A comparison bewteen Elvira software and AMIDST toolbox in environmental data: A case study of flooding risk management

Rosa F. Ropero[1] **M. Julia Flores**[2] **Rafael Cabañas**[3] **Rafael Rumí**[1]

[1]*Data Analysis Research Group* , *Mathematics Dpt., University of Almeria*, Almeria, Spain
[2]*Computing Systems Dpt., SIMD I3A* , *University of Castilla-La Mancha*, Campus Univ. Albacete, Spain
[3] *Istituto Dalle Molle di Studi sull'Intelligenza Artificiale (IDSIA)* , Lugano, Switzerland

Abstract

Bayesian networks are extensively used in different research areas, environmental modelling in particular, because of their advantages. This has encouraged the development of several tools and software. In this paper, a comparison between Elvira software and AMIDST toolbox is made using data from a flood risk modelling example. Even when Elvira model presents better results, it is computationally inefficient in large datasets, which makes necessary to explore new and more powerful tools, like AMIDST, for environmental modelling tasks.

1 INTRODUCTION

Probabilistic graphical models (PGMs) and in particular Bayesian networks (BNs) have been extensively applied from the beginning of this century in different areas [2]. This has encouraged the development of tools and algorithms able to perform learning and validation steps in BNs.

BNs are considered as a powerful tool for environmental modelling [2]. A model of this kind is composed by a qualitative or visual part configured as a Direct Acyclic Graph that allows experts and stakeholders to visualize the model structure and the (in)dependence relations between the variables. This advantage has led to the development of a multiple methodologies based on stakeholders knowledge to learn model structures [6, 7]. Besides, BNs' quantitative part is based on the Probability Theory. When variables are discrete, results are summarized through conditional probability tables of each variables according to their parents in the graph. In contrast, when the variables are continuous, results are expressed in terms of probability density functions (e.g., Gaussian distribution). In this both cases, several metrics can be obtained, as the most probable value, the probability of reaching a specific threshold, among others [24].

In this context, many commercial and open source software tools have been developed by companies or academics for probabilistic modelling. Some well-known examples are: Agenarisk [12], BayesiaLab [8], SMILE-GeNIe [11], Hugin Expert [18], Netica [22] and Elvira [14]. These are basically modelling tools embedding inference libraries and typically supporting graphical interfaces (GUI). Besides, other relevant libraries are bnlearn in R [28], AMIDST [20], Crema [15] or InferPy [5, 9].

The aim of this work is to show a comparison in terms of applicability of two of the previous software tools, namely Elvira and AMIDST. This is done considering a case of study in environmental sciences: the prediction of the flooding risk in a river of southern Spain (*Guadalhorce*) based on historical data.

The present paper is organized as follows. Section 2 gives a brief description of the software tools considered; Section 3 describes the problem, the data available and the models applied; the results obtained are analysed in Section 4 whereas Section 5 provides the conclusions and possible lines of future research.

2 SOFTWARE DESCRIPTION

2.1 ELVIRA

Elvira [14] is a Java open source tool for building and evaluating PGMs. It was developed in the early 2000s through two joint research projects (Elvira, TIC97-1135-C04, and Elvira II TIC2001-2973-C05) involving several Spanish universities. Thus, it was conceived with the aim of providing to academic researchers a common platform for developing PGMs.

This software allows to define BNs, but also other common types of PGMs such as influence diagrams, decision trees, credal networks, hybrid object oriented Bayesian networks (OOBN) [16], etc. It implements most of the traditional methods for structure and parameter learning such as K2

Accepted for the 37th Conference on Uncertainty in Artificial Intelligence (UAI 2021).

algorithm and BIC metric. Besides, a wide variety of exact and approximate methods are available: variable elimination, importance sampling, junction tree propagation, etc. They are based on *Mixture of Truncated Exponential* model (MTE) which allows the variable range to be split into a several of intervals and approximated each of them by an exponential function. For more information see [21, 23, 26].

Elvira has been programmed with Java language, so that, it could be used on any operating system. It provides a Java API for accessing to its functionality, but it also has an easy-to-use GUI, which allows model definition and inference to users without programming skills. Moreover, this interface also allows to visualize the structure of the model, which can be important when using PGMs. For this reason, Elvira has been using during the last decades for applying BNs to a variety of fields such as medicine [10] and environmental sciences [1, 13, 24, 25]. The main drawback of Elvira is that it is a tool that was intended to be used in desktop computers, and hence it cannot be executed in parallel and distributed systems.

The source code and documentation is freely available at the official webpage [1]. Additionaly for a detail explanation of this software see [14].

2.2 AMIDST TOOLBOX

The AMIDST toolbox [20] is an open source Java software for scalable probabilistic machine learning with special focus on massive streaming data. It was developed in the context of a research project funded by the European Commission (FP7-ICT-619209) and including both academic and industrial partners. As a consequence, this software has been applied in the development of autonomous cars [30] and in the finance sector [4].

AMIDST supports the specification of PGMs over continuous and discrete domains with latent (or unobserved) variables. The key point of this toolbox is that it implements, among others, approximate Bayesian inference algorithms based on variational methods [3, 31]. This allows an efficient learning of the models which is suitable in cases where the whole data cannot be stored in memory or simply because it is not available (streaming data). In general, AMIDST can be used for classification, clustering, regression, density estimation tasks and inference, even in dynamic BNs.

The implemented algorithms for learning and making inference in PGMs can be executed in a parallel and distributed manner: AMIDST uses by default the Java 8 functional programming style and map-reduce operations for exploiting multi-core CPUs [19] . Additionally, AMIDST can run in a distributed computer cluster thanks to its integration with Flink[2] and Spark[3].

This toolbox is distributed using Maven, what simplifies the installation making the interaction with external software transparent. The source code is hosted on GitHub[4] and the documentation is available at the official website [5]. Moreover, at this website it is possible to find a comparison between AMIDST and other related software for learning and making inference in PGMs. Table 1 summarizes such comparison with respect to the Elvira software.

Table 1: Comparison between AMIDST and Elvira. Source: http://www.amidsttoolbox.com/documentation/

	AMIDST	Elvira
PGMs	Yes	Yes
Stationary data	Yes	Yes
Streaming data	Yes	No
Distributed processing in a cluster	Yes	No
Bayesian learning	Yes	No
Learning with latent variables	Yes	No
Conjugate exponential family	Yes	No
Open source	Yes	Yes

3 METHODS

3.1 STUDY AREA

Guadalhorce catchment is located in Málaga province, Andalusia, in the South of Spain (Figure 1). Historically, this area is well-known for its agricultural activity and a notable population. It is limited in the North by *Sierra de Archidona* mountainous range, in the East by the *Gibalto, San Jorge, Jobo* and *Camarolos* mountainous ranges, by *Sierra de las Nieves* mountainous range in the West, and Mediterranean sea in the South.

The irregular flow regime of *Guadalhorce* river, characterized by severe droughts and flash floods, has encouraged dam constructions in its middle course. So that, from the beginning of the 20th century, several hydrological infrastructures have been constructed in order to supply water, regulate water flow, provide electricity to the cities, and reduce damages provoked by flood and drought.

Climate in this area is Mediterranean. In terms of rainfall values, autumn and spring seasons are characterized by strong storms which can provoke serious damage in infrastructures, and also, humans well-being, mostly on the upper and middle part due to the steep relief.

[1] http://leo.ugr.es/~elvira

[2] https://flink.apache.org

[3] https://spark.apache.org

[4] https://github.com/amidst/toolbox

[5] http://www.amidsttoolbox.com

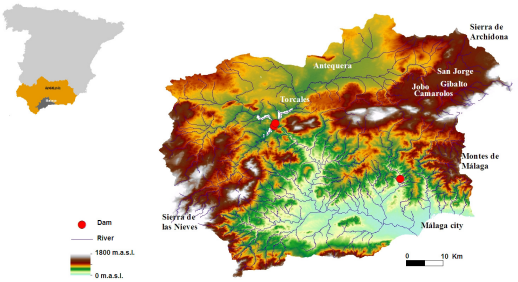

Figure 1: *Guadalhorce* catchment, its location, relief, and hydrographic systems. Dams are marked in red.

3.2 DATA COLLECTION

Data were collected per hour from October 2013 to September 2020 (both included) from the Hydrological Information Systems [6] (*Sistema Andaluz de Información Hidrológica*, SAIH). They were obtained from three different types, namely dams, meteorological and hydrological stations. Table 2 shows a summary of variables collected in each type of station used.

Table 2: Stations and variables collected in *Guadalhorce* catchment.

Station	Type	Variables collected
A130 A129 A128 A127 A104 A38 D34	Hydrological	Level, Rainfall
D126 P105 A46 P40 D33 D32	Meteorological	Rainfall
E31 E30	Dam	Level

Final dataset contains a total of 49025 observations over 15 variables. We have followed the division into hydrological years (from October to September), and divided into learning dataset, from October 2013 to September 2019, and validation dataset, from October 2019 to September 2020.

3.3 MODELS LEARNING AND VALIDATION

Herein, both models made by experts and learnt with Elvira software and AMIDST toolbox are deeply explained. Code and dataset are available in a GitHub repository [7].

3.3.1 Elvira model

The idea is to model, as accurately as possible, the river level at different points. So that, the risk of flooding can be estimated. Due to the complexity of the area, mode structure is based on a hybrid object-oriented Bayesian network (OOBN) [16]. A detailed information about model development can be found at [13]. Figure 2 shows the structure. This structure allows the catchment to be divided into five different units and models each of them independently what transforms a complex problem into a simple and easily interpretable model.

Next step consists on parameter estimation. It was done using the learning dataset (from October 2013 to September 2019) and an iterative least squares exponential regression methods [27].

Once the model is developed, a scenario is performed using dataset from October 2019 to September 2020. So that, information about rainfall is included as evidences and the river level values are achieved. Inference is based on Shenoy and Shafer algorithm [29]. Since results from MTE is expressed as a density function, we obtained the mean value from this probability distribution.

3.3.2 AMIDST model

The same problem can be modelled using the functionality available in AMIDST: in this case, we consider a Conditional Linear Gaussian (CLG) Bayesian network [17] defining the same structure than the OOBN depicted at Figure 2. All the variables are continuous and hence the parameters in model are Gaussian distributions whose means are linear combinations of the parents.

For learning the parameters, Streaming Variational Bayes (SVB) [3] algorithm with the same data was considered. This method allows to learn incrementally the parameters of a Bayesian network from a stream of data. Thus, the original dataset was divided into batches of 100 instances.

The validation of the model, the same testing data than in Elvira was used and the inference was made with the Variational Message Passing (VMP) algorithm [31].

[6] `http://www.redhidrosurmedioambiente.es/saih/presentacion`

[7] A link a Github will be provided.

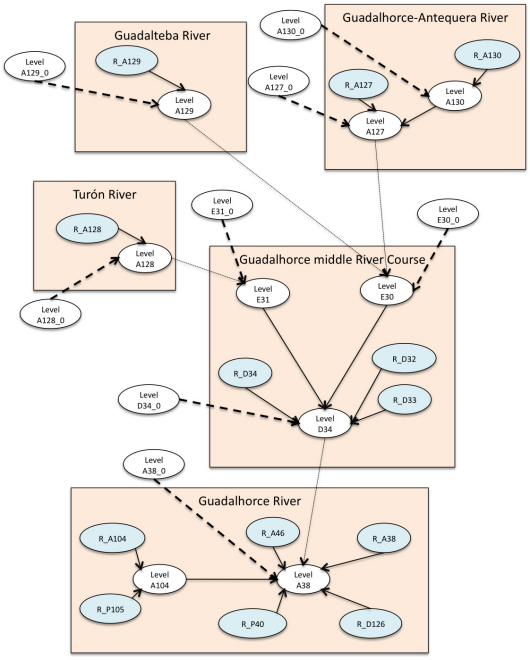

Figure 2: Model structure. Figure obtained from [13].

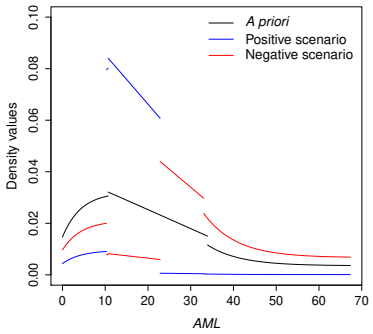

Figure 3: Example of MTE model's results. Figure obtained from [23].

4 RESULTS AND DISCUSSION

Using the same dataset for both parameter learning and inference process, Elvira and AMIDST models are compared.

Firstly, we should mark the different ways both tools achieve model parameter learning. Elvira software is based on MTE models, which means the variable range is divided in a set of intervals, each of them are approximated by a exponential function. An example can be seen in Figure 3, where data from *a priori* situation and under two different scenarios are compared. By contrast, AMIDST toolbox try to approximate data to the distribution that best fits them. In this case, all variables are approximated to a normal distribution.

In terms of computational cost, AMDIST is highly efficient in comparison with Elvira. For the same dataset, model was learnt by AMDIST in less than a minute, whilst model from Elvira took close to 40 hours [8] There are different reasons for this significant difference. First, the algorithm used was SVB, which is an efficient approximate method and its implementation in AMIDST is able to exploit the multi-core parallelism in the computer. Secondly, as this algorithm allows incremental learning, the dataset can be divided into batches. However, in Elvira the dataset must be completely loaded in memory, which makes the process more inefficient.

Figures 4 and 5 show the results of the inference process, comparing real data with those predicted by the models. Besides, Table 3 shows the root mean square error for each variable with models learnt with Elvira software and AMIDST toolbox. In general, error rates are higher in the AMIDST model for all variables. It could be explained by the fact that normal distribution are used in all variables. Very often environmental data does not follow a normal distribution, so, approximating to it can implies these higher errors. By contrast, MTE divides the range of the variable and is able to better fit to the data distribution. These differences imply that Elvira model is able to predict the moment a rainfall event takes place (Figures 5), which is not always the case of AMIDST (Figures 4).

Table 3: Root mean square error for each variable.

Variable	Elvira	AMIDST
D34	0.035	177.8
A38	0.027	12.8
A104	0.017	0.15
A127	0.097	0.39
A128	0.082	0.13
A129	0.0025	0.091
A130	0.039	0.29

Other point to consider is the applicability. Elvira software has been successfully applied into environmental studies, which means there is literature about how to use it. However, AMIDST is a really new tool that is still under development and just some applications have been published [4, 30], none of them in environmental studies. This makes that environmental data characteristics have not been totally considered yet. For example, the fact that environmental data hardly ever follows a normal distribution makes results obtained present higher errors.

In terms of interpretability, BNs have been extensively used in environmental modelling with the inclusion of expert knowledge. Thus, interpretability of the results are a key point. In this sense, Elvira software includes an interface

[8] A desktop computer with processor Intel Core i7 (2.8 GHz) and 8 GB of memory was used.

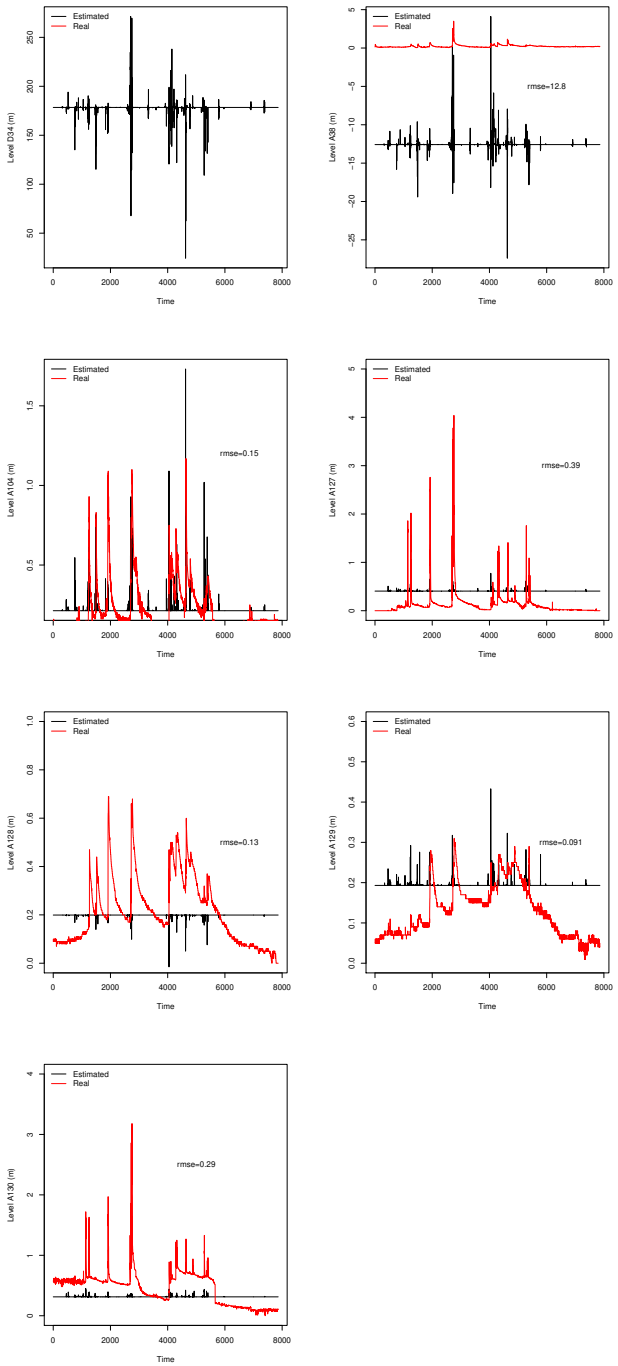

Figure 4: Real (red lines) and predicted (black lines) using AMIDST toolbox model.

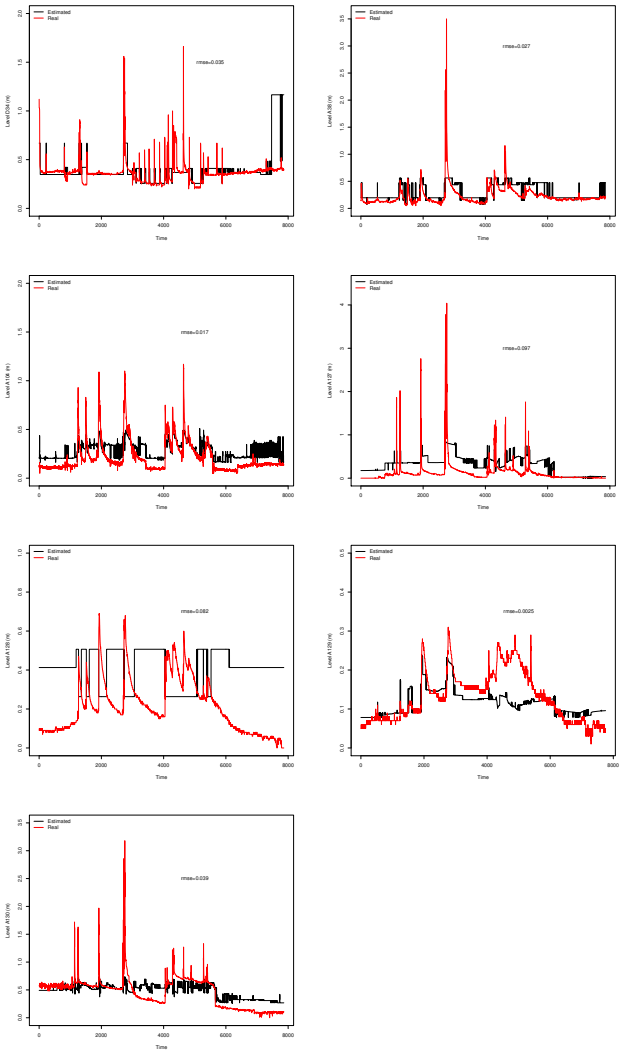

Figure 5: Real (red lines) and predicted (black lines) using Elvira software.

easy to use which allows the model learnt to be visualized. In contrast, AMIDST requires to have license for using HUGIN software to visualize the network made.

One advantage of both tools is the fact that they are published in open code, so it is possible to adapt the algorithms and methods to our own necessities. However, at the same time, it can be considered as a disadvantage since research groups need to include expert in machine learning and data mining able to deal with this programming language.

5 CONCLUSIONS

Bayesian networks are a versatile tool that have been applied in environmental modelling for the last decades. Their qualitative part allows experts to be included into the modelling process which means an advantage in the knowledge engineering area. Besides, its quantitative part has demonstrated to provide robust results in several environmental problems.

This has encouraged the development of different software and tools. In this paper, the aim was to compare between two open source tools, Elvira and AMIDST. As an example, data from flooding risk modelling were used.

In general, both models present a set of advantages. Elvira has been previously applied into environmental domains, so literature is easy to find. Besides, its easy-to-use interface helps expert to visualize the model learnt. Another advantage is the use of MTE models which split the range of variables into a set of intervals and approximate them with exponential functions, which best fit with the original data. In contrast, AMIDST is highly efficient in comparison, and perform model learning faster than Elvira.

In this paper the objective was not to decide which software is better or worst, just to compare them in terms of environmental modelling. Even when Elvira seems to present better results and more advantage *a priori*, is really computationally inefficient with large datasets. Thus, new and more powerful tools, like AMIDST, need to be deeply explored in environmental modelling tasks.

During the development of this work, some future aspects have been identified: *i)* modelling this problem with a dynamic or temporal component and compare, again, both tools, *ii)* compare their computational costs when dataset is larger, *iii)* include in AMDIST a set of latent variables with the aim of select the type of distribution that better fit with the data.

Author Contributions

Rosa F. Ropero built the Elvira model whilst Rafael Cabañas focused on the AMDIST one. Besides, M. Julia Flores created the code for AMIDST inference. Finally, Rafael Rumí supervised the study and the paper.

Acknowledgements

añana,

This study was supported by the Regional Government of Andalusia through project SAICMA (UAL18-TIC-A011-B-E); by the Spanish Ministry of Economy and Competitiveness through projects PID2019-106758GB-C31, C32/AEI/10.13039/501100011033 and C33, TIN2016-77902-C3-1-P and TIN2016-77902-C3-3-P.

References

[1] P. A. Aguilera, F. Reche, E. López, B. A. Willaarts, A. Castro, and M. F. Schmitz. Aplicación de las redes bayesianas a la caracterización del hábitat de la tortuga mora (testudo graeca graeca) en Andalucía. In *Proceedings of the I Congreso Nacional de Biodiversidad*, 2007.

[2] P. A. Aguilera, A. Fernández, R. Fernández, R. Rumí, and A. Salmerón. Bayesian networks in environmental modelling. *Environmental Modelling & Software*, 26: 1376–1388, 2011.

[3] T. Broderick, N. Boyd, A. Wibisono, A. C. Wilson, and M. I. Jordan. Streaming variational bayes. In *NIPS*, 2013.

[4] R. Cabañas, A. M. Martínez, A.R. Masegosa, D. Ramos-López, A. Samerón, T. D. Nielsen, H. Langseth, and A. L. Madsen. Financial data analysis with pgms using amidst. In *2016 IEEE 16th International Conference on Data Mining Workshops (ICDMW)*, pages 1284–1287. IEEE, 2016.

[5] R. Cabañas, A. R. Masegosa, and Salmerón A. Inferpy: Probabilistic modeling made easy. *Knowledge-Based Systems*, 2018. ISSN 0950-7051. doi: https://doi.org/10.1016/j.knosys.2018.12.030.

[6] M. J. Caley, R. A. O'Leary, R. Fisher, S. Low-Choy, and S. Johnson. What is an expert? a systems perspectie on expertise. *Ecology and Evolution*, pages 231–242, 2013.

[7] A. Castelletti and R. Soncini-Sessa. Bayesian networks and participatory modelling in water resource management. *Environmental Modelling & Software*, 22:1075–1088, 2007.

[8] S. Conrady and L. Jouffe. Introduction to bayesian networks & bayesialab. *Bayesia SAS*, 2013.

[9] J. Cózar, R. Cabañas, A. Salmerón, and A. R. Masegosa. Inferpy: Probabilistic modeling with

deep neural networks made easy. *Neurocomputing*, 415:408 – 410, 2020. ISSN 0925-2312. doi: https://doi.org/10.1016/j.neucom.2020.07.117. URL http://www.sciencedirect.com/science/article/pii/S092523122031328X.

[10] FJ Díez. Teaching probabilistic medical reasoning with the elvira software. *Yearbook of medical informatics*, 13(01):175–180, 2004.

[11] M. J. Druzdzel. Smile: Structural modeling, inference, and learning engine and genie: a development environment for graphical decision-theoretic models. In *Aaai/Iaai*, pages 902–903, 1999.

[12] N. Fenton and M. Neil. Decision support software for probabilistic risk assessment using bayesian networks. *IEEE software*, 2014.

[13] J. Flores, R. F. Ropero, and R. Rumí. Assessment of flood risk in mediterranean catchments: an approach based on bayesian networks. *Stochastic Environmental Research & Risk Assessment*, 33:1991–2005, 2019.

[14] JA Gámez, A Salmerón, et al. Elvira: An environment for creating and using probabilistic graphical models. In *Procs. of the First European Workshop on Probabilistic Graphical Models*, pages 222–230, 2002.

[15] D. Huber, R. Cabañas, A. Antonucci, and M. Zaffalon. Crema: A java library for credal network inference.

[16] D: Koller and A. Pfeffer. Object-oriented bayesian networks. In *Proceedings of the Thirteenth conference on Uncertainty in artificial intelligence*, pages 302–313, 1997.

[17] S. L. Lauritzen. *Graphical models*, volume 17. Clarendon Press, 1996.

[18] A. L. Madsen, M. Lang, U. B. Kjærulff, and F. Jensen. The hugin tool for learning bayesian networks. In *European conference on symbolic and quantitative approaches to reasoning and uncertainty*, pages 594–605. Springer, 2003.

[19] A. R. Masegosa, A. M. Martinez, and H. Borchani. Probabilistic graphical models on multi-core cpus using java 8. *IEEE Computational Intelligence Magazine*, 11(2):41–54, 2016.

[20] Andrés R Masegosa, Ana M Martinez, Darío Ramos-López, Rafael Cabañas, Antonio Salmerón, Helge Langseth, Thomas D Nielsen, and Anders L Madsen. AMIDST: A Java toolbox for scalable probabilistic machine learning. *Knowledge-Based Systems*, 163: 595–597, 2019.

[21] S. Moral, R. Rumí, and A. Salmerón. Mixtures of Truncated Exponentials in Hybrid Bayesian Networks. In *ECSQARU'01. Lecture Notes in Artificial Intelligence*, volume 2143, pages 156–167. Springer, 2001.

[22] Z. Ni, L. D. Phillips, and G. B. Hanna. Exploring bayesian belief networks using netica®. In *Evidence Synthesis in Healthcare*, pages 293–318. Springer, 2011.

[23] R. F. Ropero, P. A. Aguilera, A. Fernández, and R. Rumí. Regression using hybrid Bayesian networks: Modelling landscape-socioeconomy relationships. *Environmental Modelling & Software*, 57:127–137, 2014.

[24] R. F. Ropero, R. Rumí, and P. Aguilera. Modelling uncertainty in social-natural interactions. *Environmental Modelling & Software*, 75:362–372, 2016.

[25] R. F. Ropero, A. Maldonado, L. Uusitalo, A. Salmerón, R. Rumí, and P.A. Aguilera. Soft clustering approach to detect socio-ecological landscape boundaries using bayesian networks. *Agronomy*, 11:1–25, 2021.

[26] R. Rumí. *Modelos de redes bayesianas con variables discretas y continuas*. PhD thesis, Universidad de Almería, 2003.

[27] R. Rumí, A. Salmerón, and S. Moral. Estimating mixtures of truncated exponentials in hybrid Bayesian networks. *Test*, 15:397–421, 2006.

[28] M Scutari. Learning bayesian networks with the bnlearn r package. *Journal of Statistical Software*, 35(3), 2010.

[29] P. P. Shenoy and G. Shafer. Axioms for probability and belief functions propagation. In R.D. Shachter, T.S. Levitt, J.F. Lemmer, and L.N. Kanal, editors, *Uncertainty in Artificial Intelligence, 4*, pages 169–198. North Holland, Amsterdam, 1990.

[30] G: Weidl, A. L. Madsen, V. Tereshchenko, D. Kasper, and G. Breuel. Early recognition of maneuvers in highway traffic. In *European Conference on Symbolic and Quantitative Approaches to Reasoning and Uncertainty*, pages 529–540. Springer, 2015.

[31] J. Winn, C. M. Bishop, and T. Jaakkola. Variational message passing. *Journal of Machine Learning Research*, 6(4), 2005.