Salmerón, A. Learning hybrid Bayesian networks using mixtures of truncated exponentials. *International Journal of Approximate Reasoning* **42,** 54–68 (2006).
213. Ropero, R. F., Rumí, R. & Aguilera, P. Modelling uncertainty in social-natural interactions. *Environmental Modelling & Software* **75,** 362–372 (2016).
214. Rounsevell, M., Robinson, D. & Murray-Rust, D. From actors to agents in socio-ecological systems models. *Phil.Trans.R.Soc.B* **367,** 259–269 (2012).
215. Rubidge, E., Monahan, W., Parra, J., Cameron, S. & Brashares, J. The role of climate, habitat, and species co-occurrence as drivers of change in small mammal distributions over the past century. *Global Change Biology* **17,** 696–708 (2011).
216. Rudel, T. K. *et al.* Agricultural intensification and changes in cultivated areas, 1970-2005. *PNAS* **106,** 20675–20680 (2009).
217. Ruiz, R., Riquelme, J. & Aguilar-Ruiz, J. S. Incremental wrapper-based gene selection from microarray data for cancer classification. *Pattern Recognition* **39,** 2383–2392 (2006).
218. Ruiz-Labourdette, D, Martínez, F., Martín-López, B., Montes, C. & Pineda, F. Equilibrium of vegetation and climate at the European rear edge. A reference for climate change planning in mountainous Mediterranean regions. *Int.J. Biometeorol* **55,** 285–301 (2011).
219. Rumí, R. *Modelos de redes bayesianas con variables discretas y continuas* PhD thesis (Universidad de Almería, 2003).
220. Rumí, R. & Salmerón, A. Approximate probability propagation with mixtures of truncated exponentials. *International Journal of Approximate Reasoning* **45,** 191–210 (2007).
221. Rumí, R., Salmerón, A. & Moral, S. Estimating mixtures of truncated exponentials in hybrid Bayesian networks. *Test* **15,** 397–421 (2006).
222. Russel, S. & Norvig, P. in (ed Hirsch, M.) 542–583 (Pearson, 2002).
223. Sahami, M. *Learning limited dependence Bayesian classifiers* in *Second International Conference on Knowledge Discovery in Databases* (1996), 335–338.
224. Salmerón, A., Cano, A. & Moral, S. Importance sampling in Bayesian networks using probability trees. *Computational Statistics and Data Analysis* **34,** 387–413 (2000).

225. Salvati, L. & Carlucci, M. Towards sustainability in agr-forest systems? Grazing intensity, soil degradation and the socioeconomic profile of rural communities in Italy. *Ecological Economics* **112,** 1–13 (2015).
226. Sánchez-Picón, A., Aznar-Sánchez, J. A. & García-Latorre, J. Economic cycles and environmental crisis in arid southeastern Spain. A historical perspective. *Journal of Arid Environments* **75,** 1360–1367 (2011).
227. Santos, E. & Shimony, S. E. *Belief updating by enumerating high-probability independence-based assignments* in *Proceedings of the 10th Conference on Uncertainty in Artificial Intelligence* (1994), 506–513.
228. Scanlon, B. R., Reedy, R., Stonestrom, D., Prudic, D. & Dennehys, K. Impact of land use and land cover change on groundwater recharge and quality in the southwestern US. *Global Change Biology* **11,** 1577–1593 (2005).
229. Schmitz, M. F., Aranzabal, I. D., Aguilera, P. A., Rescia, A. & Pineda, F. D. Relationship between landscape typology and socioeconomic structure Scenarios of change in Spanish cultural landscapes. *Ecological Modelling* **168,** 343–356 (2003).
230. Schmitz, M., Pineda, F., Castro, H., Aranzabal, I. D. & Aguilera, P. *Cultural landscape and socioeconomic structure. Environmental value and demand for tourism in a Mediterranean territory* (Consejería de Medio Ambiente. Junta de Andalucía. Sevilla, 2005).
231. Schmitz, M., Matos, D., De Aranzabal, I. D., Ruiz-Labourdette, D. & Pineda, F. Effects of a protected area on land-use dynamics and socioeconomic development of local populations. *Biological Conservation* **149,** 122–135 (2012).
232. Schwarz, G. Estimating the dimension of a model. *The Annals of Statistics* **6,** 461–464 (1978).
233. Serra, P., Pons, X. & Saurí, D. Land-cover and land-use change in a Mediterranean landscape: A spatial analysis of driving forces integrating biophysical and human factors. *Applied Geography* **28(3),** 189–209 (2008).
234. Shenoy, P. P. & Shafer, G. in *Uncertainty in Artificial Intelligence, 4* (eds Shachter, R., Levitt, T., Lemmer, J. & Kanal, L.) 169–198 (North Holland, Amsterdam, 1990).
235. Shenoy, P. P. & West, J. C. Inference in hybrid Bayesian networks using mixtures of polynomials. *International Journal of Approximate Reasoning* **52,** 641–657 (2011).
236. Shenoy, P. P. & West, J. C. *Mixtures of polynomials in hybrid Bayesian networks with deterministic variables* in *Proceedings of the 8th Workshop on Uncertainty Processing (WUPES'09)* (2009), 202–212.
237. Shenton, W., Hart, B. T. & Chan, T. U. A Bayesian network approach to support environmental flow restoration decisions in the Yarra river, Australia. *Stochastic Environmental Research & Risk Assessment* **28,** 58–65 (2014).
238. Smajgl, A., G., B. D., Valbuena, D. & Hiuigen, M. Empirical characterisation of agent behaviours in socio-ecological systems. *Environmental Modelling & Software* **26,** 837–844 (2011).
239. Smith, G. & Brennan, R. E. Losing our way with mapping: Thinking critically about marine spatial planning in Scotland. *Ocean & Coastal Management* **29,** 210–216 (2012).
240. Smith, R. I., Dick, J. M. & Scott, E. M. The role of statistics in the analysis of ecosystem services. *Environmetrics* **22,** 608–617 (2011).
241. Solomon, S. *et al. Contribution of Working Group I to the Fourth Assessment Report of the Intergovernmental Panel on Climate Change* (Cambridge University Press, 2007).
242. Solstrand, M. Institutional challenges for effective governance of consumptive wildfire tourism: case studies of marine angling tourism in Icelan and Norway. *Maritime Studies* **14,** 1–27 (2015).

243. Spence, R. & Tweedie, L. The attribute explorer: information synthesis via exploration. *Interacting with Computers* **11,** 137–146 (1998).
244. Spezia, L., Futter, M. N. & Brewer, M. J. Periodic multivariate Normal hidden Markov models for the analysis of water quality time series. *Environmetrics* **22,** 304–317 (2010).
245. Spirtes, P., Glymour, C. & Scheines, R. *Causation, prediction and search* (Springer Verlag, 1993).
246. Staudhammer, C. *et al.* Predictors, spatial distribution, and occurrence of woody invasive plants in subtropical urban ecosystems. *Journal of Environmental Management* **155,** 97–105 (2015).
247. Sterzel, T. *et al.* Armed conflict distribution in global drylands through the lens of a typology of socio-ecological vulnerability. *Regional Environmental Change* **14,** 1419–1435 (2014).
248. Stocker, T. *et al. Climate Change 2013. The Physical Science Basis. Working Group I Contribution to the fifth Assessment Report of the Intergovermental Panel on Climate Change.* WMO, UNEP (2013).
249. Stone, M. Cross-Validatory Choice and Assessment of Statistical Predictions. *Journal of the Royal Statistical Society. Series B (Methodological)* **36 (2),** 111–147 (1974).
250. Stow, C., Roessler, C., Borsuk, M. E. & Bowen, K. J. D. Reckhow. Comparison of Estuarine water quality models for total maximum daily load development in Neuse river estuary. *Journal of Water Resources Planning and Management* **129,** 307–314 (2003).
251. Strand, G. H. Uncertainty in classification and delineation of landscapes: A probabilistic approach to landscape modeling. *Environmental Modelling & Software* **26** (2011).
252. Strijker, D. Marginal lands in Europe. Causes of decline. *Basic and Applied Ecology* **6,** 99–106 (2005).
253. Suh, S., Wieidema, B., Shmidt, J. & Heijungs, R. Generalized make and use framework for allocation in lyfe cycle assessment. *J.Ind.Ecol.* **14,** 335–353 (2010).
254. Tanner, M. A. & Wong, W. H. The calculation of posterior distributions by data augmentation. *Journal of the American Statistical Association* **82,** 528–550 (1987).
255. Taylor, P., Fahrig, L., Henein, K. & Merriam, G. Connectivity is a vital element of lanscape structure. *Oikos* **68(3),** 571–573 (1993).
256. Taylor, T., Dorin, A. & Korb, K. *Omnigram Explorer: A Simple Tool for the Initial Exploration of Complex Systems* Submitted to ECAL 2015. 2015.
257. Ticehurst, J. L., Newham, L. T. H., Rissik, D., Letcher, R. A. & Jakeman, A. J. A Bayesian network approach for assessing the sustainability of coastal lakes in New South Wales, Australia. *Environmental Modelling & Software* **22,** 1129–1139 (2007).
258. Tiller, R., Gentry, R. & Richards, R. Stakeholder driven future scenarios as an element of interdisciplinary management tools; the case of future offshore aquaculture development and the potential effects on fishermen in Santa Barbara, California. *Ocean & Coastal Management* **73,** 127–135 (2013).
259. Toda, M., Yokozawa, M., Emori, S. & Hara, T. More asymmetric tree competition brings about more evapotranspiration and less runoff from the forest ecosystems: A simulation study. *Ecological Modelling* **221,** 2887–2898 (2010).
260. Trincsi, K., Pham, T. T. H. & Turner, S. Mapping mountain diversity: Ethnic minorities and land use land cover change in Vietnam's borderlands. *Land Use Policy* **41,** 484–497 (2014).

261. Turner, B. *et al.* A framework for vulnerability analysis in sustainability science. *PNAS* **100,** 8074–8079 (2003).
262. Turner, R. K. *et al.* Coastal management for sustainable development: analysing environmental and socio-economic changes on the UK coast. *Geographical Journal* **164(3),** 269–281 (1988).
263. Uusitalo, L. Advantages and challenges of Bayesian networks in environmental modelling. *Ecological Modelling* **203,** 312–318 (2007).
264. Van Gerven, M., Taal, B. & Lucas, P. Dynamic Bayesian Networks as a prognostic models for clinical patient management. *Journal of Biomedical Informatics* **41,** 515–529 (2008).
265. Van der Gaag, L. C. Bayesian Belief Networks: Odds and Ends. *Computer J.* **39,** 97–113 (1996).
266. Van Holle, V. *et al.* Social and physical environmental correlates of adults' weekendd sitting time and moderating effects of retirement status and physical health. *International Journal of Environmental Research and Public Health* **11(9),** 9790–9810 (2014).
267. Varis, O. & Kuikka, S. Learning Bayesian decision analysis by doing: lessons from environmental and natural resources management. *Ecological Modelling* **119,** 177–195 (1999).
268. Verhoog, R., Ghorbani, A. & Dijkema, G. Modelling socio-ecological systems with MAIA: a biogas infrastructure simulation. *Environmental Modelling & Software* **81,** 72–85 (2016).
269. Vezeanu, C., Grigor-Pop, O., Gruia, R. & Marculescu, A. Geospatial techniques in the cartography and management of habitats in Piatra Craiului National Park. *Environmental Engineering and Management Journal* **9,** 1611–1617 (2010).
270. Vilizzi, L. *et al.* Model development of a Bayesian Belief Network for managing inundation events for wetland fish. *Environmental Modelling & Software* **41,** 1–14 (2013).
271. Villamagna, A. M., Mogollón, B. & Angermeier, P. L. A multi-indicator framework for mapping cultural ecosystem services: The case of freshwater recreational fishing. *Ecological Indicators* **45,** 255–265 (2014).
272. Virah-Sawmy, M., Gillson, L. & Willis, K. J. How does spatial heterogeneity influence resilience to climatic changes? Ecological dynamics in southeast Madagascar. *Ecological Monographs* **79(4),** 557–574 (2009).
273. Voinov, A. & Bousquet, F. Modelling with stakeholders. *Environmental Modelling & Software* **24,** 1268–1281 (2010).
274. Von Asmuth, J. R. *et al.* Software for hydrogeologic time series analysis, interfacing data with physical insight. *Environmental Modelling & Software* **38,** 178–190 (2012).
275. Vu, Q., Q.B., L., Frosard, E. & Viek, P. Socio-economic and biophysical determinants of land degradation in Vietnam: An integrated causal analysis at the national level. *Land Use Policy* **36,** 605–617 (2014).
276. Walker, B., Abel, N., Stafford-Smith, D. M. & Langridge, J. in (eds Reynolds, J. & Stafford-Smith, D. M.) 75–94 (Dahlem University Press, Berlin, 2002).
277. Walker, W. *et al.* Defining Uncertainty. A Conceptual Basis for Uncertainty Management in Model-Based Decision Support. *Integrated Assessment* **4,** 5–17 (2003).
278. Walton, A. & Meidinger, D. Capturing expert knowledge for ecosystem mapping using Bayesian networks. *Canadian Journal of Forest Research* **36,** 3087–3103 (2006).
279. Wang, R., Li, F., Hu, D. & Li, B. Understanding eco-complexity: Social-Economic-Natural Complex Ecosystem approach. *Ecological Complexity* **8,** 15–29 (2011).

280. Wang, Y. & Zhang, X. A dynamic modelling approach to simulating socioeconomic effects on landscapes changes. *Ecological Modelling* **140 (1-2),** 141–162 (2001).
281. Warmink, J., Janssen, J., Booij, M. & Krol, M. Identificaion and classification of uncertainties in the application of environmental models. *Environmental Modelling & Software* **25,** 1518–1527 (2010).
282. Wauters, E. & Mathijs, E. An investigation into the socio-psychological determinants of farmers' conservation decisions: method and implications for policy, extension and research. *J.Agric.Educ.Ext.* **19,** 53–72 (2012).
283. Webb, G. I., Boughton, J. R. & Wang, Z. Not So Naive Bayes: Aggregating One-Dependence Estimators. *Machine Learning* **58,** 5–24 (2005).
284. Webster, K. L. & McLaughlin, J. W. Application of a Bayesian belief network for assessing the vulnerability of permafrost to thaw and implications for greenhouse gas production and climate feedback. *Environmental Science /& Policy* **38,** 28–44 (2014).
285. Willaarts, B. A. *Dinámica del paisaje en la Sierra Norte de Sevilla. Cambios funcionales e implicaciones en el suministro de servicios de los ecosistemas* PhD thesis (Facultad de Ciencias Experimentales. Departamento de Biología Vegetal y Ecología. Universidad de Almería, 2009).
286. Willaarts, B. A., Volk, M. & Aguilera, P. A. Assessing the ecosystem services supplied by freshwater flows in Mediterranean agroecosystems. *Agricultural Water Management* **105,** 21–31 (2012).
287. Witten, I. H. & Frank, E. *Data Mining: Practical Machine Learning Tools and Techniques (Second Edition)* (Morgan Kaufmann, 2005).
288. Wu, J. & Hobbs, R. Key issues and research priorities in landscape ecology: and idiosyncratic synthesis. *Landscape Ecology* **17,** 335–365 (2002).
289. Wu, X., Wen, X., Li, J. & Yao, L. A new dynamic Bayesian network approach for determining effective connectivity from fMRI data. *Neural Computing & Applications* **24,** 91–97 (2014).
290. You, L., Li, Y., Huang, G. & Zhang, J. Mdeling regional ecosystem development under uncertainty- A case study for New Binhai District of Tianjin. *Ecological Modelling* **288,** 127–142 (2014).
291. Young, J. W. S. A framework for the ultimate environmental index- putting atmospheric change into context with sustainability. *Environ. Monit. Assess.* **46,** 135–149 (1997).
292. Young, O. R. *et al.* The globalization of socio-ecological systems: An agenda for scientific research. *Global Environmental Change* **16,** 304–316 (2006).
293. Young, W. A. *et al.* Modeling net ecosystem metabolim with an artificial neural network and Bayesian belief network. *Environmental Modelling & Software* **26,** 1189–1210 (2011).
294. Zhang, N. L. & Poole, D. Exploiting causal independence in Bayesian network inference. *Journal of Artificial Intelligence Research* **5,** 301–328 (1996).
295. Zhang, Y., Qu, Y., Wan, J., Liang, S. & Liu, Y. Estimating leaf area index from MODIS and surface meteorological data using a dynamic Bayesian network. *Remote Sensing of Environment* **127,** 30–43 (2012).
296. Zhang, Z. & Dong, F. Fault detection and diagnosis for missing data systems with a three time-slice dynamic Bayesian network approach. *Chemometrics and Intelligent Laboratory Systems* **138,** 30–40 (2014).
297. Zou, M. & Conzen, S. D. A new Dynamic Bayesian Network (DBN) approach for identifiying gene regulatory networks from time course microarray data. *Bioinformatics* **21,** 71–79 (2005).