Application of Bayesian networks for risk assessment in Spanish and Mexican recreational water bodies for planktonic and benthic cyanobacterial proliferations

Kristen Havrilla

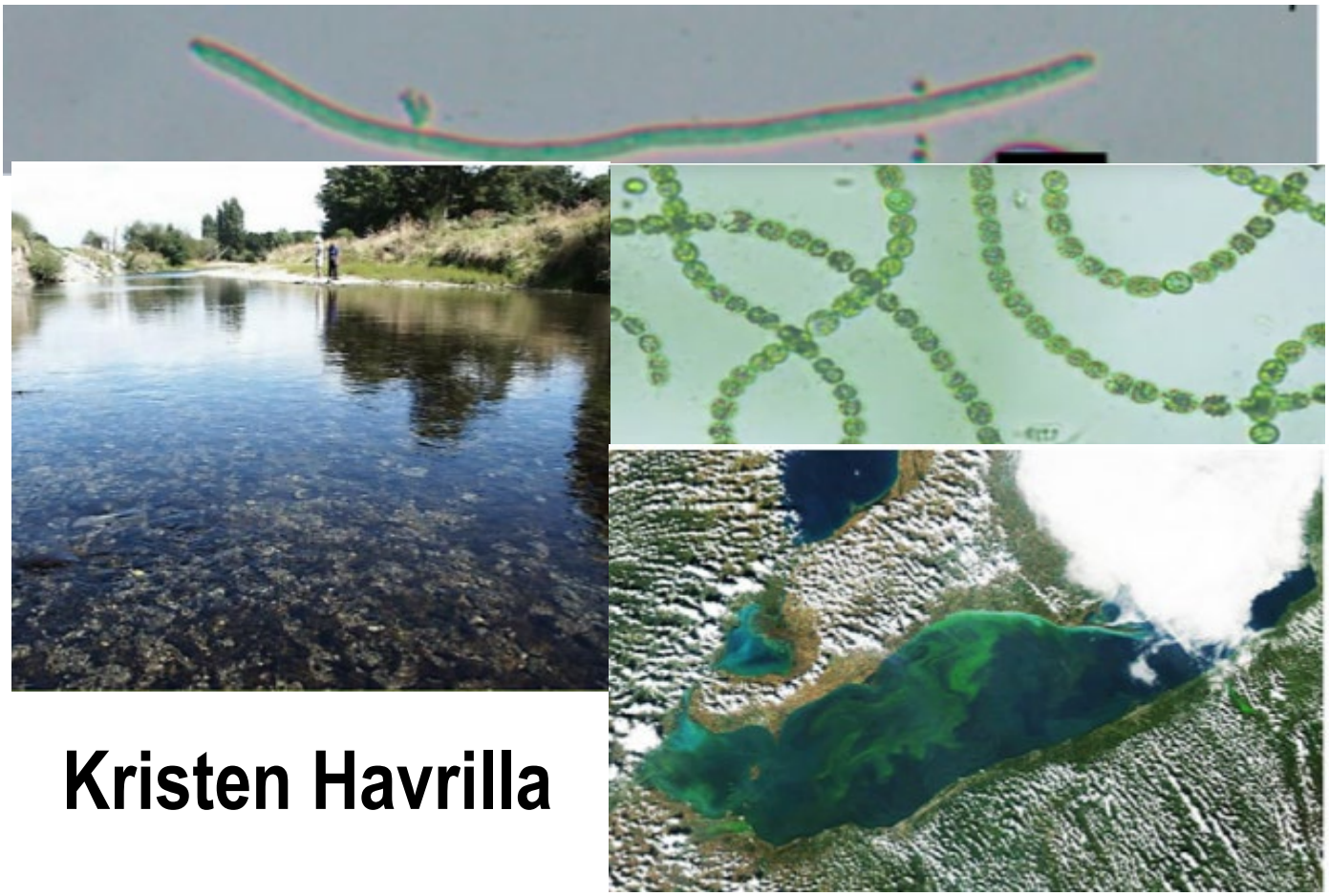
Máster en Calidad de Aguas Continentales

MÁSTERES
DE LA UAM
2019 - 2020

Facultad de Ciencias

UAM Universidad Autónoma de Madrid

Application of Bayesian networks for risk assessment in Spanish and Mexican recreational water bodies for planktonic and benthic cyanobacterial proliferations



Kristen Havrilla

Trabajo Fin de Máster. Curso 2019-2020

Director: Antonio Quesada del Corral
Tutores: Samuel Cirés and Elvira Perona
Lugar de realización: Universidad Autónoma de Madrid

Table of Contents

List of Abbreviations

Abstract

List of Abbreviations

ANA - Anatoxin-a

BWD - Bathing Water Directive

BN - Bayesian Network

CCI - Correctly Classified Instances

Chl*a*Cyano - Chlorophyll *a* from cyanobacteria

Chl*a*Total - Total Chlorophyll *a*

CPT - Conditional Probability Table

CV - Current Velocity

CyanoHAB - harmful cyanobacterial bloom

CYN - Cylindrospermopsin

DAG - Directed Acyclic Graph

DIN - Dissolved Inorganic Nitrogen

DO - Dissolved Oxygen

ELISA - Enzyme Linked ImmunoSorbent Assay

EPA - Environmental Protection Agency

Film % - Biofilm cover %

LPS – Lipopolysaccharides

Maxdom - Maximum Dominance of Cyanobacteria

MC – Microcystins

PCA – Principal Component Analysis

Presence of Toxic - Presence of Potentially Toxic Genera

RD – Royal Decree 1340/2007

SRP - Soluble Reactive Phosphorus

TDI - Tolerable Daily Intake

UAM – Universidad Autónoma de Madrid

Abstract

As the climate changes and nutrient overloading from anthropogenic activities increases, the abundance of harmful cyanobacterial blooms increases with it. These blooms can cause significant challenges to water management such as serious threats to human and wildlife health, impinging recreational uses, and impacting the aesthetics of the ecosystem. The most common cyanotoxins found in lentic and lotic systems are microcystsins (MC) and anatoxin-a (ANA) respectively. The presence of these cyanotoxins has been tentatively linked to various environmental stressors. Identifying the most influential biotic and abiotic conditions on the concentration of these cyanotoxins can help determine acceptable thresholds that correspond with benchmark values outlined in legislation for water quality. A statistical Bayesian modelling approach was taken to analyze the cause and effect relationships of the most important environmental variables with the exceedance of MC and ANA thresholds for low risk. Three independent models were developed in total, one for each of the case studies: Spanish lentic systems (using 88 samples from 76 reservoirs and lakes), Mexican lentic systems (using 65 samples from nine reservoirs and lakes), and Spanish lotic systems (46 data samples from *Phormidium* mats located in 10 rivers). Using a correlation matrix to identify strong linear and nonlinear relationships, with a validation of each of the Bayesian networks through sensitivity analyses, core parameters that will guide water managers to predict MC and ANA levels were discovered. Results propose the most important parameters for increased probability prediction of concentration and presence of toxic cyanobacteria in each case study were: total chlorophyll *a* (Chl*a*Total), chlorophyll *a* from cyanobacteria (Chl*a*Cyano), and dominance of cyanobacteria in a bloom for Spanish lentic systems; dissolved oxygen, conductivity, and temperature for Mexican lentic systems; and DIN, SRP, pH, and conductivity for Spanish lotic systems. Complying with the estimated thresholds outlined for each system, cyanotoxin risk can be effectively suppressed. The incorporation of probabilistic Bayesian modelling offers parsimonious solutions in a functional way that accommodates a certain level uncertainty for water managers to determine the likelihood of surpassing acceptable levels of risk.

Key words: Cyanobacteria, Harmful algal blooms, Bayesian network modelling, Microcystin, Anatoxin-a, Risk assessment, Water management

Acknowledgments

Spanish lentic system Bayesian network modelling was made possible by data supplied by Universidad Autónoma de Madrid for a national project with the Spanish Ministerio de Medio Ambiente y Medio Rural y Marino. Sampling was carried out by Tragsatec. Spanish lotic system Bayesian network modeling was supported by data supplied by Universidad Autónoma de Madrid, where the sampling and analyses were completed under the support of grant CGL2013-44870-R from MINECO and Cyted 2019 TALGENTOX-P919PTE0047. Also, I am grateful to Samuel Cirés and Elvira Perona for their advice and support during these difficult times. Lastly, I would also like to thank my partner Alastair Oates for his constant support and encouragement which made this project possible.

1. Introduction

1.1 Cyanobacteria and Cyanotoxins

Eutrophication (i.e. increased input of nutrients, largely phosphorus but also including nitrogen) is a constant daunting problem that is plaguing freshwater sources as a consequence of an overwhelming number of factors. These factors range from excessive nutrient runoff from fertilizers, agriculture wastes, and stormwater runoff, being additionally exacerbated by direct and indirect effects of climate change such as increased water temperature or prolonged water stability (Ibelings *et al.,* 2014). Physical pollution also can play a part, as disruption of water continuity, diversions, or salinization cause an increase in residence time and vertical stratification which enhances the formation of a harmful cyanobacterial bloom (cyanoHAB), i.e. high population densities. Figure 1 shows examples worldwide. Cyanobacteria can be found in all-natural ecosystems including soil, bare rock, freshwater, oceans, brackish water, estuarine salt lakes, salt marshes, rivers, etc. (Niamien-Ebrottie *et al.,* 2015). The increased eutrophication of the trophic state normally favors the exponential growth of cyanobacteria, as most species are adapted to hot temperatures, stratification (due to the gas vesicles which incite buoyancy), and are salt-tolerant (Paerl and Otten, 2013). These ecophysiological abilities allow these bacteria to ubiquitously exploit anthropogenic modifications to ecosystems.

Cyanobacteria are photosynthetic prokaryotes that can be found in almost every aquatic ecosystem across the globe due to their ability to adapt to a plethora of climatic and geochemical changes. This ability has evolved since their first appearance around 3.5 billion years ago. These microorganisms are thought to be the first oxygenic photoautotrophs present on the Earth that lead to a biosphere that included oxygen, paving the way for the atmosphere today. Although they are usually not visible without the use of a microscope, cyanobacteria can exist as free-living cells, as colonies, or as filaments (Quiblier *et. al*, 2013). Moreover, since these prokaryotes are ever-present in ecosystems, they can grow planktonically (in the water column, primarily in lentic systems), metaphytically (piled on the water surface), epiphytically (attached to macrophytes, other algae, or even other cyanobacteria), or benthically (attached to the sediment of the bottom of systems) (Figure 1b and 1c) (Quiblier *et. al* 2013). Although cyanoHAB are found in a large range of areas, traditionally planktonic cyanobacteria bloom formation and characterization has been given more scientific attention. However, under certain environmental conditions, cyanobacteria can proliferate massively, initiating the so-called blooms, which can reach concentrations of several thousands of cells per mL^{-1} and accumulate on the water surface in the form of scums or as benthic mats covering most of the river beds. Planktonic proliferations can cover areas of several km^2 making them even visible from satellites (Figure 1a and 1d).

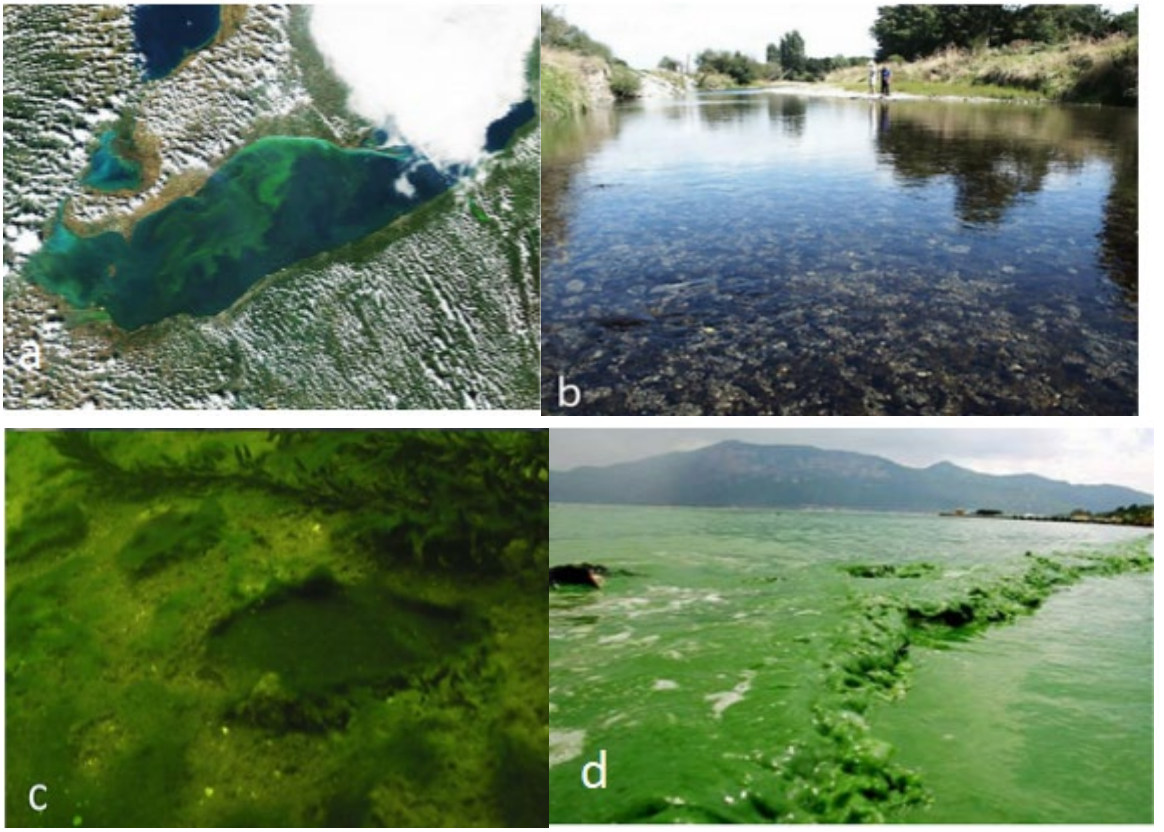
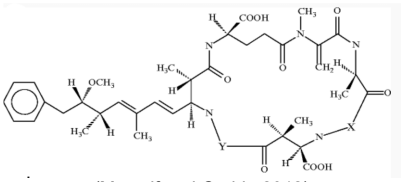
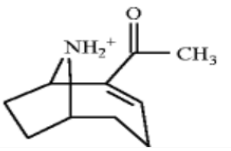
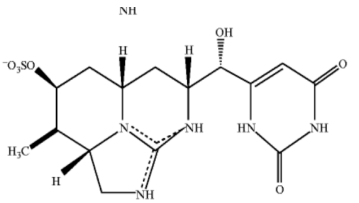
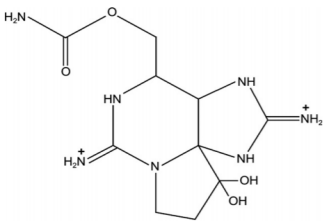
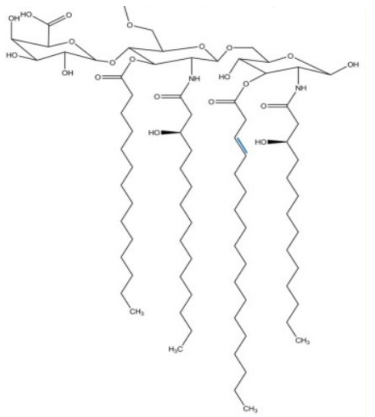


Figure 1 - **a.** CyanoHAB satellite image in Lake Erie(Pennsylvania, United States) (Paerl and Paul, 2012) **b.** *Phormidium* Benthic cyanoHAB in Waipoua River (North Island, New Zealand) (Quiblier *et al.,* 2013) **c.** Benthic mat in Lake Rotoiti, New Zealand (Quiblier *et al.,* 2013) **d.** CyanoHAB on Lake Dianchi, Yunan Province, China (Paerl and Paul, 2012)

According to Paerl and Otten (2013), along with decreased ecosystem productivity and species richness, cyanoHAB can also produce toxic secondary metabolites (the so-called cyanotoxins) which can cause serious adverse health effects in mammals, and to a lesser extent aquatic biota. These cyanotoxins have been reported in at least 66 countries worldwide with a range of 25-75% of the blooms being considered toxic (Bláha *et al.,* 2009; Ibelings *et al.,* 2014). Although the purpose and function of these cyanotoxins are unclear, it is imperative to analyze the dangers presented. This can be carried out by understanding their chemical and physical properties. The cyanobacterial toxins can be sorted into four broad categories for the toxic effects on the body that include hepatotoxic (affecting the liver), neurotoxic (affecting the nervous system), dermatotoxic (affecting the skin), or cytotoxic (which is the general inhibition of protein synthesis)(WHO, 1999). Compounding with these categories, cyanotoxins can also fall into groups based on chemical structure. Although there can be hundreds of different forms of the compounds (more than 300 have been described up to date (Cirés *et al.,* 2013), cyanotoxins can generally fall into the classification of cyclic peptides, alkaloids, and lipopolysaccharides (LPS) (WHO, 1999). Table 1 outlines the cyanotoxin groups, as well the most common genera that produce the toxin. To date, there have been about 40 genera of cyanobacteria that have been described as cyanotoxin producers (Bernard *et al.,* 2017).

Table 1. General characteristics of cyanotoxins (World Health Organization, 1999)

Toxin Group	**Most targeted organ in mammals**	**Cyanobacterial genera**	**General chemical structure**
		Cyclic peptides	
Microcystins	Liver, inhibit protein phosphatase	*Microcystis, Anabaena, Planktothrix, Nostoc, Hapalosiphon, Anabaenopsis*	 (Metcalf and Codd., 2012)
		Alkaloids	
Anatoxin-a	Nerve synapse, lungs	*Anabaena, Planktothrix, Aphanizomenon, Phormidium*	 (Metcalf and Codd., 2012)
Cylindrospermopsins	Liver, inhibits protein synthesis	*Cylindrospermopsis, Aphanizomenon, Umezakia*	 (Metcalf and Codd, 2012)
Saxitoxins	Nerve axons	*Anabaena, Aphanizomenon, Lyngbya, Cylindrospermopsis*	 (Solter and Beasley, 2013)
Lipopolysaccharides (LPS)	Affects any exposed tissue	All	 (Durai *et al.*, 2015)

The cyanobacterial toxins most commonly found in blooms are of the microcystin (MC) family (and therefore the most common species included in legislation worldwide) (Svirčev *et al.,* 2019). MCs are cyclic heptapeptides containing two variable amino acids and the unusual cyclic amino acid 3-amino-9-methoxy-2,6,8-trimethyl-10-phenyl-4,6-decadienoic acid (normally known as ADDA). Due to the different possible amino acids of the two variable positions, more than 140 variants of MCs have been discovered (Meriluoto *et al.,* 2017). In mouse bioassays, it was found that these toxins can cause death by liver hemorrhage only a few hours after acute ingestion due to the strong binding to the essential cellular enzyme: protein phosphatases (WHO, 1999).

Another common cyanobacterial toxin that is found in North America, Europe, and Australia is called anatoxin-a (ANA) which is a neurotoxic alkaloid. These can cause death quickly, usually between two and 30 minutes, by causing respiratory failure. Alkaloids are considered very diverse in their chemical structures and in their toxicities, but generally are classified by heterocyclic nitrogenous compounds that contain ring structures with at least one carbon-nitrogen bond (WHO, 1999). The World Health Organization (WHO) also claims that these are found to be common in rivers. The last common cyanotoxin that is quoted in legislation and guidelines is cylindrospermopsin (CYN). Initially CYN was considered to be a toxin mostly affecting tropical waters in Australia and Asia, but during the last two decades, it has been found in several other regions including temperate areas of Europe (including Spain) and North America. In fact, CYN is considered thus far as the second most widespread cyanotoxin group after MCs, as CYN is cited in 10% of cyanotoxin records worldwide (Svirčev *et al.,* 2019). It is a cytotoxic alkaloid which normally affects the liver, but in some variations of the structure can also affect the kidneys, spleen, thymus, or heart (Niamien-Ebrottie *et al.,* 2015).

These cyanotoxins can be either extracellular, which are released during bloom decay during the lyse of the cells, or intracellular, which are consumed by zooplankton and fish and can be bioaccumulated in some species (Pawlik-Skowrońska *et al.,* 2013). There are a few exceptions of cyanotoxins (e.g. Cylindrospermopsins) which may present active release. Common cyanotoxins in legislation include ANA and MC variants which are stored intracellularly and are usually discovered during the growth stage of the toxic blooms. These can be easier to remove from the water column than extracellular toxins due to the water-soluble properties of these compounds, and the fact that latter can be adsorbed by clays and organic material (Niamien-Ebrottie *et al.,* 2015). According to Munoz *et al. (*2019), 10-95% of all recorded cyanobacterial blooms contain cyanotoxins, emphasizing the need for proper risk management. Methods for cyanotoxin removal include chlorination, adsorption, ozonation, photocatalysis, and advanced oxidation processes. The efficiency of these processes depends on the type of toxin and the physicochemical parameters of the source water (Munoz *et al.,* 2019).

To further complicate the issue, additional knowledge gaps are present when it comes to records and data from developing countries. Although cyanotoxins are known to be present, there is barely any legislation on risk assessment or management and there are difficulties in creating monitoring programs or implementing preventive measures (Pírez *et al.,* 2012). Specifically, publications in South America are few and far between. CyanoHABs can grow out of control in places where there is uncontrolled erosion or untreated sewage that is discharged directly into the bodies of water. Furthermore, detection methods of *Microcystis* by microscope are labor intensive

and do not directly indicate toxicity. To determine if the cyanotoxin is present by HPLC or mass spectrometry (MS) are costly, making large amounts of sampling for routine monitoring unfeasible. The difficulty continues to be compounded by the fact that there are more than 240 chemical congeners of the MC family with the predominant local variants not yet being identified (Meriluoto *et al.,* 2017). Luckily, cost effective immunoassays are becoming increasingly accepted by the academic community for rapid screening of a large number of samples. Enzyme-Linked ImmunoSorbent Assay (ELISA) is an established analysis that is accepted by several regulatory agencies, including the U.S. Environmental Protection Agency (EPA). Pírez *et al.,* (2012) suggests that this tool is suitable for supplying global information on toxicity, though with the limitation that ELISA is not able to distinguish the relative proportion of each MC variant.

Due to the various limitations of MC analyses, such as their high cost and low availability of technology in some water management laboratories, it is vital that risk assessment networks include low cost parameters based on biomass that are easier to measure (such as chlorophyll *a,* hereafter referred to as Chl*a*Total) or physicochemical characteristics that can moderately help predict and anticipate blooms before cyanotoxins reach dangerous levels.

1.2 Physicochemical Variables That Affect Formation of CyanoHAB

Due to the dangers of cyanoHABs (loss of water clarity, oxygen depletion, and cyanotoxins), the physicochemical parameters that contribute to the success of common toxin producing cyanobacteria are necessary to be analyzed. Cyanobacteria can thrive and exploit both nutrient-scarce and nutrient rich terrestrial and aquatic environments around the world due to ecophysiological abilities such as the ability to fix atmospheric nitrogen (an anaerobic process), store phosphorus, iron, and other trace elements, buoyancy regulation and the formation of akinetes (resting spores) (Mantzouki *et al.,* 2018, Paerl and Otten, 2013). An example would be *Microcystis aeruginosa* which can surpass stratified layers of enhanced water column stability, as well as using the buoyancy regulation to sink down to access nutrients at deeper layers, then floating up to obtain light from the surface (Mantzouki *et al.,* 2018).

As human population density, agricultural, and industrial activities increase, the nutrient loading rates are rapidly increasing in freshwater systems (Davis *et al.,* 2009). Among these nutrients, phosphorus has routinely been quoted as the main limiting nutrient for primary producers' productivity and algal biomass in freshwater systems, and nitrogen as the catalyst for new production in marine environments. Nutrient enriched water bodies are more susceptible to cyanoHABs if they have low flushing out rates (long residence times), higher water temperatures (periodically over 20 °C), calm surface waters (absence of wind), or have prolonged, persistent vertical stratification (Paerl and Otten, 2013). Moreover, phosphorus enrichment is thought to play a larger part in the development of cyanoHAB, more so with nitrogen fixing cyanobacteria genera (*Anabaena, Aphanizomenon, Cylindrospermopsis, Lyngbya, Nodularia, Oscillatoria, Trichodesmium, etc.)* that supply their own nitrogen by converting atmospheric N_2 to biologically available ammonia (NH_3) (Paerl and Otten, 2013).

Temperature plays a large role in the development of cyanoHAB as well. As there is an increase in atmospheric carbon dioxide from the burning of fossil fuels and deforestation, it is

expected that there will be a 2-5ºC increase of global temperatures (Davis *et al.,* 2009). Usually, cyanobacteria will dominate phytoplanktonic during the warmest times of the year, especially in eutrophic ecosystems in temperate climates. In particular, *Microcystis* has been found to have the highest rate of growth and photosynthesis at their optimal temperature of/or above 25ºC (Davis *et al.,* 2009), along with an increase of cellular toxin content of multiple other genera of cyanobacteria at these temperatures. Furthermore, temperature can directly (cyanobacteria can outgrow other algal groups at higher temperatures) and indirectly alter environmental factors such as conductivity and dissolved oxygen, in addition to co-occuring with prolonged thermal stratification (Huber *et al.,* 2011).

As the water quality of an ecosystem deteriorates, cyanobacteria can use the aforementioned adaptations to outcompete other algae (Huber *et al.,* 2011). The environmental factors that influence water quality and trophic state are listed as: conductivity (measures concentration of ions in the water), Chl*a*Total (proxy for algal biomass), dissolved oxygen (in anoxic systems there will be relatively more NH_4) , and pH. According to Rahman and Jewel, 2008, alkaline pH (around 8.8), low dissolved oxygen concentration, and comparatively higher concentration of nutrients (including NH_4) were observed during the cyanoHABs. To a certain extent, toxin levels might respond to environmental conditions, meaning the toxin content per cell can vary several fold, with the proportion of different congeners changing with the changes in the environment (Ibelings *et al.,* 2014).

1.3 Risk Assessment and Legislation

As previously stated, cyanobacteria can occur in both aquatic and terrestrial environments with the potential to create toxins that are harmful to human and animal health. Svirčev *et al.,* (2019) states that major cyanotoxins have been found in 66 countries worldwide (with 1118 recorded identifications in 869 freshwater systems). Of these cyanotoxins, the most common were MC, which accounted for 63% of the records globally, CYN: 10% of records, ANA: 9%, and saxitoxins: 8%. Due to the increasing awareness of the ubiquitous and dangerous nature of these cyanotoxins, countries are addressing cyanotoxin hazards and implementing regulatory approaches for their respective conditions (Ibelings *et al.,* 2014). Regulations and guidelines up to date cover the three vehicles of possible oral exposure: ingestion of cyanotoxins through drinking water, recreation use, or consumption of seafood from freshwater bodies (only Australia, Denmark, France, and the United States of America). The literature on drinking water legislation is much more substantial than that of the other routes of exposure, especially recreation, probably owing to the fact that it is a necessity of life. However, in countries with water that is treated well, recreation can be the major exposure route (Ibelings *et al.,* 2014).

Recreational water risk assessment is usually defined by many countries as a two or three tier alert level framework, although there is a discrepancy between countries in terms of legislation and scientific literature, specifically a lack thereof in some countries from South/Central America and Africa (Ibelings *et al.,* 2014). These tiers normally are comprised of a low level ‘Surveillance mode’ where authorities continue monitoring, a moderate level of ‘Alert mode’ which is usually the presence of an indicator of cyanoHABs (such as MC or cyano-chlorophyll *a*, referred to Chl*a*Cyano hereafter) where the public is warned, and finally the highest level of ‘Action mode’ which is based

on the presence of persistent scums or surpassing the last tier in the decision tree. An example of a decision tree for distinguishing the probability of cyanoHAB proliferation in Spain is outlined in Figure 2. Countries such as Australia have an additional two levels of 'Action mode' dealing with the likelihood of adverse health effects. Other countries like Canada have only a single guideline value of MC for recreational waters, or in Singapore which uses a single value of Chl*a*Total (being >50 µg/L) for 'Action mode (Ibelings *et al.,* 2014). The United States of America, for example, does not have a national guideline but rather depends on the state with only 21 out of the 50 states implementing a guidance value for recreational waters (Ibelings *et al.,* 2014). It should be noted that in literature and legislation, either the general term of cyanobacteria is used, or the specific cyanotoxin of MC. As aforementioned, MC and CYN are the most commonly found cyanobacteria in lentic systems and ANA being most commonly found in lotic systems (benthic). For this reason, these cyanotoxins can be found in decision trees such as that for Spain.

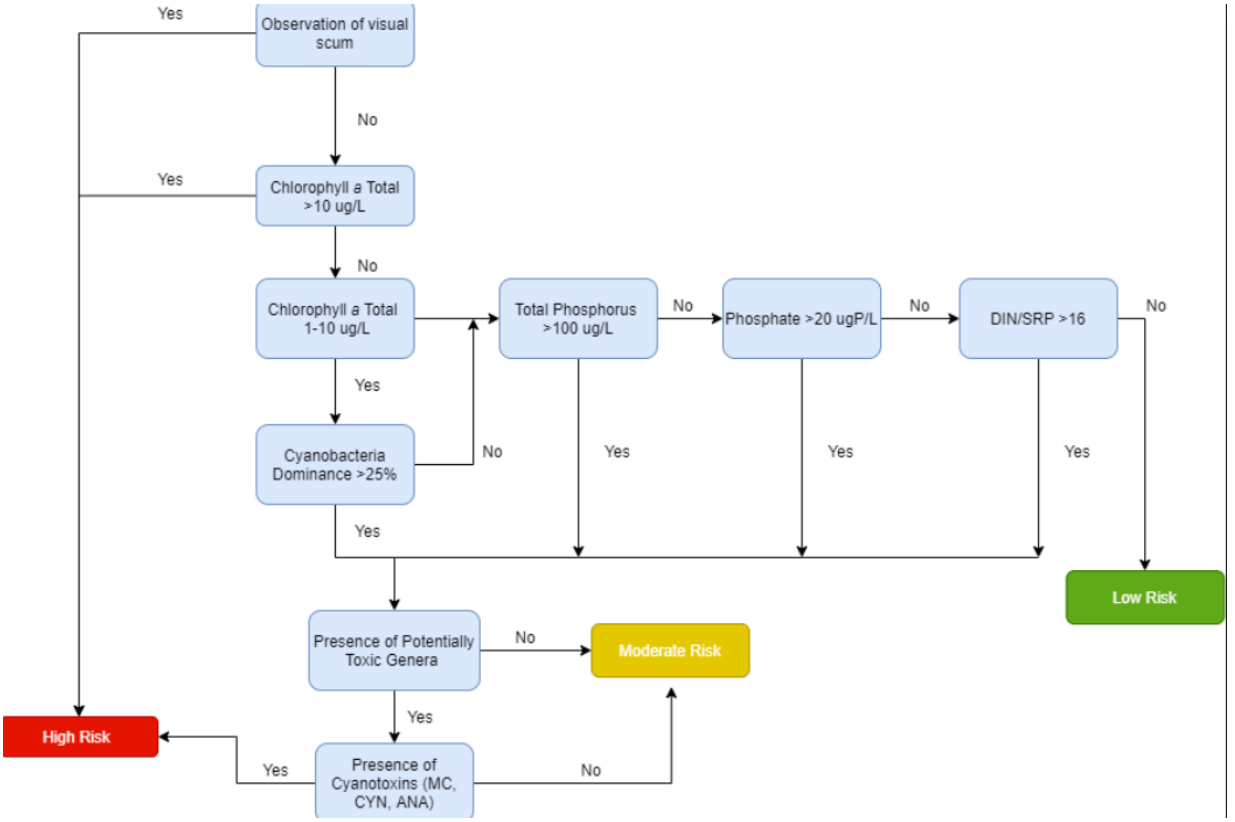


Figure 2 outlines the current decision-making tree for risk of proliferation of cyanoHAB in Spanish recreational waters (Wörmer, *et al.,* 2011)

For members of the European Union, the Bathing Water Directive 2006/7/EC (BWD) provides the foundation for classification guidance values of bathing sites in relation to the probability of cyanobacterial proliferation (low probability, medium probability, high probability). The risk-assessment framework used calls for authorities to analyze potentially contaminating conditions in a timely manner (Ibelings *et al.,* 2014). Due to the general terms of the BWD, many countries have adopted thresholds defined by a general framework of guidelines set by the WHO that is used internationally, either directly incorporated such as in Czech Republic, France, Japan, Korea, New Zealand, Norway, Poland, Brazil and Spain, or translated into different thresholds depending on the country (Australia, Canada) (Burch, 2008). For recreational waters, three potential routes of

exposure are outlined: direct contact, accidental swallowing, and inhalation of water. The hazards of cyanotoxin uptake, specifically MC which is the cyclic peptide cyanotoxin outlined by WHO due to the ubiquitous nature of this cyanobacterial toxin and also because it is widely regarded as the most serious potential source of human injury globally (Burch, 2008), is directly related to the level of toxins in the water and how much of that water is ingested. Therefore a safe range needs to be derived for guidelines of safe use of bathing sites (Table 2), but it should be borne in mind that these guidelines have not been adjusted for benthic populations and therefore cannot be used for data for rivers (WHO, 1999). In this project, MC concentrations were adapted from the WHO guidelines for risk classification for lentic waters.

To assess safe water quality for recreational waters, drinking water guidelines of Tolerable Daily Intake (TDI) can be applied due to the unavoidable ingestion of water (approximately 100-200 mL of water in one session, with water-sport athletes probably ingesting more). TDI represents the dose level in humans that when taken daily over a lifetime will result in no adverse effects. For example, the most quoted guideline value globally is the threshold of 1.0 μg L^{-1} for drinking water which is much lower than the thresholds listed for recreational waters (WHO, 1999). This is due to the fact that drinking waters are associated with chronic injuries (from low doses being ingested daily) whereas recreational waters are more an acute risk (high doses with short exposure times). There are many variables that can lead to a higher acute risk than is outlined, such as a child with less body weight ingesting more water than estimated, wind sweeping scums to shorelines, or benthic mats breaking off after a storm.

In addition, risk assessment has commonly been performed only for these planktonic organisms despite benthic cyanobacteria being responsible for multiple animal deaths (Quiblier *et al.,* 2013, Wood *et al.,* 2007). Legislation and guidelines are primarily developed for planktonic species in reservoirs and lakes for recreational and drinking water supply use with limited information on sampling, monitoring and managing benthic cyanobacteria, even with dangerous mats being found in reservoirs and lakes (Uriza *et al.,* 2017). According to Quiblier *et al.,* (2013), only two countries have established guidelines for risk and management of these species: Cuba and New Zealand, although they are based on preliminary research. Both frameworks incorporate a percentage of coverage of potentially toxic cyanobacteria that is attached to the substrate. Data related to the growth of benthic species and toxin production is scarce, leaving a large knowledge gap that is incomplete.

Table 2. Range of guidelines outlined by World Health Organization, 2003

Probability of Adverse Health Effects	Health Effects	Guideline Values	Recommended Advisory
Mild/Low	Irritation or allergic effects	20,000 cyanobacterial cells mL^{-1} <10 µg L^{-1} chlorophyll *a* (with cyanobacterial dominance) 2-10 µg L^{-1} of MC	Provide visitors with information about low level risk Authorities are informed to initiate further surveillance
Moderate	Irritation, may cause health impact	<100,000 cyanobacterial cells mL^{-1} <50 µg L^{-1} chlorophyll *a* (with cyanobacterial dominance) 10 - 20 µg L^{-1} of MC	Intensification of surveillance Daily inspection for scum formation Intervention and restriction of bathing
High	Severe health hazards	<10,000,000 cyanobacterial cells mL^{-1} <5,000 µg L^{-1} chlorophyll *a* <2,000 µg L^{-1} of MC Appearance of cyanobacterial scum formation	Possible closing of the bathing site/inform public Continued intense surveillance Risk Management/inform public Public health follow-up investigation

1.3.1 Spanish Recreational Waters

Legislation for recreational water and drinking water differ, with drinking water being outlined in the Royal Decree 140/2003 which integrates the European Directive 98/83/CE into Spanish legislation. Recreational waters have a national implementation of the European Union BWD for water quality that has been translated into the Royal Decree (RD) (1341/2007), without much amending. In article 6 of the RD, it is outlined that cyanobacteria risk is to be analyzed and managed, corresponding to Annex III 1c of the BWD which establishes a profile for bathing waters. One of the parameters that is analyzed in these defined profiles is the “propension to cyanobacterial proliferation” (BWD, 2009). For this, Universidad Autónoma de Madrid in collaboration with the Spanish Ministry of Environment created a decision tree (Figure 2) to qualify bathing waters’ risk of developing a cyanoHAB.

As concentration of cyanotoxin depends on the concentration of cyanobacterial biomass, the formation of scum (accumulation of floating cyanobacteria metaphytically) can increase toxin levels by orders of magnitude. Therefore, in the Spanish risk assessment decision tree, the observation of scum or mats (associated with benthic populations) is the first-tier level, along with the WHO parameter of Chl*a*Total. If there is presence of scum, or if the Chl*a*Total value is over 10, then the bathing site is automatically labelled high risk. Although this is considered moderate by WHO, the

value was set with conservative measures in mind relating to the multitude of outside variables aforementioned. The second tier involves variables such as total Phosphorus (>100 µg/L), phosphate (>20 µg P/L), the molar ratio of dissolved inorganic nitrogen (DIN)/ total soluble reactive phosphorus (SRP) (>16), and the percentage of cyanobacterial dominance (>25%). If these thresholds are not reached then the risk is considered low, if any of the thresholds are reached it then moves to the third tier which includes the presence of toxic genera (which if not present is moderate risk), if present then the presence of cyanotoxins (MC, ANA, or CYN) is analyzed. Ultimately, if these toxins are not present the risk should be considered moderate, and if present then it is considered high risk (Wörmer *et al,* 2011). This decision-making tree is outlined in Figure 2.

As stated before, information regarding MCs primarily deals with planktonic species in lentic waters, as it was thought that lotic conditions would not be favorable for mass growth of cyanobacteria. However, studies such as Uriza *et al.,* (2017) have found that toxic species that can produce MCs were found in all sampled Spanish habitats, such as creeks, streams, springs, and rivers, with the highest concentration being 1.87 µg/g of dry weight. Currently, there are no distinct protocols in Spanish literature for evaluating benthic cyanobacteria populations in lentic waters.

Most risk assessment decision systems are based on the WHO recommendations listed in Table 2 or on qualitative decisions that are based on literature reviews. However, it's becoming increasingly common to apply a quantitative modelling approach that is rooted in data, such as Bayesian networks (BN) in environmental modelling, risk assessment, resource management, or ecosystem services. Due to the probabilistic qualities of the outputs, results can be interpreted as the risk of proliferation and can support risk decision making process (Moe *et al.,* 2016). For instance, the EPA employs a BN for assessing large amounts of data and analyzing the relationship between nutrients and cyanobacterial abundance (EPA, 2015).

1.3.2 Mexican Recreational Waters

Mexican legislation also quotes the WHO standards for cyanobacteria, concentrations of MC, and Chl*a*Total in the Official Standard from 2014 (Tomasini-Ortiz, 2012). The water is governed by the National Water commission (CNA or Conagua) which is under the authority of the State Ministry of Public Works. Legislation also contains an Official Standards for Water Quality that regulates the conditions of the water for human consumption and utilization. To date, the office which oversees the distribution of cyanotoxins in water and possible poisonings is the Secretary of Health and Assistance which relies on the Federal Commission for the Prevention of Health Risks, specifically in relation to seafood. Besides this, cyanotoxins are excluded from water quality analysis (Diario Oficial de la Federación, 2017).

1.4 Bayesian Networks Background

Probabilistic frameworks models are becoming more commonly utilized to predict future changes in freshwater quality, such as BN. BNs are a powerful tool for many reasons, such as the ability to link abiotic process based models to biological quality indicators quantitatively (Moe *et al.,* 2019). This can then be viewed visually and communicated easily without the need of mathematical

background. Along with the ability to infer cause-and-effect relationships, it can also integrate multiple sources of information without relying on a single deterministic outcome that does not encapsulate real ecosystem interdependability and variability, as well as incorporate expert knowledge (Moe *et al.,* 2016). A simple example is outlined in Figure 3.

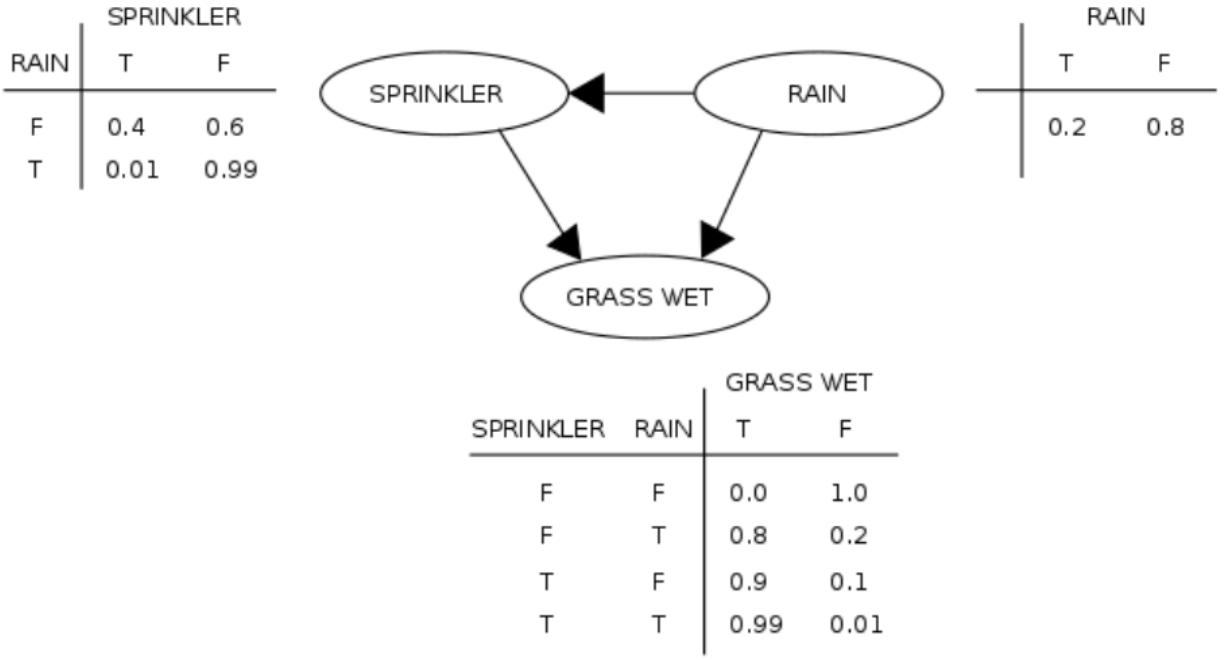


Figure 3. Simple example of Bayesian Network with probabilities (Goronto, 2017) T stands for true and F stands for false

BNs, also called Bayesian belief networks, are a multivariate probabilistic inference model which is able to demonstrate quantitatively and qualitatively conditional (in)dependencies between random variables (also referred to as nodes). Nodes can be continuous or discrete and has a finite set of states (Phan *et al.,* 2016). Qualitatively, the BN is able to show visually through a directed acyclic graph (DAG) the structure of the relationships between the nodes, which carries an advantage of being easy to understand while avoiding the necessity of computing any numerical calculations to show which variables are relevant and irrelevant (Aguilera *et al.,* 2011). This graphical structure can be defined as $G = (\mathbf{V}, A)$ where **V** is the node set (X_1, X_2, …, X_n) and A is the arc (or edge) that represents the probabilistic dependencies between them (known as probability tables i.e. takes an input and gives an output as a probability) (Scutari, 2010). In Figure 3, the nodes are Rain, Sprinkler, and Grass Wet, with the arrows pointing in the direction of the dependencies. So, Rain is the parent of Sprinkler and Grass Wet, etc. Once the structure is defined, it is important to understand the strength of the connections between the variables (represented by conditional probability tables (CPT)). A CPT is able to quantify the probability distribution of a node given the realized states of the parent nodes, or in other words, finding the probability of an outcome given known inputs. This is done by implementing Bayes' theorem which can determine a single probability distribution for each variable (Equation 1):

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)} \qquad \textbf{Equation 1}$$

where P(A) and P(B) are the probabilities of observing A or B without any regard to one another, P(A|B) is the conditional probability of A given that B happens, and P(B|A) is the conditional probability of B given A; additionally P(B|A)/P(B) represents the likelihood ratio, or also called Bayes' Factor (Phan *et al.,* 2016). Furthermore, to find the probability of a final event given all the other dependent event utilizes a joint probability distribution formula:

$$p(x_1, \ldots, x_n) = \prod_{i=1}^{n} p(x_i|pa(x_i)) \qquad \forall x_1, \ldots, x_n \in \Omega_{X1}, \ldots, X_n \qquad \textbf{Equation 2}$$

where Ω_{Xi} symbolizes all possible outcomes of variable X_i, and $\forall$ defines all possible values of x. A conditional distribution is also calculated for each variable $p(x_i|pa(x_i))$ for each variable X_i $i = 1, \ldots, n$ in relation to its parents that are denoted as $pa(x_i)$. Assuming X_i is the variable of interest, and X_E is the set of variables with known values, then X_i given X_E can be predicted by computing the probability of each possible value of X_i given each possible configuration of X_E (Aguilera *et al.,* 2011). In Figure 3, the CPT are to the side of each node (Rain with only two probabilities as it does not have parent nodes) that outline the probability of an outcome given the predecessor is true or false. For example, if the question of interest is "Given the grass is wet, what is the probability that it is raining?", by plugging the values from the CPTs into equation 2, it's possible to calculate each term of the numerator and denominator:

$$\Pr(R=T|G=T) = \frac{\sum_{S\in\{T,F\}} P(G=T,S,R=T)}{\sum_{S,R\in\{T,F\}} P(G=T,S,R)} \qquad \textbf{Equation 3}$$

For example, the $P(G=T, S=T, R=T) = P(G=T \mid S=T, R=T)\, P(S=T \mid R=T)\, P(R=T)$ which looks like 0.99 x 0.01 x 0.2 = 0.00198. Now plugging in all the values would give the equation:

$$\Pr(R=T|G=T) = \frac{0.00198_{TTT} + 0.1584_{TFT}}{0.00198_{TTT} + 0.288_{TTF} + 0.1584_{TFT} + 0.0_{TFF}} = 35.77\,\% \qquad \textbf{Equation 4}$$

Which gives a 35% chance that it is raining given the grass is wet (Gales, 2005, Goronto, 2017).

The statistical principle called Markov property regulates the structure of the DAG, in that every random variable X_i directly and solely depends only on its parents and is independent of the other nodes. The direction of the arrows (or arc) will represent which node depends on the other, with the dependent being the child node and the one who influences the child node being the parent node (Vieira *et al.,* 2017). Moreover, each stochastic variable has a conditional probability table associated with it that identifies the conditional probability distribution that relates to the different value combinations of the parent nodes (Egmont-Peterson *et al.,* 2005).

An important advantage of BN is the ability to predict the value of a target node from variables whose values can be more easily known. For example, if the variable of interest is X_i, and a data set X_E is known, then the value of X_i given X_E can be calculated using the probability of each possible value of X_i given each possible configuration of X_E. The probability distribution can be figured out through the joint distribution Equation 1. Another advantage of this type of statistics is the numeric values are inherently tied to the relationships of the nodes in the structure, meaning the probability of a particular hypothesis can be calculated automatically (Aguilera *et al.,* 2011).

The first step to the creation of a BN is structure learning, which is the creation of the DAG. This step needs to occur before quantitative probability distributions can be computed. There are two main ways that the structure can be learned which are: expert knowledge where the network is defined by specialists, or data-based learning (Scutari, 2010). Aguilera *et al.,* (2011) recommends using a combination (partly fixed) of the two for optimization of the model, especially in environmental modelling, which represents only 4.2% of papers that utilize BN. Data-based learning can be broken down into two main categories: score-based learning (which is finding the highest scoring network structure) or constraint-based structure learning (finding the network that best explains the dependencies and independencies in the data). The latter all depend on the Inductive Causation algorithm (examples implementing this algorithm: Grow-Shrink Markov Blanket, Incremental Association Markov blanket, Max-Min Parents and Children), whereas structure-based learning uses various heuristic search algorithms such as hill-climbing greedy search (Scutari, 2010). This project implemented hill-climbing score-based learning, which had several options for the scoring function such as log-likelihood (or entropy measure), Akaike Information Criterion, Bayesian Information Criterion, or K2 method. In the end, maximum log-likelihood was used to model the parameters of the network, which is a common method in literature (Galanti, 2015, Aguilera *et al.,* 2011, Scutari, 2010) and also is the scoring associated with the hill-climbing greedy search in the software applied in this thesis.

To strengthen the validity of the model, validation methods should be utilized, especially in models that are used to perform inference. If the model has a target node (such as in the case of this project), a sensitivity analysis can be carried out. This shows which variables or states of the variables are the most influential on the target and if small changes in the probability of the states will return large changes in the probability distribution of the target (Aguilera *et al.,* 2011). Furthermore, a BN is learned from a set amount of data, so data can be separated into two data sets: one for network learning (training set), and another for validation of the model (test set). Multiple commercial or free software exist and are listed extensively in Korb and Nicholson, 2003. Due to the general advantages for quantitative relationships between random variables to predict unobservable effects of a system, BN are becoming more frequently studied for water management and risk assessment analysis.

2. Objectives

The general objective of this project is to analyze risks for human health due to toxic cyanobacteria based on statistical evidence using Bayesian data modelling in three case studies. The three datasets that will be used will cover a range of representative situations including different geographical locations (Spain and Mexico), and different types of systems (lentic and lotic). The conclusions obtained after Bayesian analyses will be compared with the present decision tree in Spanish legislation for recreational waters in order to provide recommendations for future risk assessment for cyanobacterial proliferation in Spain (lentic and lotic) and Mexico (only lentic).

Specific objectives are as followed:

- Assess the current decision tree for reservoirs/lakes (planktonic cyanobacteria) in Spain and provide recommendations for its improvement based on new assessment methodologies.
- Provide recommendations for the creation and future elaborations of risk assessment for benthic cyanobacteria populations in Spanish rivers.
- Provide recommendations for the creation and future elaborations of a low-cost risk assessment of cyanobacterial proliferations in Mexico.

3. Materials and Methods

3.1 Data Organization for Risk Assessment Analyses

The data that was used to develop this master's thesis came from three distinct data sets from separate sources. The three groups of data described water parameters and presence of cyanotoxins are as follows: a large group of Spanish reservoirs used as bathing sites, a second group that includes benthic cyanobacteria populations from rivers, and finally another group with data from Mexican reservoirs with recreational usage.

3.1.1 Data Organization for Spanish Lentic Bathing Sites

The first data set used included bathing sites obtained from a national study that Universidad Autónoma de Madrid (UAM) and a public company called Tragsatec carried out for the Spanish Ministerio de Medio Ambiente y Medio Rural y Marino over a 15 month period between July 2008 to October 2009. The data contained 212 bathing sites, which were located in reservoirs, rivers, and one lake (Lake Sanabria), are all used recreationally during the bathing season, with each sampling point being tested three times (once in 2008 and twice in 2009) (Wörmer *et al.,* 2011). However, due to the fact that sampling techniques used during this time were only suitable for lentic sites, the data containing information about rivers had to be discarded (Samuel Cirés, personal communication). Out of the 212 bathing sites, only 88 reservoirs/lakes sampling points were used.

The 88 sampling points contained 76 individual lentic bodies of water, some of which served as sources for irrigation, energy, and/or drinking water for human consumption (23 of the 76 were used as the latter) (SNCZI-Inventario de Presas y Embalses, 2020). Figure 4 below shows the locations of each reservoir as well as the associated risk of cyanobacterial proliferation (red meaning high, yellow is moderate, and green is low) classified by the authors of the study following the decision tree mention before (Wörmer *et al., 2011).*

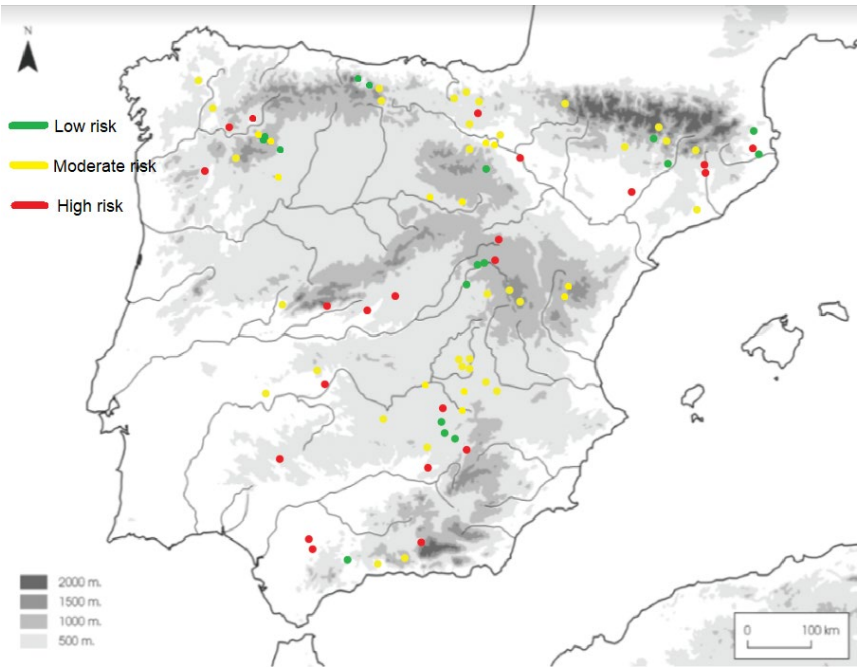


Figure 4. Location of 88 Spanish lentic bathing sites taken (Modified from Wörmer *et al.,* 2011), each point represents a sampling point. Red dots mean high risk of proliferation, yellow is moderate, green is low risk.

The physicochemical variables used in this project are listed in Table 3. Nitrates and phosphates were tested in situ using spectrophotometry by colorimetric methods using a DREL-2010 Portable Laboratory (portable spectrophotometer), following APHA 1992, while Chl*a*Total and Chl*a*Cyano concentrations were measured by fluorometry (Moldaenke BBE Algae Analyzer Fluorimeter). Taxonomy for the identification of potential toxic cyanobacteria genera (PresenceofToxic) was analyzed in water samples after 24-hour flotation using an optical microscope (Olympus BH2 equipped with a Leica DF300 FX camera). Species identification was then classified based on morphological characteristics according to Geitler (1932) and Kmareck and Anagnostidis (1989, 1999, 2005). For the dissolved fraction, MC was analyzed through a graphite carbon cartridge and then inspected by mass spectrometric analysis (Varian 500MS Ion Trap Mass Spectrometer), following Wörmer *et al.,* (2010). Sestonic MC was eluted by washing the cartridge with methanol 90% and then concentrated under a vacuum until inspection by mass spectrometric analysis, following Wörmer *et al.,* (2010) once again.

3.1.2 Data Organization from Mexican Lentic Sites

Table 3. Parameters included in the Bayesian Network for each of the three datasets analyzed

Variables included in BN	Spanish Reservoirs and Lakes	Mexican Reservoirs and Lakes	Spanish Rivers
Number of data points	88	65	46
Number of waterbodies	76	8	10
Season	X	✓	✓
Month	X	✓	X
Temperature (°C)	X	X	X
Conductivity	X	✓ (mS/cm)	✓ (µS/cm)
pH	X	✓	✓
Dissolved Oxygen (mg/L)	X	✓	X
Chla Total (µg/L)	✓	X	X
Chla Cyano (µg/L)	✓	X	X
Maximum Dominance of Cyanobacteria (%)	✓	X	X
Presence of Toxic Genus	✓	X	✓
Nitrate (NO_3^- mg/L)	✓	✓	✓
Phosphate (PO_4^{3-})	✓	✓	✓
Light	X	X	✓
Depth (cm)	X	X	✓
Biofilm size (%)	X	X	✓
Current Velocity (m/s)	X	X	✓
Microcystins (µg/L)	✓	✓	
Anatoxin-a (µg ANATX/ mg dw)	X	X	✓
Risk	✓	X	X

Data from Mexico was collected using a literature review of papers that included physicochemical parameters of a reservoir and the corresponding MC (ug/L) concentrations. This was accomplished by doing an advanced keyword search on the Bun search engine of UAM using terms “cyanobacteria”, “physicochemical”, “microcystin”, and “Mexico”. A list of potential resources regarding MC concentrations were provided by UAM (Munoz *et al.,* 2020). In the end, the variables used were those that were found consistently in the literature resulting in a larger dataset that produced a more accurate model; the academic journals that supplied the data are outlined in Annex 2. The variables included (Table 3): the month sampled, the season, phosphates, nitrates, temperature, conductivity, pH, dissolved oxygen, and concentration of MCs. The MC was analyzed in each study by using the ELISA technique (EnviroLogix, USA). Conductivity was measured in mS/cm due to the relatively high values encountered. Dissolved oxygen (DO) was also kept in mg/L (instead of converting to percentage) due to that being the unit found ubiquitously in the literature. Compounding the relatively common nature of these variables, they were also selected for their parsimonious nature (highest explanatory power with the lowest cost). Finally, the data included 65 sampling sites consisting of eight reservoirs and lakes used for drinking water, irrigation, collection of seafood, recreation, or aquatic sports, with the emphasis of this project being recreational activities. These reservoirs ranged from mesotrophic to hypertrophic on the trophic index. Figure 5 displays the positioning of the eight reservoirs.

The reservoirs included are as follows: Valle de Bravo reservoir, Los Berros reservoir, Villa Victoria lake, Lake Chapultepec, Lake Alameda, Pista Olímpica de Remo y Canotaje artifical lake, Lake Bosque de Aragón, and Lake Zumpango. Lake Zumpango comprises 52% of the 65 sampling sites collected due to an extensive study done by Figueroa-Sanchez *et al.,* (2020). This lake is used

for irrigation, aquaculture, and recreation. It was tested in three spots during the course of one year: one where partially treated wastewater enters, another in the open water, and the last at the river mouth (Figueroa-Sanchez *et al.,* 2020). Valle de Bravo reservoir makes up 17% of the data. This reservoir contributes 38% of the water belonging to the Cutzamala Hydraulic System, and supplies water to Mexico City and the City of Toluca (~six million people) as well as supporting water sports. Four rivers feed the reservoir, meaning deterioration of the water quality due to diffuse pollution causing concerns over the quality of the drinking water (Alillo-Sanchez *et al.,* 2012, Figueroa-Sanchez *et al.,* 2014). The remaining 31% of reservoir data used was found in Valle de Mexico in Arzate-Cardenas (2008).

Figure 5. Map of eight lentic bathing sites, each point represents separate bodies of water (map found at Mexico Elevation map – SpeakLounge, 2020)

3.1.3 Data organization of Spanish lotic sites

New data samples from ten rivers (Mediano, Manzanares, Lozoya, Eresma, Jarama, Escabas, Tajo, Guadiela, Caldarés, Brazato), located along a longitudinal gradient of Spain, were collected from 2011 to 2016 (Quesada *et al.,* 2016). Each river had a specific sampling point, except Manzanares having two. Furthermore, the Manzanares and Mediano rivers were sampled in three different moments, obtaining different samples during different seasonal periods associated with the Mediterranean climate (spring, summer, and autumn). The other samples were extracted only during the one Mediterranean seasons. 46 specific biological samples (cyanobacterial mats) were considered during the compilation of data for the construction of the BN for Spanish rivers. These samples correspond with potentially *Phormidium* dominated mats located along the ten rivers

(Ramos, 2012, Kønig, 2013, Haya, 2016, Quesada *et al.,* 2016 and Perona *et al.,* 2017). All biological samples were used in order to obtain the largest dataset possible, which in turn will provide a more accurate inference diagram. Figure 6 shows the geographic location of these ten rivers as well as their sampling points in relation to each other. Parameters are outlined in Table 3.

Conductivity and pH were measured in situ using a specific portable probe for each parameter, as well as current velocity, depth, and light (measured using the quadrant technique outlined in Necchi *et al.,* (1995). The size of the biofilm in percentage was estimated by using a quadrant to determine the percentage of surface area covered (Haya, 2016). Inorganic nutrients were measured using a Hach portable laboratory (following APHA, 2005 and methodology laid out in Perona *et al.,* 1999). Mat morphology and composition were determined by a Leica Z7 stereoscope, while the cyanobacteria taxonomy was identified using Komarek and Anagnostidis (2005). ANA (µg ANA/mg dw) (ANA) was measured using Receptor Binding Assays (described in Haya, 2016) and analyzed due to its being the most common cyanotoxin present in benthic populations with the presence of *Phormidium spp.* and being the cyanotoxin used in the limited legislation.

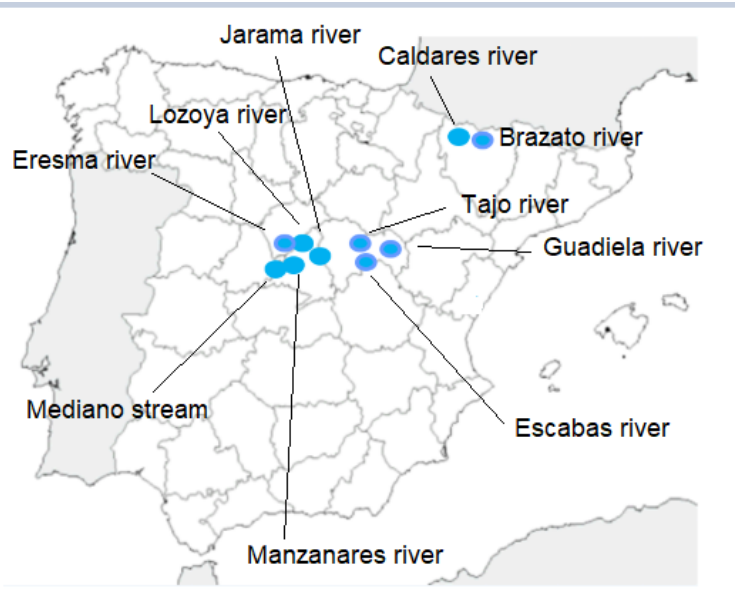


Figure 6. Geographic distribution of the ten rivers with potential *Phormidium* mats

3.2 Data Preprocessing for Creation of Bayesian Models

Once the data was gathered, it was distributed into three individual excel sheets for each group of data: Spanish lentic systems, Mexican lentic systems, and Spanish lotic systems. Presence/absence data in Spanish systems was changed into a binary 1/0. These data sets were then uploaded to Rstudio (version 3.6.3) to perform statistical analyses. The candidate explanatory variables are outlined in Table 3. These variables were used to determine the relationships between abiotic and biotic data and the corresponding concentration of cyanotoxins, which was determined through linear correlation, and scatter plot smoothing. The function “ggpairs” was used in the R package “GGally” to analyze the linear correlation through a scatter plot matrix. The explanatory variables with a positive correlation coefficient show that as the independent variable changes, then the response variable changes in the same direction, and vice versa with negative correlation. The

highest the value of the correlation coefficient, then there is a stronger relationship between the two variables. If the value is low, then it indicates that the variables are hardly related. All r values will range between -1 to +1, where 0 symbolizes no relationship at all. A correlation matrix was carried out in order to decipher if any variables could be omitted from the BN, due to low correlation.

To summarize and visualize trends between observations and variables, a Principal Component Analysis (PCA) was implemented through RStudio, utilizing packages “ggplot2” and the function “prcomp”. Analyzing these relationships, it’s possible to also distinguish if some variables are so correlated that they would duplicate information that is already being used to describe the system. To carry out BNs, it is important to reduce redundant variables as with the more variables included, the more data is needed in order to observe the conditional dependencies. Aguilera *et al.,* (2011) affirms that increasing the number of variables will increase the complexity of the model, and more data is needed to calculate the probability distributions. A limitation of PCAs is the inability to analyze data rows with missing data. Mexican lentic systems were missing conductivity values from Los Berros reservoir (5 of the 65 data points), which were removed from the analysis. Spanish lotic systems were missing biofilm size, depth, and cv from 6 of the 46 samples, which accounted for three entire rivers (Guadiela, Caldarés, and Brazato) being omitted from the Spanish lotic PCA. Scaling was used in all the datasets due to large discrepancies in the standard deviation of the variables due to the magnitude of units. Non-numerical data such as light availability, month, and season were not included in the PCAs due to the inability for this analysis to handle numeric, binary, and categorical together.

The excel sheets need to be converted into .txt data files in order to be uploaded to the free learning software: Genie 2.5.R4 (BayesFusion LLC). While BN can use continuous nodes or discrete nodes, there are more algorithms and scoring methods available for that of discrete. Furthermore, continuous BN have more limitations that are restrictive for structure learning as well as over the probability distributions. Due to these reasons, there are more traditional discrete BN in literature with continuous BN only being found in 4.4% of all papers using this type of statistics (Aguilera *et al.,* 2011). In this project, a discrete BN was implemented, meaning discretization of the 14/19 continuous variables (with only 5 variables being discrete: light, season, month, presence of toxic genera, and risk). Two approaches were taken for this process. Eleven of the 14 variables were discretized with unsupervised hierarchical k-means clustering, which calculates continuous distance-based similarity measures to group the data points together in relation to the calculated centers of the clusters into as many clusters as required (Dash *et al.,* 2011). The remaining 3 variables (MC concentration for both lentic systems, ANA concentration for the lotic system, and biofilm coverage %) were bounded by thresholds found in legislation and recommended worldwide guidelines suggested by WHO. MC concentration was discretized based on WHO standards for low, moderate, and high risk of human health problems after recreational exposure, as well as the guideline for drinking water. However, due to the frequency of samples in both systems under the standard of 1 µg/L, both systems were discretized into three more intervals for the sake of sensitivity. ANA was bounded by legislation for benthic populations in New Zealand drinking water, and biofilm % was discretized based on recreational legislation in Cuba and New Zealand (Ibelings *et al.,* 2014). All ranges for continuous and discrete variables used in each of the three BN are listed in Annex 1. Additionally, Chl*a*Total was not discretized based on literature, due to quantitatively assessing the cause and effect between predictors of blooms and actual concentrations. In general, the more

discretization bands there are, the smaller the error. In this paper, six discrete bins were used, as this led to the highest score of all the networks tried but was still small enough to fit the data.

Originally, only 80% of the dataset was used for the creation of the BN, removing 20% in Rstudio through the function “sample” for a validation method called Train and Test Cross Validation (Rigosi *et al.,* 2015). However, in Aguilera *et al.,* (2011), it was stated that an estimated model should be learned with the complete dataset even if the Train and Test validation method is used, especially if the end goal is user friendly application. This is so that the model is as accurate as possible, which is done by using the most amount of data available. Data division was carried out at the end in order to validate the model.

The final step of preprocessing the data is to distinguish if any of the datasets contain missing values. Although there are some learning algorithms that can handle missing information (Expectation Maximization algorithm), most traditional statistical techniques for learning BN cannot deal with. In Aguilera *et al.,* (2011), it is reported that only 9.6% of all BNs constructed in the literature review were able to function with missing values. The software (Genie) implemented for this project was not able to handle missing data, and offered three choices for correction: 1) delete the rows with missing data, 2) replace with a specified value, or 3) replace with the average value of the entire column. Out of the three datasets, two were missing data, which were Mexican lentic systems, and Spanish lotic systems. Due to the lack of data and the effort to retain all data available, option 3 was utilized. Mexican lentic systems were missing conductivity values (5 cells) for Los Berros reservoir in the State of Mexico, and a value of 0.91 was used to replace these cells. Spanish lotic systems were missing biofilm size, depth, and current velocity for six rivers: one point on Mediano river, one point on Manzanares river, Jarama river, Guadiela river, Caldarés river, and Brazato river. The values used to replace the missing information was 47%, 10.08 cm, and 0.25 m/s respectively. The choice to replace the data instead of deleting the rows was made due to the fact that Jarama, Guadiela, Caldarés, and Brazato rivers only consisted of one data row, and if this was removed then the BN would lose 40% of the river diversity, significantly weakening the model. Data preprocessing is a vital step for the success of an accurate BN.

3.3 Creation of Bayesian Models and Statistical Analysis

The goal of a BN is to determine the posterior conditional probability distribution of all the possible unobserved outcomes given observed evidence. As mentioned before, BN can be learned through data-driven processes or through expert knowledge. To learn the structure there are several learning algorithms that can be used on discrete variables, and these are usually broken down into two main categories: constraint-based algorithms (such as Grow-Shrink) or score-based algorithms. Score-based algorithms maximize a heuristic search algorithm, which can help identify the network with the highest probability distribution (the best network maximizes the posterior probability). Genie makes use of a few learning algorithms, but this project used “Bayesian Search” which is a score-based algorithm that uses hill climbing procedure that applies log-likelihood scores. Arcs are scored by using the BDe score for the prior link probability.

Of the three networks, only the Spanish lentic system was able to be derived completely from data. Mexican lentic systems and Spanish lotic systems both had arcs that were forced through

the background knowledge option. These connections were determined through analytical analysis and through literature reviews (Noges *et al.,* 2010, Huber *et al.,* 2012, Wagner and Adrian, 2009, Pawlik-Skowrońska and Toporowska, 2016, Rigosi *et al.,* 2015). Ten of the 12 arcs in Mexican lentic system were forced out of 9 nodes. Six out of 13 arcs in Spanish lotic systems were coerced out of the 11 nodes. More arcs were needed to be forced in the Mexican dataset due to having the least amount of different water bodies, which caused the data to be too similar to see a clear conditional dependency.

The final stage of the creation of a network is the evaluation of the predictive ability through validation, scenario analysis, and sensitivity analyses. Each network was tested for their predictive performance by using a random 20% of the data points to assess the Correctly Classified Instances or CCI. Generally, a model can be considered 'good' if it has a CCI of at least 0.7 (Shan *et al.*, 2019). CCI is calculated by dividing true positives and true negatives by all true positives, true negatives, false positives, and false negatives. Next, a sensitivity analysis was carried out on the end node (or goal node) to determine which factors affected the concentration of cyanotoxin the most. This technique validates the probability parameters of the BN. It is done by analyzing the effect of small changes in inputs on the output parameters (or the posterior probability) (BayesFusion, 2020). This is done in Genie by setting "target nodes" which in this case is the cyanotoxin concentration and presence/absence of toxic genre. Then an algorithm proposed by Kjaerulff and Van Der Gaag (2000) calculates a set of derivatives of the posterior probability distribution of the target nodes over each of the input parameters. If the derivative is large for the parameters, then small changes can give rise to a large change in the posterior distribution of the target, and vice versa (BayesFusion, 2020). This analysis then generates nodal percentages which can be compared within the same network to indicate which parameters have the greatest effect on the target node. As Wang, 2006 claims, it is an elicitation process that can help rid of parameters that do not affect the target node and shows the most influential which will cut down on costs and increase accuracy. More accurate values of targeted probability parameters can be obtained by refining the BN to only contain influential parameters, and after performing the sensitivity analysis, more attention can be focused on the probabilities in which the network's behavior exhibits highest sensitivity. The uninfluential variables can have fewer discretization intervals with more rudimentary estimates (Wang, 2006). Validation of the model is an important step, although in the literature review done by Aguilera *et al.,* (2011), 37.7% of models have no validation technique. For a model to be used for inference or predictability, validation is crucial. Finally, potential scenarios are run to establish possible future events and how cyanotoxin probability changes. The pertinent steps necessary for the construction of a BN is shown in Figure 7.

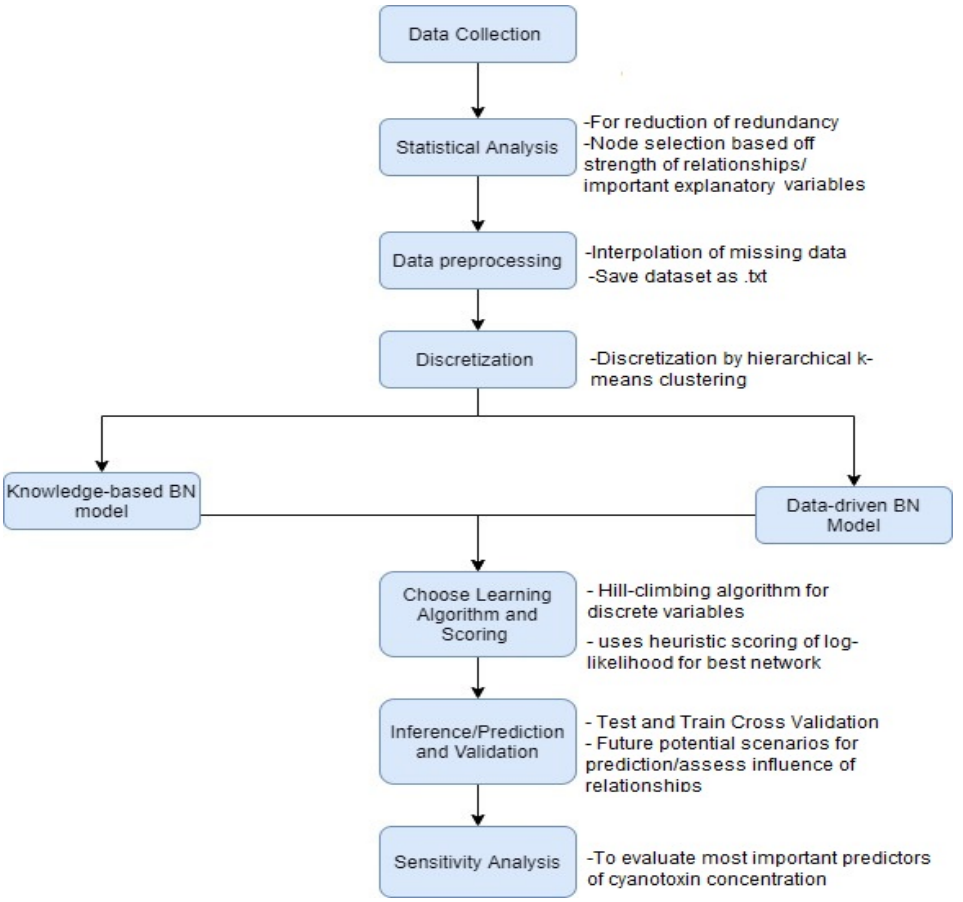


Figure 7. Flowchart of steps necessary for the creation of a BN for gauging the relationship between predictors/indicators of cyanotoxins and cyanotoxin concentration by knowledge-based BN and Data-driven BN

.

4. Results

4.1 Physicochemical and Biological Characteristics of Water Bodies in Spain and Mexico

The physical, chemical, and biological variable ranges are outlined in Table 4. This table also shows the average value of each sample, as shown in the parenthesis. The only parameters that were present in all of the three datasets were nitrates (NO_3^- mg/L) and phosphates (PO_4^{3-} mg/L). Spanish lentic systems had the highest maximum value of nitrates (9.5 mg/L), and the highest mean value as well (0.8 mg/L). Mexican lentic systems had the highest maximum value for the phosphate range (14 mg/L), and highest average value (6.02 mg/L). Mexican lentic systems and Spanish lotic systems had two more comparable parameters: pH and conductivity. The pH in Mexican systems was more alkaline than in Spanish lotic systems, with also a higher minimum, maximum, and average values of conductivity.

Table 4. Maximum and minimum values of all numerical parameters used in BN for each dataset. Values in parenthesis are the mean value

Variables	Spanish Reservoirs and Lakes	Mexican Reservoirs and Lakes	Spanish Rivers
Nitrate (NO_3^- mg/L)	0.0 - 9.5 (0.8)	0.01 - 1.5 (0.59)	0.08 - 3.1 (0.36)
Phosphate (PO_4^{3-} mg/L)	0.0 - 4.04 (0.18)	0.003 - 14.00 (6.02)	0.002 - 0.08 (0.038)
Chl*a* Total (µg/L)	0.2 - 178.8 (9.6)		
Chl*a* Cyano (µg/L)	0.0 - 89.7 (3.9)		
Maximum Dominance of Cyanobacteria (%)	0.0 - 95 (28)		
Conductivity µS/cm		139 - 4400 (900)	105 - 1369 (106)
pH		6.8 - 11.00 (9.05)	5.5 - 8.4 (7.0)
Current Velocity (m/s)			0.0 - 1.52 (0.2)
Biofilm size %			5.0 - 100 (47)
Depth (cm)			0.0 - 34.9 (10.1)
Temperature (ºC)		14.8 - 24.4 (20.1)	
Dissolved Oxygen (mg/L)		1.4 - 18.3 (6.7)	
Microcystins (µg/L)	0.0 - 123.6 (1.4)	0.0 - 12 (1.3)	
Anatoxin-a (µg ANA/mg dw)			0.0 - 38.8 (2.9)

MC concentration was measured in the two lentic systems, with only ANA being measured in lotic systems. Between the two systems that had MCs, Spanish lentic systems had an overall higher maximum of 123.66 µg/L, but the two averages are comparably the same with 1.4 µg/L (Spanish lentic systems) and 1.3 µg/L (Mexican lentic systems). Furthermore, only 13 of the 88 sampled sites in Spanish reservoirs and lakes tested any concentration of MC at all, and only 2 of those exceeding the drinking water guideline of 1.0 µg/L. Only one site tested MC levels over the recommended WHO recreational threshold for high risk (20 µg/L), with the rest being classified as low risk. Fifty of the 65 Mexican reservoirs and lakes gathered in the literature review had MC concentration, with 26 of those surpassing the recommended drinking water guideline. However, only one site surpassed the WHO recreational guidelines for low risk (10 µg/L) and would be classified as moderate risk. It should be kept in mind that the literature review was conducted on articles that records of MC concentrations, as this was the goal. ANA in Spanish lotic systems had a maximum value of 38.8 µg ANA/mg dw that was found in autumn in the Escabas river.

Risk of proliferation in Spanish lentic systems, classified based on Spanish legislation, consisted of 18 of 88 (20.5%) samples being classified as low, 46 (52.3%) being classified as moderate, and 24 are classified as high risk (27.3%). As stated before, neither Mexican recreational waters nor Spanish lotic systems have official legislation related to risk.

4.2 Statistical Relationships Amongst Environmental Factors, Cyanotoxin Concentration, and Presence of Toxic Genera

Linear regression models were carried out to determine the most significant explanatory variables for predicting cyanotoxin concentration, which could then be implemented in the knowledge-based BN (Mexican lentic and Spanish lotic). All linear correlation coefficients for the pairs are shown in the upper right panels of Figures 8, 11, and 14, while scatterplots with linear regression are displayed in the lower left-hand panels.

4.2.1 Statistical Analysis of Spanish Lentic Systems

Figure 8 outlines the correlation coefficients for all pairs in the dataset referring to Spanish lentic systems. The explanatory variables with the highest linear correlation with concentration of MC were ChlaTotal (r = 0.85), and ChlaCyano (r = 0.78). The variable with the lowest correlation was PO_4^{3-} (r = 0.007), with the second lowest being the other nutrient indicator of NO_3^- (r = - 0.067) with a negative correlation. Although the two nutrients had a smaller linear correlation with MC than the rest of the parameters, the scatterplots showed a stronger non-linear increase in MC at low levels of PO_4^{3-} and NO_3^-.

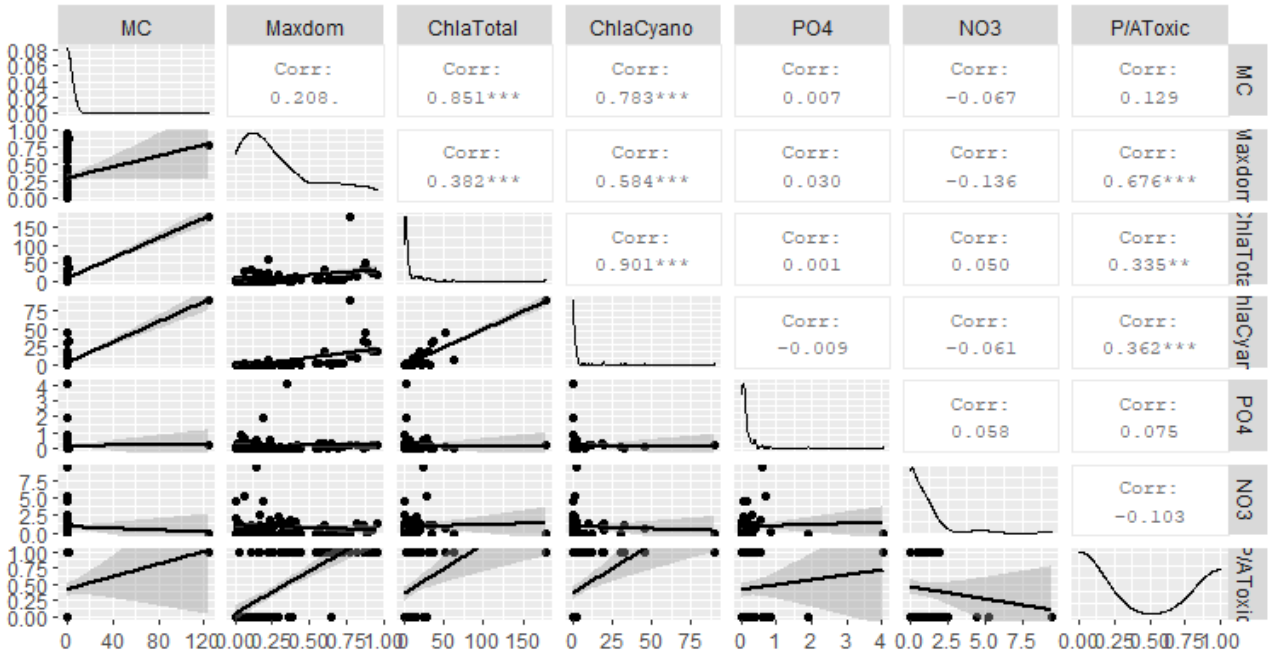


Figure 8. Correlation analysis of possible variables for BN, with regards to variables of interest being MC and presence/absence of toxic genera. Upper panel corresponds to the correlation coefficient (r) of each pair, values with the highest correlation are marked with three stars. The lower panel shows the linear regression curve of the scatter plot. Abbreviations are as follows: MC (microcystin), Maxdom (Maximum dominance of cyanobacteria), ChlaTotal (Total Chlorophyll *a*), ChlaCyano (Chlorophyll *a* from cyanobacteria), PO4 (Soluble reactive phosphorus), NO3 (Nitrates), P/AToxic (Presence or absence of toxic genus).

The variables with the highest correlation with the presence or absence of toxic genus (another goal node) was maximum dominance of cyanobacteria (r = 0.676), and ChlaCyano (r = 0.362). Therefore, the most predictive variables for MC concentration are ChlaTotal, and ChlaCyano, with a slightly higher correlation with NO_3^- (r = -0.067) than PO_4^{3-} (r = 0.007).

Additionally, the possible set of predictor factors for presence or absence of toxic genus would be maximum dominance of cyanobacteria, and ChlaCyano.

The dominant environmental variables that describe each sampling point for Spanish lentic systems were then analyzed by a PCA (Annex 3). The groups were classified based on risk of cyanobacterial proliferation (classified by the Spanish decision tree outlined in Figure 2), as low, moderate, and high risk. About 65% of the variance of the sampling sites can be explained by Annex 3. The First Principal Component (PC1) is associated with MC concentration, ChlaTotal, ChlaCyano, Maximum dominance, and Presence of toxic genus which describes the variability between the sampling sites the most. This can also be described as PC1 is related with the biological indicators, and also shows that a majority of the high risk sites are connected with these indicators. Low and moderate risk sites are located along PC2 which is an indicator of nutrient overloading. PC2 also indicates the negative correlation between nitrates and the presence of toxic genera, while phosphates are positively correlated. From the positioning of the loadings, it can be determined that all variables give information about the variability of the sampling sites and are not considered redundant.

4.2.1.1 Spanish Lentic Bayesian Network

Spanish lentic BN was data-derived, solely from the data of 88 sampling points throughout the various lentic systems. The discrete probability distributions in the CPTs were calculated for each node; Table 5 outlines the CPT for the target node of MC. The CPT of each node was calculated by the frequency distribution of the variable across each state of the parent node. For example, for MC concentration to be at its lowest value of below 0.01 µg/L with a parent node of ChlaTotal given a state of below 5 µg/L, the probability value of 0.91 is calculated by taking the count of observed values of MC in the interval ChlaTotal under 5 µg/L (56), and dividing by the total number of observations of ChlaTotal >5 µg/L (61) which the probability can be written as 56/61 = 0.91 (shown in the upper left-hand cell in Table 5). This also follows Equations 1 and 2 from section 1.4. Of the data samples for Spanish lentic sites, 86.4% fell in the range of 0 - 0.01 µgMC/L, 11.4% in 0.01 - 1 µgMC/L, 1.14% in 1 - 10 µgMC/L, 0% in 10 - 20 µgMC/L,, and 1.14% >20 µgMC/L.

Table 5. Conditional Probability Table for the node of interest "Microcystins". Each column represents the probability distribution of MC based on the given states of the parent node ChlaTotal. ChlaTotal = Chlorophyll *a* Total, MC = Microcystin

ChlaTotal (µg/L)	below 5	b 5 to 10	b 10 to 20	b 20 to 44	b 44 to 100	above 100
MC concentration (µg/L)						
below 0.01	0.91	0.90	0.65	0.62	0.39	0.083
b 0.01 to 0.4	0.083	0.018	0.02	0.12	0.39	0.083
b 0.4 to 1.0	0.002	0.018	0.27	0.12	0.05	0.083
b 1.0 to 10	0.002	0.018	0.02	0.12	0.06	0.083
b 10 to 20	0.002	0.018	0.02	0.01	0.06	0.083
above 20	0.002	0.018	0.02	0.01	0.05	0.58

The BN developed for this system is outlined in Figure 9. The best score in 19 iterations was -718.544, which was increased as the intervals went from three to six. The Log Likelihood was at -476.906, which is better (larger) than the others (Clauset, 2011). It should be noted however, that the larger the number of parameters, the more flexible and complex the model will be, but also the likelihood assigned will be lower. However, this was the best score of all the discretization attempts while keeping a degree of sensitivity and of the heuristic searching for the network. In all possible models found, nitrates were the child node of ChlaTotal, and phosphate the child node of ChlaCyano. Maximum dominance was always connected to the presence of toxic genus, although sometimes as the parent and sometimes as the child. Risk was always the child node of MC and ChlaTotal. No background knowledge was used to develop the outline of the model.

The arrows in Figure 9 show the strength of influence between the nodes. The thicker arrows represent a stronger influence connection, where all influences are calculated from the CPT of the child node, and basically shows the distance between the conditional probability distributions of the child node depending on the state of the parent. It's calculated using average Euclidean distance (Genie, 2020). The strongest connections are between the node of interest (MC) and ChlaCyano (0.48) and ChlaTotal (0.43). The third highest strength of influence is between ChlaCyano and PO_4^{3-} (0.41).

A BN is run for prediction or inference of a node of interest by changing the probability distribution of one or more of the nodes. This will then update all the probability distributions of all the nodes that are linked through their CPTs. This is normally done by setting evidence of one of the parent nodes, which will then show 100 % probability, as there is a 100% probability of being in the state if it is set accordingly. Two "what-if" scenarios were carried out to portray the combined effects of low-cost parameters of ChlaTotal and ChlaCyano on the concentration of MC and whether toxic genera are present. Two examples are outlined in Figure 9: a) using WHO low risk recommended thresholds for ChlaTotal <10 µg/L and the lowest interval of ChlaCyano <4.8 µg/L and b) using the next discrete interval higher than the previous model to analyze changes in MC.

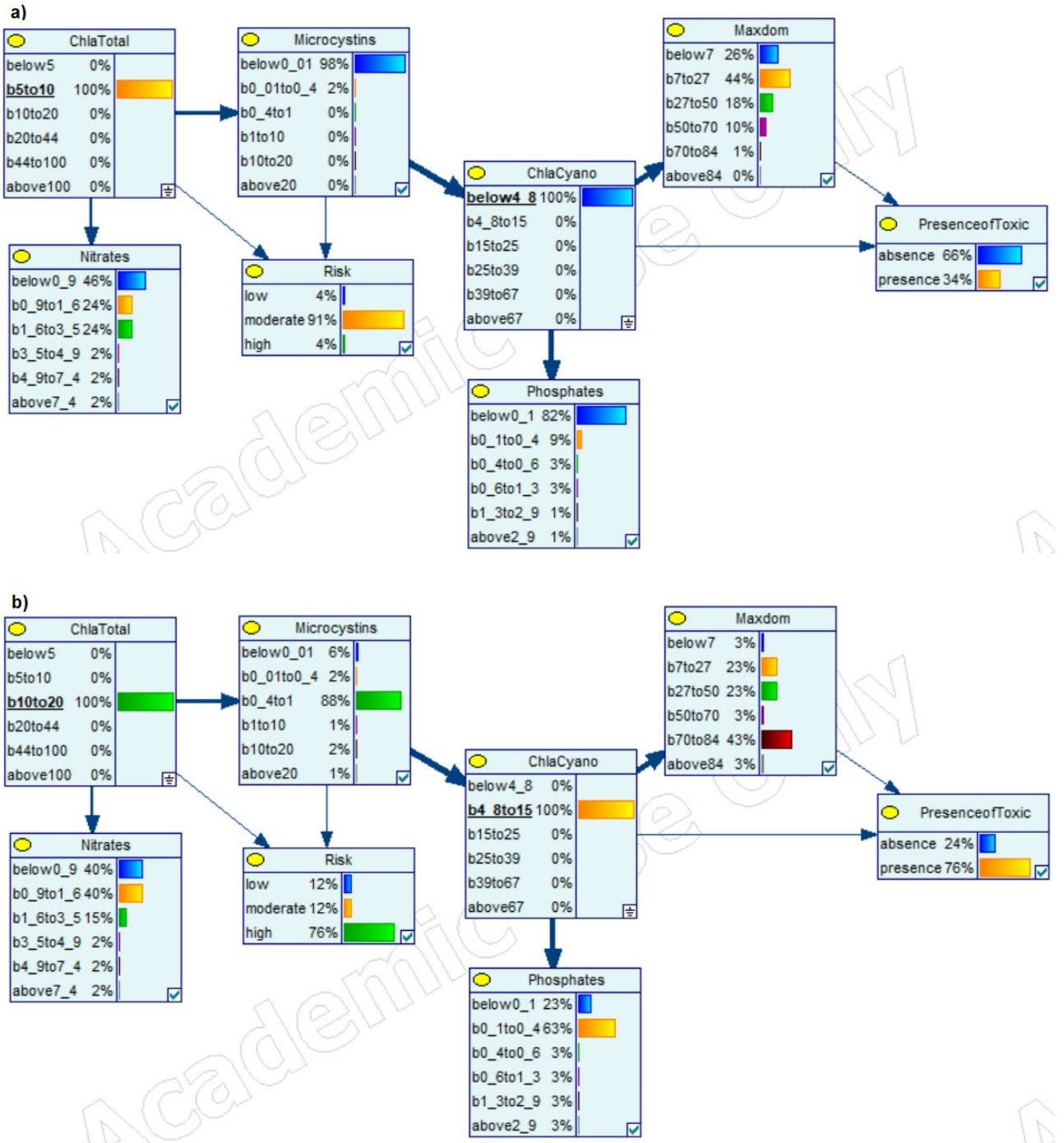


Figure 9. BN model structure for MC concentrations, Risk, and presence/absence of toxic genus (PresenceofToxic) in 88 sampling sites throughout Spain. The node states consist of the either below: meaning less than, b __ to ___ : meaning between a value to another value, or above: meaning more than. Decimals are replaced by underscores as demanded by the software (example: 4.8 is 4_8 in the model). Screenshots of scenarios a) WHO low risk thresholds for ChlaTotal b) WHO moderate risk thresholds for ChlaTotal

The WHO thresholds for low risk represented in Figure 9a give a 98% probability of having 0 - 0.01 µg/L and a 68% probability of the absence of any toxic genera. According to the Spanish risk assessment decision tree, this would qualify as moderate risk which contradicts the WHO

guidelines. Increasing the discrete range by one bin, and therefore progressing into the WHO moderate risk thresholds, the MC concentration probability range increases to 0.4 - 1 µg/L with a probability of 88%, with a 76% probability of having toxic genera present. The MC range still falls within low risk parameters for MC according to WHO, although is categorized as high risk by the Spanish decision tree. This MC range is still considered potable under Spanish legislation. In addition, it can be seen that when ChlaCyano increases, phosphate concentration as well as Maxdom also increase.

4.2.1.2 Validation of Spanish Lentic Bayesian Networks

Sensitivity analysis leads to the conclusion if it is pertinent to acquire more accurate estimates for the probabilities. The sensitivity analysis for Spanish lentic BN is shown in Figure 10.

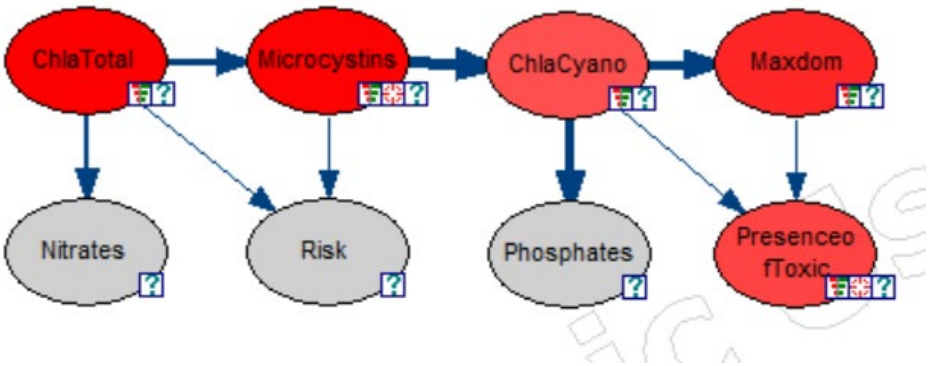


Figure 10. Sensitivity Analysis of Spanish Lentic system BN, darker red indicating stronger influence

The sensitivity analysis was performed on the two target nodes: presence of toxic genera and on concentration of MCs. Risk was overlooked due to the aim of this project to suggest new quantitative thresholds for risk assessment. The darker red nodes represent the more influential parameters and are important for the calculation of the posterior probability distribution of the target node (where a small change will cause a larger shift in the target node), and the lighter the color red gets, the less influential the node. Gray symbolizes a lack of sensitivity altogether, or that they are not used in the calculation of the target node. As has been shown in each analysis, nitrates and phosphates were not influential on MC or the presence of toxic genera. ChlaCyano is also one of the least influential due to the direction of the arc between MC and ChlaCyano. The average sensitivity coefficient (calculated on each of the target nodes CPT) for MC was 0.02, maximum was 0.8 which was for ChlaTotal parent state of >5 µg/L and a minimum of 0. While for presence of toxic genera average was 0.014, maximum was 0.35 which is for parent states of Maxdom between 7% to 27% and ChlaCyano under 4.8 µg/L, and the minimum of 0. Altogether, MC's conditional probabilities were more sensitive to small changes in the parameters than presence of toxic genera, and ChlaTotal was the parameter that most affected MC concentration probability. For elicitation purposes, the strongest predictors of MC concentration and presence of toxic genera is ChlaTotal, ChlaCyano, and Maxdom. Phosphates and nitrates more strongly correlate with concentrations of chlorophyll. If both chlorophyll concentrations are known, nutrients can be removed from the model. However, to predict biological parameters they should be kept in the model.

The last validation of the model for evaluation of predictive performance is CCI. Due to a limited data set, the model was made with the entire data set, as outlined in Aguilera *et al.,* (2011). However, 20% of the dataset was used for verification of the mode. A validation scenario was considered "accurate" if the MC concentration, risk, and presence of toxic genera had the highest posterior probabilities in the correct discretization range. Seventeen of the 88 data sets were used for testing the model. Five of these seventeen had one of the three end nodes predicted incorrectly (as in it predicted a value that was different than the observed value in the dataset), but never all three at once. Table 6 outlines which parameters were predicted incorrectly by the model at each station, with the model giving a different value than was measured. The CCI for this model was 0.70, which corresponds to the value outlined in Shan *et al.,* (2019) as considered a good model. In Embalse Encinarejo, the BN predicted a range of 1 - 10 µg/L for MC, when in reality the system had 0.0 µg/L. However, CYN was also measured in the dataset (although not used in the BN for comparability with Mexican lentic systems and with WHO recommended thresholds) and measured 3.5 µg/L CYN in this reservoir. Presence of toxic genera had less accuracy predicting when presence of a toxic genera when Maxdom was not input. Risk was always classified incorrectly in the testing dataset when the risk was low, but 100% correct with high/moderate risk.

Table 6. Incorrectly Classified Instances of Spanish Lentic BN based on end nodes of interest: MC= microcystins, Presence of Toxic = presence/absence of toxic genera

Site	MC concentration	Risk	Presence of Toxic
Embalse Encinarejo	✗	✓	✓
Embalse Cubillas	✓	✓	✗
Laguna Salvadora	✓	✗	✓
Lago Carucedo	✗	✓	✓
Embalse Entrepeñas	✓	✗	✓

4.2.2 Statistical Analysis of Mexican Lentic Systems

The variables found in a literature review for Mexican lentic systems is outlined in Table 4. A linear correlation scatter plot (Figure 11) for the 65 data points showed that the explanatory variables with the highest correlation with MC (µg/L) concentration were temperature (r = 0.19), NO_3^- (r = 0.09) and conductivity (r = 0.07). However, the scatterplots showed a strong nonlinear increase in MC with high levels of conductivity and with low levels of DO, despite having smaller linear correlation with MC than temperature, and NO_3^-, and in the case of DO, smaller than PO_4^{3-} (r = 0.06). Comparatively, these correlations are relatively low, as the maximum can be 1.0. Therefore, this BN was knowledge-based instead of data-driven, with two arcs learned from data. The correlation matrix was taken into account while giving background knowledge, connecting the arcs based on linear and nonlinear correlation as well as a literature review. For example, the highest correlation through the entire matrix is the linear correlation between conductivity and temperature (r = 0.4), and temperature with PO_4^{3-} (r = -0.38). Therefore connecting temperature and PO_4^{3-} with conductivity in the BN will allow information to be translated through CPTs to the end node of MC. Due to the low levels of correlation, it does not make sense to revise the possible set of predictor values and therefore based on this analysis all parameters were left in the dataset.

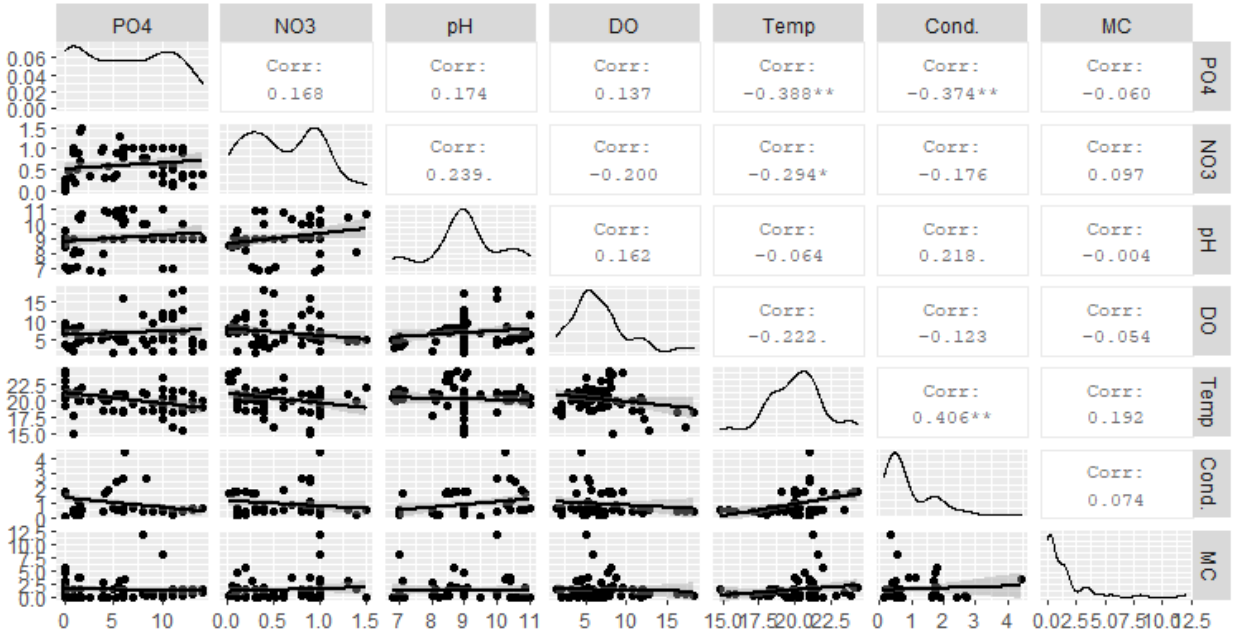


Figure 11. Figure 5. Correlation analysis of possible variables for Mexican Lentic BN, with regards to variables of interest being MC. Upper panel corresponds to the correlation coefficient (r) of each pair. The lower panel shows the linear regression curve of the scatter plot. Abbreviations are as follows: MC (microcystin), PO4 = phosphates, NO3 = nitrates, DO = dissolved oxygen, Temp. = temperature (ºC), Cond. = conductivity

The eight lentic systems used to create the Mexican BN are grouped in the PCA (Annex 4). The systems are grouped together based on which lentic system they came from, and in the case of Lake Zumpango which sampling site. Only 47% of the variance of the systems can be described by the first two PCs, which is relatively low for recognizing patterns. PC1 is most associated with temperature, PO_4^{3-}, and conductivity meaning these three variables describe 28.3% of the variability of the systems. PC2 is related to NO_3^-, and DO, meaning these two variables make up 19% of the total variance of the system. This analysis shows none of the variables are redundant and therefore all would be beneficial to include.

-

4.2.2.1 Mexican Lentic Systems BN

Similar to the Spanish lentic system, each node has an associated CPT table which lays out the discrete probability distributions across the states of the two parent nodes, DO and conductivity. Table 7 portrays the first 12 rows of the node of interest MC. The entire CPT contains 36 rows, due to having two parent nodes each with six discrete intervals. The CPT is calculated from Equations 1 and 2 from section 1.4.

Table 7. Conditional Probability Table for the node of interest “Microcystins”. Each column represents the probability distribution of MC based on the given states of the parent nodes dissolved oxygen and conductivity. Abbreviations are DO = dissolved oxygen, Cond = Conductivity.

DO	below 4.3 mg/L						between 4.3 to 6.4 mg/L					
Cond mS/cm	<0.3	0.3 - 0.8	0.8 - 1.3	1.3 - 2.1	2.1 - 3.5	>3.5	<0.3	0.3 - 0.8	0.8 - 1.3	1.3 - 2.1	2.1 - 3.5	>3.5
MC µg/L												
<1.0	0.6	0.57	0.8	0.1	0.6	0.2	0.4	0.47	0.55	0.6	0.73	0.1
1 - 4	0.1	0.35	0.05	0.6	0.1	0.2	0.4	0.32	0.3	0.1	0.06	0.6
4 - 6.8	0.1	0.02	0.05	0.1	0.1	0.2	0.06	0.01	0.05	0.1	0.06	0.1
6.8 - 10	0.1	0.02	0.05	0.1	0.1	0.2	0.06	0.09	0.05	0.1	0.06	0.1
>10	0.1	0.02	0.05	0.1	0.1	0.2	0.06	0.09	0.05	0.1	0.06	0.1

The BN found for Mexican Lentic systems is displayed in Figure 12. The software ran 18 iterations, with the best score being -992.92, with a log likelihood score of -643.9. The model was run on 65 samples, with a link probability of 0.1 and a prior link probability of 0.001. These scores are higher than Spanish lentic systems due to the increased number of parameters. The strength of the influence is outlined by the thickness of the arcs. This strength is important to see for modifying parameters and for the testing phase. It is calculated from the CPT of the child node (conditional on the parent node’s state) using average Euclidean distance. The strongest connection is between month and season (0.68), with the second strongest being between conductivity and pH (0.46). The strength between the parent nodes (conductivity and DO) and the end node of interest (MC) is 0.23 and 0.22 respectively, having a maximum possible value of 1. MC was discretized into five categories due to the distribution of the data, and five intervals led to a better log likelihood score.

Two “what if” scenarios were run to observe the predictive ability of the model, or the predicted probability distributions based on potential future events. Figure 12 illustrates the relationship between the concentration of MC based on the effects of stressors of cyanoHAB. Evidence was set by inputting the cost-effective variables that are able to be measured by a single probe (for example then H198199 Hanna meter). The month of July was set for each situation, as well as two scenarios a) temperature (19.6 - 22.4 ^{o}C) and conductivity at it;s lowest interval <0.3 mS/cm and b) both parameters are increased to the next highest discretization bin which is temperature >22.4^{o}C and conductivity 0.3 - 0.8 mS/cm.

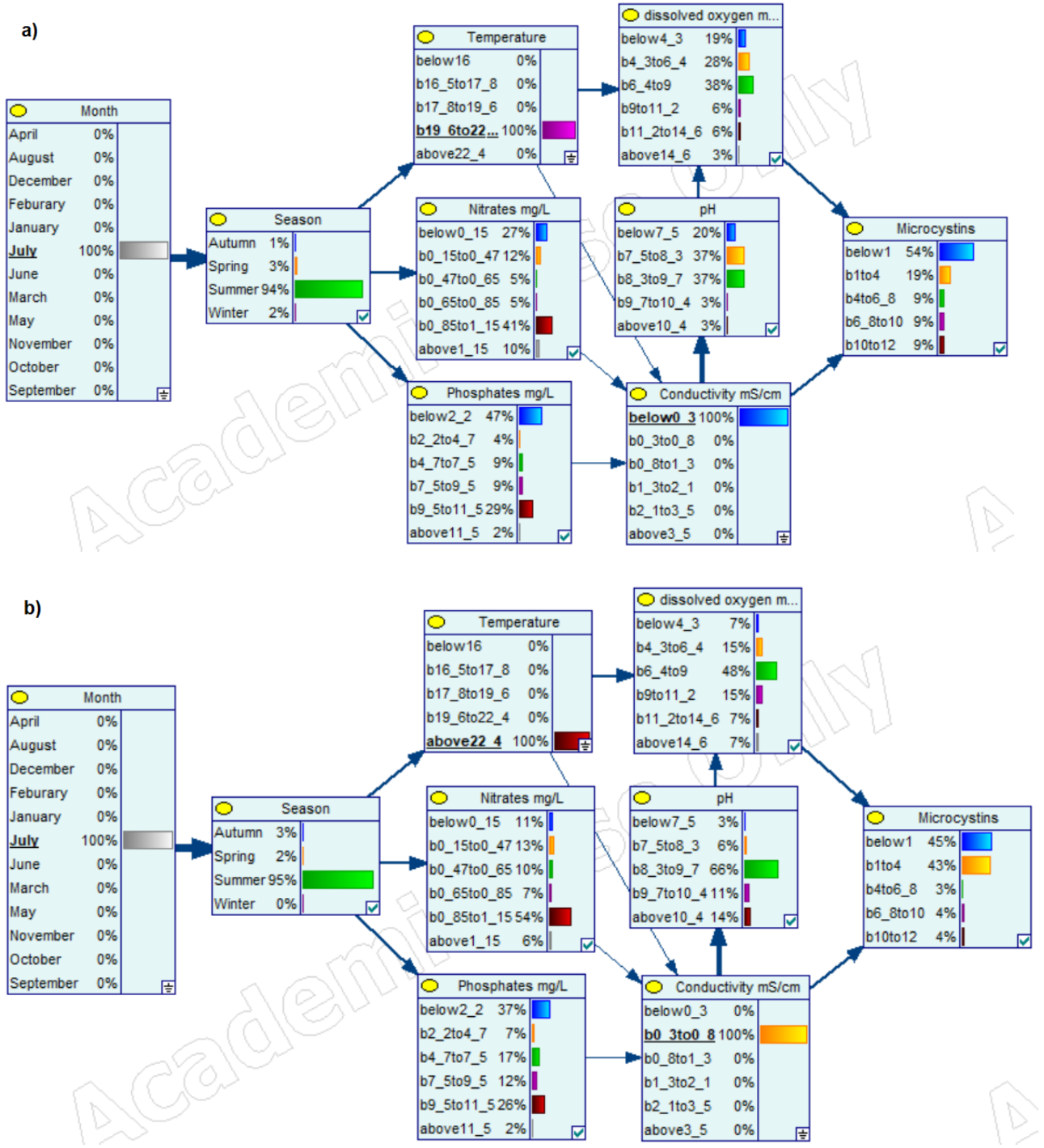


Figure 12. BN model structure for predictive MC concentrations for 9 lentic systems in Mexico. The node states consist of the either below: meaning less than, b __ to ___ : meaning between a value to another value, or above: meaning more than. Decimals are replaced by underscores as demanded by the software (example: 2.2 is 2_2 in the model). The thicker arrows show the strength of influence between nodes.. Screenshots of dynamic scenario a) warm water temp with low conductivity b) highest water temp with next highest conductivity

The predicted probability distribution for MC in Figure 12a shows that with colder summer weather and lowest conductivity, there's a 54% probability of having MC concentrations within the WHO recommended threshold for low risk, and within the range for potability (<1 µg MC/L). When

temperature and conductivity increase the probability distribution of MC shifts from 54% to 45% probability of being <1 µg/L, while the predicted probability distribution of MC from 1 -4 µg/L (WHO recommended low risk) increases from 19% to 43% showing the combination of high conductivity and high water temperature was linked to higher levels of MC. This has been studied extensively in literature, with the results supporting each other. High conductivity has been said to be related to persistent low flow conditions, whereas DO can be an indicator of photosynthetic rates, eutrophication or even of past algal blooms.

4.2.2.2 Validation of the Bayesian Network for Mexican Lentic Systems

A sensitivity analysis was performed to validate and evaluate the Mexican lentic systems BN by analyzing how small changes in the explanatory parameters can affect the target probability of interest (or the conditional probabilities). The end node of the cyanotoxin µg/L MC was identified as the “target node” and the parameters that most affect the CPT are highlighted in darker red. As the red becomes lighter, those parameters affect the posterior probabilities of MC less (Figure 13). The sensitivity coefficient of MC was lower than its Spanish counterpart at an average of 0.006 and a maximum of 0.08 (in the CPT of DO 6.4 - 9 mg/L and conductivity being 0.3 - 0.8 mS/cm). This indicates that the conditional probabilities of MC are not affected greatly by small changes in the input variables but of the parameters, MC concentration is more sensitive to small changes in DO. Conductivity, pH, and temperature were comparatively similar in the degree of which they affected MC. Lastly, nutrients PO_4^{3-} and NO_3^- and month were the last parameters to have moderate effect of MC, with season being the least influential. Based on these results, for elicitation from a sensitivity analysis, season could be removed from the BN.

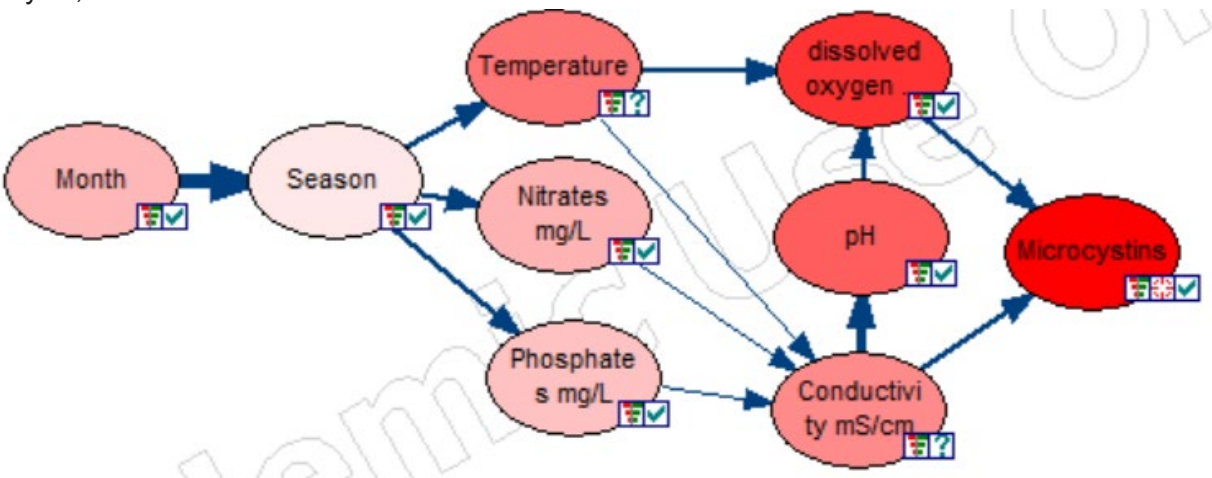


Figure 13. Mexican Lentic BN sensitivity analysis. Dark red implies greater sensitivity on the target node, with lighter colors meaning less.

To verify the predictive qualities of the model, and to ensure optimal structure, a CCI was performed on 20% of the 65 data points from Mexican lentic systems (13 samples). Input evidence was set for each node, with the only parameter being inferred being MC µg/L. Twelve of the 13 testing sites were predicted accurately. The misclassified site, Lake Zumpango in October, had the highest concentration of MC out of the entire data set with 12 µg/L, which was misclassified to the interval 1.0 - 4.0 µg/L (48% probability). The network was relatively better at predicting lower amounts of MC rather than high concentrations. The CCI was calculated at 0.92, which would be considered a good model according to Shan *et al.,* (2019).

4.2.3 Statistical Analysis of Spanish Lotic Systems

Due to a limited amount of data to yield a data-based BN for Spanish lotic systems, a correlation matrix was implemented on the key explanatory variables for cyanobacteria (Figure 14) to determine diagnostic relationships between the variables. The parameters with the highest linear correlation with concentration of ANA were DIN (r = 0.63) and conductivity (r = 0.45). Presence/absence of the biosynthetic gene cluster for ANA-producing cyanobacteria known as ANA also considered as an end node, as this can tell the potential of ANA production. The variables most correlated with the presence of the gene were biofilm size (r = 0.35), depth (r = - 0.27) and DIN (r = 0.21). There were no non-linear relationships found. All variables appear to have relatively strong correlation with other variables, or the variables of interest, and therefore the revised variable set of predictive parameters for the lotic BN included all.

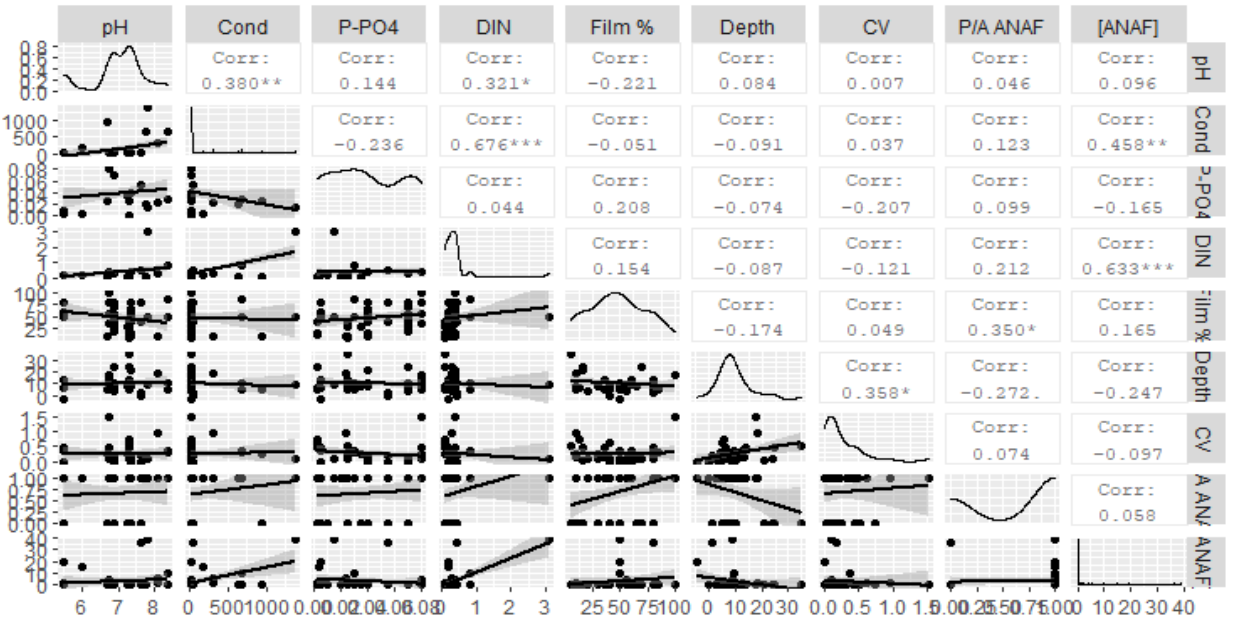


Figure 14. Correlation matrix for exploratory analysis of variables for the BN for Spanish lotic systems, with emphasis on variables of interest being P/A ANAF, and concentration of ANA. Upper panel corresponds to the correlation coefficient (r) of each pair. The lower panel shows the linear regression curve of the scatter plot. Abbreviations are as follows: Cond = conductivity, $P\text{-}PO_4^{3-}$ = Soluble reactive phosphorus, DIN = dissolved inorganic nitrogen, Film % = Biofilm %, CV = Current velocity, P/A ANAF = Presence/absence of ANA

A PCA was carried out to look for redundancy in the data, to decrease the number of variables in the BN due to the lower quantity of data. The first two PCs accounted for 48.75% of the variance of the seven rivers. The first PC was dominated by DIN, conductivity, and concentration of ANA. The second PC was most associated with depth, presence/absence of ANAF, and biofilm size. According to the PCA, depth and CV overlap in describing the variance of the data, with depth being moderately more significant due to a longer arrow, and therefore one could be removed from the dataset without losing too much data. Furthermore, pH was not important in the first two PC structure creations describing the systems, and therefore removing this parameter could be considered for elicitation sake.

4.2.3.1 Bayesian Network for Spanish Lotic Systems

Spanish lotic system was a hybrid network, created by using background knowledge for six arcs, and the remaining eight arcs were determined through data. The CPT of the final end node ANA µgANA/mg dw is outlined in Table 8**.** This outlines the first 12 columns (out of 36) of the probabilistic dependencies between the child node and the two parent nodes: conductivity and DIN. This table is needed to run the model, and will calculate the probability distribution, given the states of the parents.

Table 8. Examples of Conditional Probability Table for end node Anatoxin-a µgANA/mg dw. Each column expresses the probability distribution of the child node given certain states of the parent nodes. The table contains the first 12 columns / out of 36

Cond.	<105 µS/cm						between 105 - 273 µS/cm					
DIN (mg/L)	<0.14	1.4 - 0.22	0.22 - 0.35	0.35 - 0.63	0.63 - 1.9	>1,9	,<0.14	1.4 - 0.22	0.22 - 0.35	0.35 - 0.63	0.63 - 1.9	>1.9
[ANAF] µgANA/mg dw												
<1.0	0.74	0.17	0.94	0.07	0.16	0.16	0.69	0.08	0.93	0.82	0.16	0.16
1 - 3	0.16	0.4	0.01	0.01	0.16	0.16	0.19	0.08	0.01	0.01	0.16	0.16
3 - 6	0.02	0.02	0.02	0.82	0.16	0.16	0.02	0.08	0.01	0.07	0.16	0.16
6 - 17	0.02	0.02	0.01	0.01	0.17	0.17	0.02	0.58	0.01	0.01	0.17	0.17
17 - 27	0.02	0.35	0.01	0.01	0.17	0.17	0.02	0.09	0.01	0.01	0.17	0.17
>27	0.02	0.02	0.01	0.07	0.16	0.16	0.02	0.08	0.01	0.07	0.16	0.16

The Spanish lotic system BN (Figure 15) had a score of -706.66 which was the best score out of 18 iterations performed. The log likelihood was -391.6. The link probability between nodes is 0.1, with the prior link probability being 0.001. The strongest connection of influence is between the node season and DIN with a value of 0.47, and the second strongest being season with SRP (0.46), and pH with DIN (0.43).The strength of influence on the end nodes ANA µgANA/mg dw were lower than the previous BN target nodes with a value of 0.19 (conductivity with ANA), and 0.17 (DIN with ANA). The presence of ANA had a slightly stronger influence with 0.25.

Two “what-if” scenarios were run to test the predictive ability of the BN for evaluation. Evidence was set (100% probability of seeing the value because set) based on the parameters with the most influence, and parameters that increase as eutrophication increases. DIN and conductivity were both placed at the a) lowest intervals to represent pristine waters and b) highest intervals representing pollution to gauge the effect on the remaining variables.

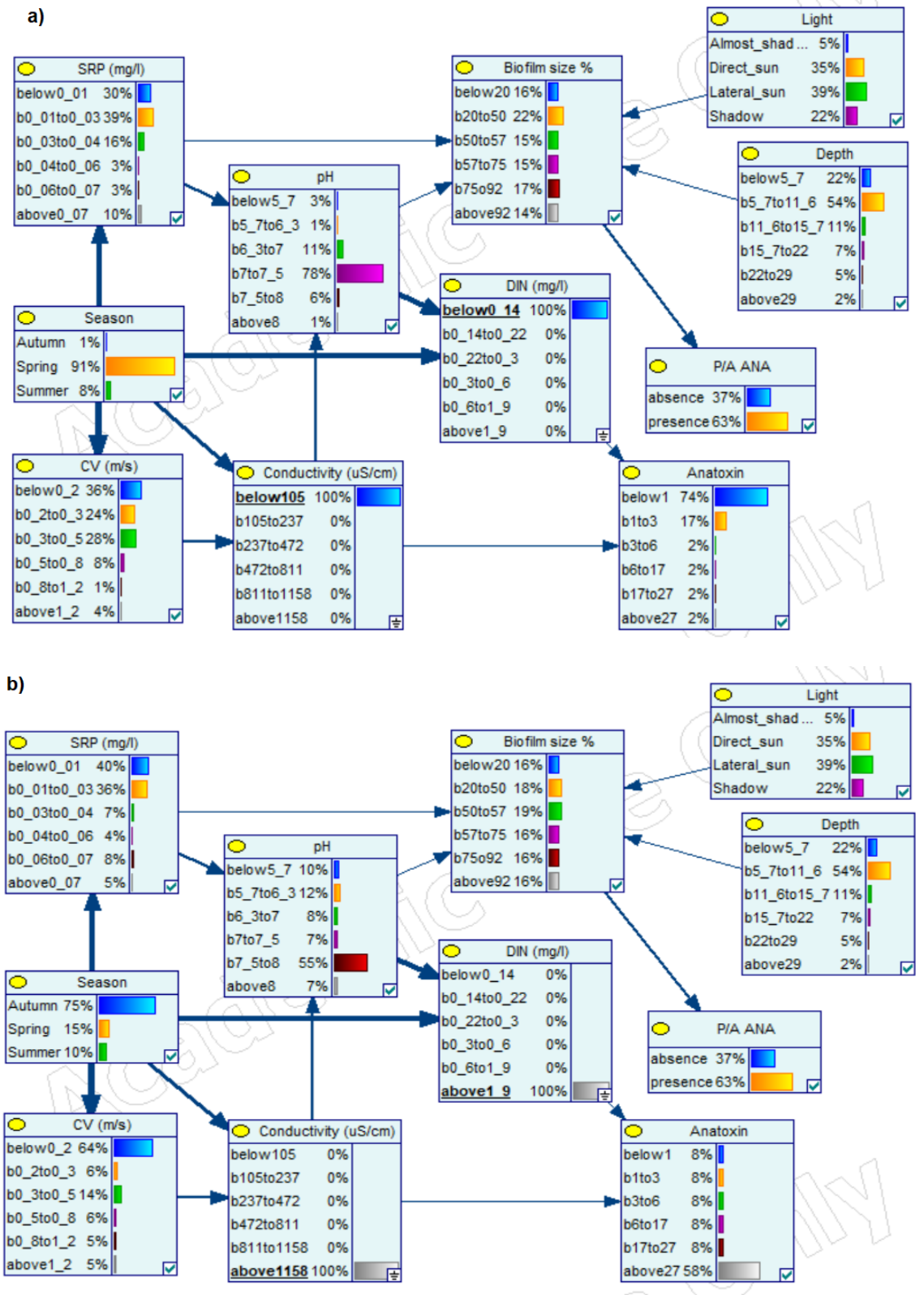


Figure 15. BN model structure for predictive ANA concentrations for 10 rivers in Spain. The node states consist of the either below: meaning less than, b __ to ___ : meaning between a value to another value, or above: meaning more than. The thicker arrows show the strength of influence between nodes. Screenshots of what-if scenarios eutrophication predictors conductivity and DIN evidence being set a) lowest intervals b) The highest discretization levels were set

Figure 15a shows the posterior probability of ANA being 74% of having a concentration <1 µgANA/mg dw, with a 38% probability of having biofilm size <50%, and a 52% probability of having CV between 0.2 - 0.5 m/s. Figure 15b portrayed more eutrophicated waters where the posterior probability of ANA increased from <1 µgANA/mg dw to a 58% probability of being >27 µgANA/mg dw. CV also decreased in speed to a probability of 64% being <0.2 m/s, and the highest biofilm probability being between 50 - 57% in size. Although there is no proposed legislation for recreational waters for ANAF concentration, this value can be considered high due to being substantially over the recommended drinking water threshold of 6 µg/L. Moreover, the benthic recommended guidelines for biofilm state that above 50% is high risk for recreational waters, in conjunction with river flow. In addition, if the evidence is reset as only the CV node in the fastest interval (>1.2 m/s), the posterior probability of ANA decreases from 27 µgANA/mg dw (58%) to below 1.0 µgANA/mg dw (32%).

4.2.3.2 Validation of Spanish Lotic System Bayesian Network

To perform the sensitivity analysis, the presence of ANA and ANA µgANA/mg dw were selected as target nodes. The sensitivity coefficients were 0.08, and 0.005 respectively. Based on the CPT of ANA µgANA/mg dw, the node is most sensitive to changes in conductivity, specifically in the ranges of 9.8 - 170 µS/cm and DIN below 0.63 mg/L. The analysis is outlined in Figure 16. The darker red nodes represent parameters that affect the target nodes more intensely (small changes leading to different posterior probabilities), which are season, CV, and conductivity in terms of Spanish lotic systems. DIN is the second, along with SRP and pH. Biofilm size and depth are the least sensitive, although this could be due to the structure of the BN, as usually biofilm is quoted for risk of benthic populations. Based on these results, the only parameter that could be removed for elicitation is light, however since this is an easy variable to measure, it was kept as a part of the BN.

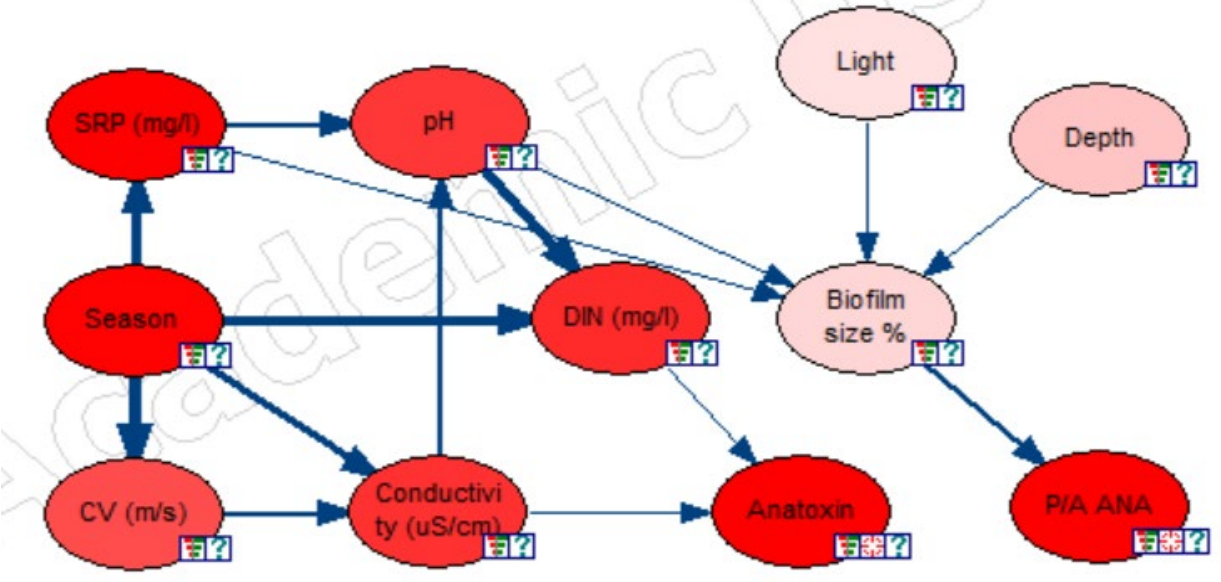


Figure 16. Sensitivity analysis of Spanish lotic systems BN: darker red indicates more influence of node of interest, lighter is less

Lastly, 10 sites out of the 46 data points were used to calculate CCI (20% of the data). Two parameters were used in order to classify the site as “correct” which were ANAF µgANA/mg dw and

presence of ANA. Eight of the 10 sites were completely categorized correctly. Of the two incorrectly identified sites, neither were completely incorrect. The first site in the Mediano river correctly predicted the presence of ANA while incorrectly placing the concentration as 1 - 3 µgANA/mg dw, when it was 3.8 µgANA/mg dw. The second site also in the Mediano river correctly predicted the concentration of ANAF as 0 µgANA/mg dw, but incorrectly classified the system as having ANA present. The Spanish lotic system was moderately better at correctly predicting the presence of ANA in place of absence. Overall, this led to a CCI of 0.8, which would be associated with a 'good model'.

5. Discussion

5.1 Risk Assessment Recommendations for Spanish Lentic Systems

Understanding country-specific (or even site-specific) mechanisms that can cause cyanoHABs or MC threshold concentration violations are important for determining water quality, risk assessment, and ultimately water management processes. Moe *et al.,* (2019) outlines the expectation of more frequently occurring cyanoHABs, and how the development of these site-specific management plans can help thwart and adapt to future predicted changes. Due to variable response of MC and casual relationships between environmental factors and MC concentration, a pragmatic approach based on probability and predictive models with uncertainty analysis can give insight into potentially important predictor variables and the related cyanotoxin. To determine this, the final node's (MC) evidence was set for each of the corresponding risk thresholds outlined by WHO (2003) (low 2-10 µgMC/L; 10-20 µgMC/L moderate; and >20 µgMC/L high)**,** then observing the environmental predictor approximate thresholds for each risk level. Values were only considered if they had over 50% probability in each predicting node. The potentially predictive factors are outlined in Table 9. Of the 88 sampling sites studied in this project, there were no locations with MC concentrations ranging from 10 - 20 µg/L, meaning the BN could not give a probability distribution of more than 50% for any of the predicting factors. This moderate risk range was therefore determined by interpolating between the values calculated for low risk and high risk.

Table 9. Recommended thresholds for environmental factors that drive risk levels recommended by WHO (2003) for safe recreational waters when applied to Spanish lentic systems. Color coding: blue is low risk for recreational waters but falls within the WHO guideline for drinking water, green is low risk, yellow is moderate, and red is high risk. Respective probability distributions in parentheses

Guideline values for Risk	NO_3^- mg/L	PO_4^{3-} mg/L	Chl*a*Total µg/L	Chl*a*Cyano µg/L	Maxdom of Cyanobacteria
0.0 - 0.01 µg/L	<0.9 (67%)	<0.1 (78%)	<5 (86%)	<4 (92%)	<27% (66%)
0.01 - 1 µg/L	<0.9 (50%)	0.1 - 0.4 (50%)	10 - 20 (65%)	4 - 15 (84%)	<27% (58%)
1 - 10 µg/L	0.9 - 1.6 (54%)	<0.1 (65%)	20 - 44 (71%)	15 - 39 (61%)	>84% (59%)
10 - 20 µg/L	<0.9	<0.4	44 - 100	39 - 67	>84
>20 µg/L	<0.9 (54%)	0.1 - 0.4 (58%)	>100 (55%)	>67 (61%)	70 - 84 (62%)

By inputting the values (Table 9) into the BN, the corresponding observed probability distribution for MC is as follows: 95% probability the input variables for the guideline values of 0.0 -

0.1 µg/L will be in this range, 91% probability the input variables for the guideline values of 0.01 - 1 µg/L will be in this MC interval, 85% probability the input factors for the guidelines values of 1 - 10 µg/L will be in this MC range, 30 - 34% probability the input values of the guideline thresholds of 10 - 20 µg/L will be in this MC bin, 93% probability the input variables of the guideline thresholds of >20 µg/L will be in this MC range. The lowest probability is associated with the bin that had zero corresponding data in the dataset, with the range being for different intervals of Chl*a*Total (44 - 56 µg/L is 30%, while 56 - 100 µg/L is 34%). Using the Bayesian model to identify the critical levels of predictor variables that will exceed the threshold of MC low risk, the results indicate that Chl*a*Total concentrations ≥ 44 µg/L , Chl*a*Cyano concentrations ≥39 µg/L, Maxdom ≥70% will significantly increase the probability of reaching moderate levels of MC. NO_3^- and PO_4^{3-} concentrations by themselves failed to accurately predict the surpassing of the different levels of risk, which coincides with Kelly *et al.,* (2019) and Shimoda *et al.,* (2016).

The core parameters that are most accurate at predicting MC risk concentrations are therefore Chl*a*Total, Chl*a*Cyano and maximum dominance of cyanobacteria. It's long been established that Chl*a*Total concentration is a strong predictor of MC concentrations, as it is used as an indicator of phytoplankton biomass (Kelly *et al.,* 2019). The exceedance of low risk thresholds found for Chl*a*Total concentrations closely resemble concentrations found in other research. Hollister and Kreakie 2016 found that when Chl*a*Total concentration reached 68 and 104 µg/L then there was a 50% chance of surpassing the health advisory levels of 1 and 2 µg/L respectively. Yuan *et al.,* (2014) concluded that when TIN levels were low (around 570 µg/L) then a Chl*a*Total concentration of 37 µg/L caused a 10% chance of finding a MC concentration about 1 µg/L.

Maximum dominance had a significant increase from interval 0.01 - 1.0 µgMC/L and then to 1.0 - 10 µgMC/L (<27% to >84% respectively), then decreased to 70 - 84% for the interval >20 µgMC/L. There is a degree of uncertainty in the prediction statistically due to the observed frequency in the data set of sampled with MC concentrations (13). Of the four sites that observed Maxdom level of >84%, only two of the samples consisted of any concentration of MC. Alternatively, ecologically there is uncertainty as well as outlined in Agha *et al.* (2012), which describes the variability of toxicity (chemotypes) intra-bloom. Blooms are composed of both toxic and non-toxic strains in variables concentrations. Furthermore, the amount of toxin per dry weight of a cell is highly variable even within the same strain in response to biotic and abiotic factors.

Levels of PO_4^{3-} and NO_3^- do not vary in between levels of risk of proliferation with a probability distribution that remained almost the same. The sensitivity analysis for Spanish lentic waters showed the same results. It has been a continuous debate about the link between MC production and nutrient enrichment (Scott *et al.,* 2013). However, it's been highly researched that phosphorus and nitrogen are normally dominated by feedback processes (Guignard *et al.,* 2017) which are impossible to include in a traditional BN, as they cannot handle loops due to the acyclic nature of the graphical structure. This could be a disadvantage to using a BN in the application to analyze and predict potential harmful cyanobacteria in lentic systems in addition to the uncertainty surrounding nutrient loading and the relationship to MC.

The thresholds outlined by the BN for moderate risk (exceedance of 10 µg MC/L) differs from the values outlined in the current Spanish decision tree (Chl*a*Total concentrations ≥10 µg/L,

maximum dominance ≥20%), but are more closely related to the values found by conditional probability approaches outlined in literature (Shimoda *et al.,* 2016, Hollister and Kreakie, 2016, Yuan *et al.,* 2014). Furthermore, Kelly *et al.,* (2019) states that Chl*a* is used as an indicator of remediation in the Bay of Quinte, with a target concentration of 10 - 12 µg/L which indicators that the water has been restored. These values contrast with the value in the Spanish decision tree which at these values would indicate moderate risk.

Furthermore, moderate risk for Chl*a*Total is considerably higher than the thresholds laid out by WHO (2003), which is <50 µg/L with cyanobacterial dominance. This BN defined moderate risk as 44 - 100 µg/L due to the lack of data in the moderate threshold range which limits the confidence interval of the probability distribution. Therefore, it cannot be said whether the range is definitively between 44 - 56 µg/L or 56 - 100 µg/L and was therefore left at 44 - 100 µg/L. The BN was able to validate the risk level of the current Spanish decision tree for cyanobacterial proliferation for only moderate and high risk levels. The BN was unable to classify 90% of the sites as low risk although this level accounted for 21% of the total samples. The discrepancy was probably due to inadequacies of connecting the causal relationships between environmental factors and risk, in that in the current Spanish decision tree, if one parameter exceeds the threshold, the entire system is considered at a moderate risk of proliferation. The BN considers all factors interconnected and views the system's variability as a whole, and can have a difficult time creating the relationships without consistency, for example in Embalse Palmaces all predicting factors are well below low risk level, but however contains 0.024 µgMC/L which would classify it as high risk level in the current risk assessment decision tree. Nevertheless, the BN takes into account all nodes to construct a final risk node CPT without giving a single importance to only one node, leading to a mismatch between prediction and observation.

While Chl*a*Total, Chl*a*Cyano and maximum dominance are the best predictors for this sites-specific BN, which is shown through the BN and through the correlation matrix, other research done on conditional probability models indicate temperature as a driving factor for the promotion of MC concentrations (Kelly *et al.,* 2019, Shan *et al,* 2019, Davis *et al.,* 2009).This Low-cost parameter can be easily detected. As climate change increases the temperatures globally, there is a growing need to incorporate this variable into decision trees, as it can be representative of stratification and water column stability.

5.2 Risk Assessment Recommendations for Mexican Lentic Systems

Through Bayesian conditional probability techniques, the critical levels of the physicochemical parameters of the Mexican lentic systems in the State of Mexico that increase the probability of exceeding low levels of risk were identified to be NO_3^- concentrations 0.8 - 1.15 mg/L, PO_4^{3-}concentrations 7.5 - 9,5 mg/L, DO concentrations 4.3 - 6.4 mg/L, temperature 19 - 22ºC, conductivity ≥3.5 µS/cm, with a 55% probability of taking place in October. These thresholds reflect warmer weather with low flushing out rates, most likely during the decomposition of algal blooms. Arazate-Cardenas *et al.,* (2010) corroborate the pH and temperature intervals being most associated with the presence of *Microcystis* genus. Of these low-cost parameters, the core predictors that will have the highest likelihood of prediction for water management are conductivity, DO, and temperature. With values of conductivity ≥3.5 µS/cm, DO between 4.3 - 6.4 mg/L, and

temperature between 19 - 22ºC there is a 60% chance of having a MC concentration over 6.8 µg/L. Apart from having a direct effect on MC production, temperature plays a part in causation of concentration of DO and conductivity. Shan *et al.,* (2019) determines that ammonium in place of nitrate has a more important role in the prediction of MC. In the nitrogen cycle, it is known that the lower amount of DO generally increases the concentration of ammonium. Thus, DO can act as a proxy for the reduced form of nitrate. Conductivity is seen as an indicator of low flushing out rates of water bodies (Moe *et al.,* 2016, Elliot, 2010), which low flushing out rates and higher temperatures lead to an increase of pronounced cyanoHAB for prolonged periods of time (Kelly *et al.,*2019).

39 of the 65 samples were within the range <1 µg MC/L, quantifying this as the largest frequency of observed data (60% of the data). The frequency of observed MC concentration >10 µg/L included only one sample at 12 µg/L (making up 1.5% of the data), as well as only one sampling site between concentrations 6.8 – 10 µg MC/L (1.5% of the data), both in Lake Zumpango. This implies that the CPT for higher concentrations of MC are related to high levels of uncertainty. As a result, the probabilities of the physicochemical states did not greatly vary, therefore, 20 - 40% probability for predicting node values when the final end node of MC was set to the corresponding ranges were considered the most valid. There were no values over the designated high risk associated with the WHO recommendations and is therefore not included in Table 10, which outlines the recommended predicting physicochemical predicting factors for MC risk assessment in Mexico. Table 10 also includes the interval 0.0 - 0.01 µg MC/L as an indicator of no MC for comparison.

Table 10. Thresholds for environmental factors that drive risk levels recommended by WHO (2003) in Mexican lentic systems. Color coding: blue is low risk for recreational waters that's acceptable concentration for drinking water, green is low risk, yellow is moderate

Guideline Values for Risk	NO_3^- mg/L	PO_4^{3-} mg/L	Temp (ºC)	pH	Cond. µS/cm	DO mg/L
0 - 0.01 µg/L	0.85 - 1.15 (89%)	<2.2 (52%)	19.6 - 22.4 (89%)	>9.7 (50%)	0.8 - 1 (32%)	4.3 - 6.4 (45%)
< 1 µg/L	0.85 - 1.15 (32%)	<2.2 (32%)	19.6 - 22.4 (57%)	8.3 - 9.7 (37%)	0.3 - 0.8 (32%)	4.3 - 6.4 (30%)
1 - 4 µg/L	0.15 - 0.4 (31%)	<2.2 (30%)	19.6 - 22.4 (54%)	8.3- 9.7 (46%)	0.3 - 0.8 (40%)	6.4 - 9 (30%)
4 - 6.8 µg/L	0.15 - 0.4 (40%)	<2.2 (35%)	19.6 - 22.4 (44%)	8.3 - 9.7 (32%)	1.3 - 2.1 (30%)	6.4 - 9 (26%)
6.8 - 10 µg/L	0.85 - 1.15 (32%)	9.5 - 11.5 (32%)	19.6- 22.4 (43%)	9.7 - 10.4 (30%)	>3.5 (21%)	4.3 - 6.4 (22%)
10 - 12 µg/L	0.85 - 1.15 (32%)	7.5 - 9.5 (30%)	19.6 -22.4 (43%)	9.7 - 10.4 (30%)	>3.5 (21%)	4.3 - 6.4 (22%)

By inputting the values found in Table 10 into the Mexican Lentic BN, the corresponding probability of concentration of MC are as follows: a 47% probability of MC <1 µg MC/L with the recommended values of predicting nodes, 60% probability of MC being between 1 - 4 µg MC/L with the recommended thresholds, 22% probability of MC being between 4 - 6.8 µg MC/L with the recommended values, 30% probability of being between 6.8 - 10 µg MC/L with the recommended values, and a 30% probability of having between 10 - 12 µg MC/L with the corresponding values. The predicting variables showed weak changes in between ranges above 4 µg MC/L, as did the MC ranges showing weak response to changes in thresholds due to the limited amount of data, and the low frequency of samples above this value. Furthermore, all data found for Mexican recreational waters was found in the State of Mexico which lead to all samples having the same water temperature for each risk level, and generally, weak correlation between the environmental factors.

The Mexican dataset had the least number of water bodies, which is reflected in the low correlation due to less variability and the lowest certainty in probability. A larger monitoring program with sampling campaigns around Mexico would lead to a better performing model and more accurate statistical analysis, which in turn leads to more adequate risk assessment procedures. Arazarte-Cardenas *et al.,* (2010), Mercado-Borrayo (2008), and Figueroa-Sanchez *et al.,* (2020) all suggest that the same three core parameters for predicting MC recommended by this project are also the most correlated with phytoplanktonic growth and MC production.

5.2.1 Economic Analysis of Implementation in Mexico

CyanoHABs can affect not only human health and ecological integrity, but also economic opportunities for a country as well as economic losses such as: rendering a body of water unsuitable for swimming, fishing, and aquatic sports. These all affect tourism which can lead to long term costs. Monitoring recreational waters can have a high immediate cost but can mitigate the larger costs from restoration and the cost of closing down access to the water body. Cell counts of cyanobacteria can be used for monitoring but are time consuming and unsuitable for a large number of samples. Instrumental analysis through mass spectrometry can be expensive and not suitable for routine testing. A BN uses free software to investigate the conditional probabilities between physicochemical parameters such as conductivity, temperature, pH, and DO with concentrations of MC which would lead to risks which could be costly to the country or lead to dangers for human health.

The BN works more efficiently with more data from lentic bodies of water from all over the country. Using a YSI 556 probe (as was used in Figueroa-Sanchez *et al.,* 2020), which measures pH, DO, conductivity, temperature, and oxidation-reduction potential and has a single cost of around 1,000 euros, a BN can be created that would allow the prioritization of recreational waters with the highest probability of having moderate to high risk of MC to be tested using ELISA immunoassays, which is considered moderate cost. With this, water managers can focus on probable moderate to high risk water bodies using parsimonious variables.

5.3 Risk Assessment Recommendations for Spanish Lotic Systems

The Spanish Lotic BN was run by setting the evidence of the node of interest (µg ANA/mg dw of *Phormidium* mats) to the six discrete intervals. The intervals were determined by a literature review and translation of legislation in New Zealand for drinking water thresholds (provisional maximum acceptable values: Anatoxin-a 6 µg/L and Anatoxin-a(s) as 1 µg/L) as well as recreational water guidelines (three risk levels outlined by biofilm coverage <20%, 20-50%, and >50%) (Section 3. The guidelines | Ministry for the Environment, 2020). The predictive environmental nodes were considered valid if the probability ranged over 30%. The lower probability is due to the fact that the lotic dataset had the least amount of data (46 samples from benthic mats), with 78% of the samples being below 1 µg ANA/mg dw. As with the Mexican BN, this suggests that the CPT for ANA is linked with higher levels of uncertainty, especially with ranges 17 - 27 µg ANA/mg dw, as it only has an observed frequency of 2% (1 sample) in the dataset. Therefore, the probabilities of the predictive nodes do not differ immensely in between different states, and as a result 30% probability was sufficient for calculating the observed thresholds for each risk level in Table 11. Furthermore, ranges

3 to 6 µg ANA/mg dw, 6 to 17 µg ANA/mg dw, and >27 µg ANA/mg dw all only had two samples with observed corresponding concentrations. As a result, the BN had less certainty about the probabilities of the predictive environmental variables. Ranges were chosen based on which of the two probabilities was founded in literature in order to create the risk assessment.

Table 11. Recommended thresholds for environmental factors that drive risk levels recommended by provisional legislation outlined by New Zealand for Spanish lotic systems. color coding: blue is low risk and also acceptable for drinking water, red is high risk based on biofilm size

Guideline Values for Risk	DIN mg/L	SRP mg/L	pH	CV m/s	Cond. mS/cm	Biofilm Size
<1 µg ANA/mg dw	0.2 - 0.3 (36%)	0.6 - 0.7 (38%)	6.3 - 7 (41%)	<0.2 (64%)	<105 (91%)	20 - 50 (31%)
1 - 3 µg ANA/mg dw	0.1 - 0.2 (42%)	<0.01 (50%)	<5.7 (41%)	<0.2(35%)	<105 (70%)	20 - 50 (32%)
3 - 6 µg ANA/mg dw	0.3 - 0.6 (80%)	0.01 - 0.03 (42%)	6.3 - 7 (43%)	<0.2 (65%)	<105 (83%)	50 - 57 (31%)
6 - 17 µg ANA/mg dw	0.1 - 0.2 (33%)	0.01- 0.03 (54%)	5.7 - 6.3 (40%)	0.2 - 0.3 (32%)	105 - 237 (40%)	50 - 57 (36%)
17 - 27 µg ANA/mg dw	0.1 - 0.2 (51%)	<0.01 (49%)	<5.7 (41%)	<0.2 (35%)	<105 (60%)	50-57 (58%)
>27 µg ANA/mg dw	0.3 - 0.6 (37%)	<0.01 (27%)	7.5 - 8 (30%)	<0.2 (43%)	>1158 (42%)	75-92 (42%)

Depth and Light were not included in Table 11 as a consequence of not being necessarily predictive but more advantageous for cyanoHAB growth. There was no direct correlation between the end node of ANA and these two variables, so when ANA was set to the different intervals, depth and light did not alter their probabilities. Possibly, doubling depth with temperature to portray stratification could lead to more insights, as stated in Rigosi *et al.,* (2015). Everything above 3 µg ANA/mg dw was considered high risk for a more conservative risk assessment, although the provisional value in New Zealand states 6 µg ANA/mg dw (Ibelings *et al.,* 2014). This was decided based on the biofilm size of the *Phormidium* mat, with anything over 50% being considered likely to contain cyanotoxin and a proliferation of cyanobacteria. Inserting the corresponding probabilities from Table 11 into the BN, the predicted probability distribution of the end node of ANA went as follows: a 93% probability of inputting the values outlined for <1 µg ANA/mg dw will be in this range, a 40% probability of the associated values for 1 -3 µg ANA/mg dw will be in this range of ANA, an 82% probability that the values for 3 - 6 µg ANA/mg dw will be in this range of ANA, a 58% probability that the values for 6 - 17 µg ANA/mg dw will end up in this range of ANA, a 36% probability that the values for 17 - 27 µg ANA/mg dw will fall in this range, and a 58% probability that the values for >27 µg ANA/mg dw will be in this range of ANA.

Most of the environmental factors do not follow a clear pattern for increasing levels of ANA. DIN maintained around 0.1 to 0.6 mg/L for most ranges of ANA, but skyrocketed in the last range, as did conductivity; a possible reason explaining this could be most of the data from the *Phormidium* mats were obtained in oligio-mesotrophic systems. Data from other eutrophic rivers is highly recommended for a more comprehensive analysis in the future. Electrical conductivity has been shown to have significant influence on *Phormidium* cover in the river due to the fact that EC is influenced by flow rate variations, can represent persistent low flow conditions, or is a reflection

macro and micro nutrients that possibly influence cyanobacterial growth (McAllister *et al.,*2017, Wood *et al.,* 2017). In this case, the highest concentrations of ANA were found in Tajo River which overlays a geological calcareous part of the river, accounting for the high levels of conductivity in this area. Based on the correlation matrix and sensitivity analysis as well as Bayesian predictive techniques, the critical parameters that affect the concentration of ANA surpassing the recommended threshold for low risk are DIN (0.1 - 0.6 mg/L), SRP (0.01 - 0.03 mg/L), pH (5 - 9), and conductivity (35 - 1158 mS/cm). These predictor concentrations corroborate closely with those identified McAllister *et al.,* (2017), including the negative effect of observed *Phormidium* cover at higher concentrations of SRP, known as subsidy-stress concept. Furthermore, biofilm size coincided with results found in Wood *et al.,* (2017) where a positive relationship was found between ANA concentrations and cover %. However, these are related to a high degree of uncertainty, as the 46 samples were insufficient statistically with the number of parameters to identify any clear correlation between ANA concentration and environmental variables. This is reflected in the large discrepancies between the risk levels, due to some levels being defined by only two samples, i.e. a 50% probability between two discrete bins for environmental factors.

With this insufficient data, the only clear trend is that higher concentrations of ANA occur at more extreme levels of the environmental variables, for example, ANA levels 17 - 27 µg ANA/mg dw has a 41% probability of occurring with pH <5.7 with a 60% probability of being below 105 µS/cm. According to McAllister *et al.,* (2017), this is possibly due to competition with other benthic algae and that *Phormidium* can outcompete at these extreme levels. In the same paper, water temperature was determined to have an integral part in predicting *Phormidium* cover and ANA production. This variable would be beneficial to add to the lotic BN. The environmental variables had an interactive relationship as opposed to an additive effect, where if one node's evidence is set it will change the probabilities of all other nodes and could more adequately predict the concentration of ANA given the evidence of the other nodes, due to the combination of effects to reach a specific outcome.

5.4 Bayesian Network Applicability and Future Improvements

BNs offer a flexible system that can be applied to risk assessment tailored to each use of water. In the present project, recreational uses were mostly focused on due to a higher availability of risk data. However, within our dataset some of the lentic water bodies in both Spain and Mexico are used not just for recreational activities but also for drinking water supply. Our BN could easily be adapted to drinking waters by setting a different risk threshold, for instance based on the WHO guidelines for drinking water. The WHO outlines a guideline level of 1 µg/L of MC, which each of the lentic BNs contain in the node of interest MC discrete interval. The Lotic BN can also be adapted to drinking water levels, as it contains the provisional values for ANA levels that evaluate the quality of the system for parallel uses of the river (such as fishing). Therefore, each BN can adapt to drinking water by concentrating on the discrete interval of interest. For future analyses for risk assessment using BNs for drinking water, it is necessary to include official water treatment techniques applied to water supply locations, and how efficient these are at removal of cyanotoxins.

There are a few disadvantages related to BNs, which directly affect environmental application, such as the inability to incorporate feedback loops. Although BNs are able to handle a large amount of variables, the more variables there are, the more data is required to build the

structure of the network and estimate the parameters (ie, as the variables increase, the data must as well). In terms of this project, it directly hindered the structure learning of the network from Mexico, whose data is scarce due to a lack of raw data that was able to be found. Another drawback is that BNs have a more difficult time managing hybrid (discrete and continuous variables) or continuous variables. Even if it is possible to find an algorithm that can handle this type of data, the limitations are very restrictive. The solution to this, as was done in this project, is to discretize the variables (modifying continuous values into discrete ones) (Aguilera *et al.,* 2011). There are several ways in which to do this which include equal width or equal frequency binning, k-means, entropy minimization and hierarchical. This project focuses on hierarchical unsupervised discretization that involves clustering the number of data into the number of desired discrete variables (Vieira *et al.,* 2017). The discretization is made by using the maximums and minimum values of the data set, calculating the cluster centers and the midpoints between the clusters to produce the output (BayesFusion, 2020). An important point to emphasize is that the narrower the discretization band is, the less error will be added to the model, although it can make the model less organized, and have a much larger CPT.

However, Bayesian methods have gained momentum currently due to present trends of big data and artificial intelligence/machine learning. This modeling can be useful as it can handle large and heterogeneous data (from various sources of information including handling missing data), predict cause and effect relationships, infer linear and non-linear relationships between variables, and account for uncertainty. In addition to the many advantages, the model can also learn parameter values and avoid any reliance on a single node or deterministic outcome, as this is more representative of a natural environment. Because of this, BN models aid water managers to make realistic decisions of the probability of desired outcomes based on water quality strategies (Shan *et al.,* 2020).

Finally, future recommendations for risk assessment using BN, extensive data is needed in order to calculate clear correlation and trends for each level of risk. Large uncertainty was associated with the higher concentrations of cyanotoxins due to the low frequency of observed values in those ranges. To create a more successful model, optimal data collection would include multiple lentic bodies of water (for lentic BN) and multiple points along lotic systems (for lotic BN) with data spanning an entire year. Rigosi *et al.,* (2015) suggested physical chemical and biological variables being tested at least bimonthly, ideally for several years of observations. Multiple lakes would ensure the causal relationships between cyanotoxin concentration and environmental factors are not site-specific. In the case of the lotic systems BN, each sample in this project was associated with mats with potential for toxic cyanobacteria, with many of the samples being taken from the same place. Similar to the recommendations for lentic systems, increasing the lotic systems would create more effective BN modeling.

Secondly, discretization of variables in a model can lead to a loss of information. Algorithms to maintain continuous variables as that should be studied and researched to increase precision and efficiency.

Lastly, according to Moe *et al.,* (2019) the Water Framework Directive requires links between abiotic and biotic factors of an ecological system, which can be demonstrated in BN through cause

and effect arcs. In the next cycle of this directive (2021- 2026) it is required that potential impacts of climate change be incorporated into the river basin management plans. BN statistics and models can combine the expert knowledge of water managers with the probabilistic manner of BN models into predictive models given different climatic what-ifs. Incorporating biological variables with their physicochemical counterparts, determining through sensitivity analyses which variables affect the system most significantly, can give clear management decisions for achieving 'good status'.

6. Conclusion

This study found that in Spanish lentic systems, the critical variables for predicting MC were Chl*a*Total and Chl*a*Cyano, which were able to accurately predict observed concentrations in a majority of the 88 sampling sites. Mexican lentic systems core parameters that will help water managers predict MC levels were temperature, pH, DO, and conductivity in the absence of biological variables. Further investigation of the relationship between abiotic and biotic factors are required to identify stronger correlation to MC concentrations. Spanish lotic systems' ANA concentrations were influenced by time of year, DIN, conductivity, SRP, and pH, although further investigation and sampling is needed to produce a more robust model. The creation and organization of probabilistic water quality criteria for water quality management can be an effective way to cope with the levels of uncertainty about synergistic relationships in natural ecosystems. With the outlined biological and physical conditions examined in the BNs, more realistic (and statistically supported) water quality standards can be developed to decrease chances of false conclusions while still effectively monitoring risk.

7. References

Agha, R., Cirés, S., Wörmer, L., Domínguez, J. A., & Quesada, A. (2012). Multi-scale strategies for the monitoring of freshwater cyanobacteria: Reducing the sources of uncertainty. *water research*, *46*(9), 3043-3053.

Aguilera, P.A., Fernández, A., Fernández, R., Rumí, R. & Salmerón, A. (2011). Bayesian networks in environmental modelling*. Environmental Modelling and Software*, 26(12): 1376-1388.

Arzate Cardenas, Mario Alberto. (2008). Detección de cianobacterias toxigénicas pertenecientes al género *Microcystis* mediante marcadores moleculares y ensayos biológicos. Tesis Maestría en Ciencias Químico Biológicas. Instituto Politécnico Nacional

Arzate-Cárdenas, M.A., Olvera-Ramírez, R. & Martínez-Jerónimo, F. (2010). Microcystis toxigenic strains in urban lakes: A case of study in Mexico City. *Ecotoxicology*, 19(6): 1157-1165.

APHA. (2005). Standard Methods for the Examination of Water and Wastewater, 18th Edition, American Public Health Association, Washington D.C

BayesFusion, LLC (2020). Genie Modeler: User Manual. *Version 3.0.R1*

Bernard, C., Ballot, A., Thomazeau, S., Maloufi, S., Furey, A., Mankiewicz-Boczek, J., Pawlik-Skowrońska, B., Capelli, C., Salmaso, N., (2017). Appendix 2: Cyanobacteria Associated With the Production of Cyanotoxins, in: Handbook of Cyanobacterial Monitoring and Cyanotoxin Analysis. John Wiley & Sons, Ltd, Chichester, UK, pp. 501–525.

Bláha, L., Babica, P., Maršálek, B., (2009). Toxins produced in cyanobacterial water blooms - toxicity and risks. *Interdiscip. Toxicol.* 2.

Burch, M.D. (2008). Effective doses, guidelines & regulations. *Advances in Experimental Medicine and Biology*, 619: 831-853.

Cirés, S.; Wörmer, L.; Carrasco, D.; Quesada, A. Sedimentation Patterns of Toxin-Producing *Microcystis* Morphospecies in Freshwater Reservoirs. *Toxins* 2013, *5*, 939-957

Clauset, Aaron., (2011). Inference, Models and Simulation for Complex Systems.Sante Fe Institute. *CSCI 7000-001.* Lecture 5

Dash, R., Paramguru, R.L., Dash, R., (2011). Comparative Analysis of Supervised and Unsupervised Discretization Techniques. *International Journal of Advances in Science and Technology.* Vol. 2 No. 3, 2011.

Davis, T.W., Berry, D.L., Boyer, G.L. & Gobler, C.J. (2009). The effects of temperature and nutrients on the growth and dynamics of toxic and non-toxic strains of Microcystis during cyanobacteria blooms. *Harmful Algae*, 8(5): 715-725.

Diario Oficial de la Federación (DOF) Norma Oficial Mexicana Proy-Nom-127-SSA1-2017 (2017) Agua para uso y consumo humano. Límites máximos permisibles de la calidad del agua, control, vigilancia de los sistemas de abastecimiento. Secretaria de Salubridad y Asistencia. México.

Durai, P., Batool, M. and Choi, S., (2015). Structure and Effects of Cyanobacterial Lipopolysaccharides. *Marine Drugs*, 13(7), pp.4217-4230.

Bathing Waters Directive (2009). *Bathing Water Profiles: Best Practices and Guidance*. [online] Available at: <https://ec.europa.eu/environment/archives/water/report2011/profiles_dec_2009.pdf>

Egmont-Petersen, M., Feelders, A. & Baesens, B. (2005). Confidence intervals for probabilistic network classifiers. *Computational Statistics and Data Analysis*, 49(4): 998-1019.

Vieira, Elamara Marama de, Araujo, Jonhatan Magno Norte, Da Silva & Luiz Bueno, D.S. (2017). Modeling Bayesian Networks from a conceptual framework for occupational risk analysis. *Produção : Uma Publicação Da Associação Brasileira De Engenharia De Produção*, 27(0)

Elliott, J., (2010). The seasonal sensitivity of Cyanobacteria and other phytoplankton to changes in flushing rate and water temperature. *Global Change Biology*, 16(2), pp.864-876.

Environmental Protection Agency (EPA), (2015). *Algal Toxin Risk Assessment and Management Strategic Plan for Drinking Water.* Strategy Submitted to Congress to Meet Requirements of P.L. 114-45.

Environmental Protection Agency (EPA), (2014). *Cyanobacteria and Cyanotoxins: Information for Drinking Water Systems.* Office of Water- EPA 810F11001

Figueroa-Sánchez, M., Nandini, S. and Sarma, S., (2020). Zooplankton community structure in relation to microcystins in the eutrophic Lake Zumpango (State of Mexico). *Fundamental and Applied Limnology / Archiv für Hydrobiologie*, 193(3), pp.213-225.

Galanti, Tomer. (2015) Probabilistic Graphical Models: Parameter Estimation. Malaysian Communications and Multimedia Commission

Gales, Mark. (2005) Statistical Pattern Processing and Graphical Models and Bayesian Networks. University of Cambridge, Engineering Part IIB and EIST Part II

Guignard, M.S., Leitch, A.R., Acquisti, C., Eizaguirre, C., Elser, J.J., Hessen, D.O., Jeyasingh, P.D., Neiman, M., Richardson, A.E., Soltis, P.S., Soltis, D.E., Stevens, C.J., Trimmer, M., Weider, L.J., Woodward, G. & Leitch, I.J. (2017). Impacts of Nitrogen and Phosphorus: From Genomes to Natural Ecosystems and Agriculture. *Frontiers in Ecology and Evolution*, 5

Haya, R., (2016). Assessment of Anatoxin-a of Benthic Communities from Rivers. Effects on Macroinvertebrates. Final master project, Sciences Faculty, Biology Department, UAM

Hollister, J.W., Kreakie, B.J., (2016). Associations between chlorophyll a and various microcystin-LR health advisory concentrations. F1000Research 5, 151.

Huber, V., Wagner, C., Gerten, D. & Adrian, R. (2012). To bloom or not to bloom: contrasting responses of cyanobacteria to recent heat waves explained by critical thresholds of abiotic drivers. *Oecologia*, 169(1): 245-256.

Ibelings, B.W., Backer, L.C., Kardinaal, W.E. & Chorus, I. (2014). Current approaches to cyanotoxin risk assessment and risk management around the globe. *Harmful Algae*, 40: 63-74.

Kelly, N.E., Javed, A., Shimoda, Y., Zastepa, A., Watson, S., Mugalingam, S. & Arhonditsis, G.B. (2019). A Bayesian risk assessment framework for microcystin violations of drinking water and recreational standards in the Bay of Quinte, Lake Ontario, Canada. *Water Research (Oxford)*, 162: 288-301

Kjærulff, Uffe & Linda C. van der Gaag (2000). Making Sensitivity Analysis Computationally Ef icient. Proceedings of the Sixteenth Annual Conference on Uncertainty in Artificial Intelligence (UAI 2000), pages 317-325

Komárek J, Anagnostidis K (2005) Cyanoprocaryota 2. Teil: Oscillatoriales. In: Sußwasserflora Von mitteleuropa 19/2. Elsevier Spektrum, Heidelberg, p 759

Konig, S, 2013. Polyphasic anatoxin-a and homoanatoxin-a screening of benthic cyanobacteria communities of the Manzanares river and Mediano stream (Spain): Firstreport of the potential benthic anatoxin-a producers in Spanish riverine systems. Bachelor thesis. Universidad Autónoma de Madrid

Korb, K.B., Nicholson, A.E., (2003). Bayesian Artificial Intelligence. Chapman & Hall

Mantzouki, E., Lürling, M., Fastner, J., de Senerpont Domis, L., Wilk-Woźniak, E., Koreivienė, J., Seelen, L., Teurlincx, S., Verstijnen, Y., Krztoń, W., Walusiak, E., Karosienė, J., Kasperovičienė, J., Savadova, K., Vitonytė, I., Cillero-Castro, C., *et al.* (2018). Temperature Effects Explain Continental Scale Distribution of Cyanobacterial Toxins. *Toxins*, 10(4)

McAllister, T.G., Wood, S.A., Atalah, J. & Hawes, I. (2018). Spatiotemporal dynamics of Phormidium cover and anatoxin concentrations in eight New Zealand rivers with contrasting nutrient and flow regimes. *The Science of the Total Environment*, 612: 71-80.

Mercado Borrayo, Bertha Maria. (2007). Estudio sobre la remoción de cianobacterias y sus metabolitos en la planta potabilizadora “Los Berros” sistema cutzamala. Tesis para maestro en ingeniería ambiental-agua. Instituto de ingeniería; Universidad Nacional Autonoma de Mexico.

Meriluoto, J., Spoof, L., Codd, G.A. (Eds.), (2017). Handbook of cyanobacterial monitoring and cyanotoxin analysis. John Wiley & Sons.

Metcalf, J. and Codd, G., (2012). Cyanotoxins. *Ecology of Cyanobacteria II*, pp.651-675.

Moe, S.J., Haande, S. & Couture, R. (2016). Climate change, cyanobacteria blooms and ecological status of lakes: A Bayesian network approach. *Ecological Modelling*, 337: 330-347

Moe, S.J., Couture, R., Haande, S., Lyche Solheim, A. & Jackson-Blake, L. (2019). Predicting Lake Quality for the Next Generation: Impacts of Catchment Management and Climatic Factors in a Probabilistic Model Framework. *Water (Basel)*, 11(9): 1767

Muñoz, M., Nieto-Sandoval, J., Cirés, S., de Pedro, Z.,M., Quesada, A. & Casas, J.A. (2020). Degradation of widespread cyanotoxins with high impact in drinking water (microcystins, cylindrospermopsin, anatoxin-a and saxitoxin) by CWPO.

Nechhi Jr. O., Branco L. H. Z. & Branco C. C. Z. (1995): Comparison of three techniques for estimating periphyton abundance in bedrock streams. - Arch. Hydrobiol. 134: 93-402.

New Zealand Govt Mfe.. (2020). *Section 3. The Guidelines | Ministry For The Environment*. [online] Available at: <https://www.mfe.govt.nz/publications/fresh-water/guidelines-cyanobacteria/section-3-guidelines>

Niamien-Ebrottie J.E., Bhattacharyya S, Deep P.R., Nayak B (2015). Cyanobacteria and cyanotoxins in the World: Review. *International Journal of Applied Research.* 1. 563-569.

Noges, P., Mischke, U., Laugaste, R., (2010). Analysis of changes over 44 years in the phytoplankton of Lake Vortsjarv (Estonia): the effect of nutrients, climate, and the investigator of phytoplankton-based water quality indices. *Hydrobiologia.* 646: 33-48

Goronto. (2017) ODU Model Engineering Genie for Bayesian Networks. Modeling, Stimulation, Visualization Engineering

Paerl, H. W., N. S. Hall, and E. S. Calandrino (2011), Controlling harmful cyanobacterial blooms in a world experiencing anthropogenic and climatic-induced change, *Science Total Environment*, 409(10), 1739–1745.

Paerl, H. & Otten, T. (2013). Harmful Cyanobacterial Blooms: Causes, Consequences, and Controls. *Microbial Ecology*, 65(4): 995-1010.

Paerl, H.W, Paul, V.J. (2012). Climate change: Links to global expansion of harmful cyanobacteria. *Water Research.* 46 (2012) 1249-1363

Pawlik-Skowrońska, B., Kalinowska, R. & Skowroński, T. (2013). Cyanotoxin diversity and food web bioaccumulation in a reservoir with decreasing phosphorus concentrations and perennial cyanobacterial blooms. *Harmful Algae*, 28: 118-125.

Perona, E., Bonilla, I & Mateo, P. (1999). Spatial and temporal changes in water quality in a Spanish river. *Sci. Total Environ.* 241: 75 - 90

Perona, E. Haya, R., Alemla, P., Konig, S., Cirés, S., Asens, G., Martin-Munos, MA, Mateo, P., Quesada, A., (2017), Cyanobacterial benthic communities as producers of anatoxin-a in rivers: *Phormidium* as the main actor. 5º CIC Congreso Iberoamericano de Cianotoxinas.

Phan, T.D., Smart, J.C.R., Capon, S.J., Hadwen, W.L. & Sahin, O. (2016). Applications of Bayesian belief networks in water resource management: A systematic review. *Environmental Modelling and Software*, 85: 98-111

Pírez, M., Gonzalez-Sapienza, G., Sienra, D., Ferrari, G., Last, M., Last, J.A. & Brena, B.M. (2013). Limited analytical capacity for cyanotoxins in developing countries may hide serious environmental health problems: Simple and affordable methods may be the answer. *Journal of Environmental Management*, 114: 63-71.

Quesada, A., Haya, R., Cubero, S., Konig, S., Cirés, S., Ramos, M., Munoz, M., Mateo, P., Perona, E., Ecology and community structure of river mats of anatoxin-a benthic producers. ICTC congress, China, 2016

Quiblier, C., Wood, S., Echenique-Subiabre, I., Heath, M., Villeneuve, A. & Humber, J.F. (2013). A review of current knowledge on toxic benthic freshwater cyanobacteria – Ecology, toxin production and risk management. *Water Research*, 47(15): 5464-5479.

Ramos, M., (2012), Environmental Influences in *Phormidium* Biofilms Development in Rivers. Final master project, Sciences Faculty, Biology Department, UAM.

Rahman, S. and Jewel, M., (2008). Cyanobacterial blooms and water quality in two urban fish ponds. *University Journal of Zoology, Rajshahi University*, 27(ISSN 1023-6104), pp.79-84.

Rigosi, A., Hanson, P., Hamilton, D.P., Hipsey, M., Rusak, J.A., Bois, J., Sparber, K., Chorus, I., Watkinson, A.J., Qin, B., Kim, B. & Brookes, J.D. (2015). Determining the probability of cyanobacterial blooms: the application of Bayesian networks in multiple lake systems. *Ecological Applications*, 25(1): 186-199.

Scott, J.T., McCarthy, M.J., Otten, T.G., Steffen, M.M., Baker, B.C., Grantz, E.M., Wilhelm, S.W., Paerl, H.W., (2013). Comment: an alternative interpretation of the relationship between TN:TP and microcystins in Canadian lakes. Can. J. *Fish Aquatic Science* 70, 1265e1268

Scutari, M. (2010). Learning Bayesian Networks with the bnlearn R Package. *Journal of Statistical Software*, 35(3).

Shan, K., Wang, X., Yang, H., Zhou, B., Song, L. & Shang, M. (2020). Use statistical machine learning to detect nutrient thresholds in Microcystis blooms and microcystin management. *Harmful Algae*, 94: 101807.

Shimoda, Y., Watson, S.B., Palmer, M.E., Koops, M.A., Mugalingam, S., Morley, A., Arhonditsis, G.B., (2016). Delineation of the role of nutrient variability and dreissenids (Mollusca, Bivalvia) on phytoplankton dynamics in the Bay of Quinte, Ontario, Canada. *Harmful Algae* 55, 121e136.

Sig.mapama.gob.es. (2020). *SNCZI-Inventario De Presas Y Embalses*. [online] Available at: <https://sig.mapama.gob.es/snczi/> .

Solter, P. and Beasley, V., (2013). Phycotoxins. *Haschek and Rousseaux's Handbook of Toxicologic Pathology*, pp.1155-1186.

SpeakLounge. 2020. *Elevation Map – Speaklounge*. [online] Available at: <https://speaklounge.wordpress.com/category/elevation-map/>

Svirčev, Z., Lalić, D., Bojadžija Savić, G., Tokodi, N., Drobac Backović, D., Chen, L., Meriluoto, J. & Codd, G.A. (2019). Global geographical and historical overview of cyanotoxin distribution and cyanobacterial poisonings. *Archives of Toxicology*, 93(9): 2429-2481

Tomasini-Ortiz, Ana Cecilia, Moeller-Chávez, Gabriella, Sánchez Chávez, José Javier, Bravo Inclán, Luis Alberto (2012) Cyanobacteria and Cyanotoxins in Lake Patzcuaro, Michoacan, Mexico. *Revista AIDIS de Ingeniería y Ciencias Ambientales.* 5 (2), 93-101

Uriza, E.A.C., Asencio, A.D. & Aboal, M. (2017). Are We Underestimating Benthic Cyanotoxins? Extensive Sampling Results from Spain. *Toxins*, 9(12): 385.

Wagner, C. & Adrian, R. (2009). Cyanobacteria dominance: Quantifying the effects of climate change. *Limnology and Oceanography*, 54(6): 2460-2468.

Wang, Haiqin. (2006). Using Sensitivity Analysis to Validate Bayesian Networks for Airplane Subsystem Diagnosis. *Mathematics and Computing Technology. Boeing Phantom Works.*

Wood, S.A., Rasmussen, J.P., Holland, P.T., Campbell, R. & Crowe, A.L.M. (2007). First Report of the Cyanotoxin Anatoxin-a From Aphanizomenon Issatschenkot (Cyanobacteria)1. *Journal of Phycology*, 43(2): 356-365

Wood, S.A., Borges, H., Puddick, J., Biessy, L., Atalah, J., Hawes, I., Dietrich, D.R., Hamilton, D.P., (2017). Contrasting cyanobacterial communities and microcystin concentrations in summers with extreme weather events: insights into potential effects of climate change. *Hydrobiologia* 785, 71e89

World Health Organization (WHO), (1999). Guidelines for Drinking - Water Quality, second ed. Addendum to Vol 2, Geneva, Switzerland.

World Health Organization (2003) Guidelines for safe recreational water environments. Volume 1, Coastal and fresh waters. ISBN 92 4 154580 1, 33pp. Available on line at: http://www.who.int/water_sanitation_health/bathing/srwe1/en/

Wörmer, L., Agha, R., Cirés, S., Galán, E., Ratón, C., Al-Ismail, S., Quesada, A., (2011). Informe de los análisis realizados en las zonas de baño continentales durante las temporadas 2008 y 2009. In: Ministerio de Medio Ambiente, Medio Rural y Marino (MMAMRM). Cianobacterias. Madrid, Spain.

Yuan, L.L., Pollard, A.I., Pather, S., Oliver, J.L., D'Anglada, L., (2014). Managing microcystin: identifying national-scale thresholds for total nitrogen and chlorophyll a. *Freshwater Biology*. 59, 1970e1981

8. Annexes

***Annex 1** Discretization Intervals for all BN*

Overview of discrete intervals of the nodes used in each BN. Spain res stands for Spanish lentic systems, Mexico res stands for Mexican lentic systems, and Spain river is the Spanish lotic systems *Month not included for Mexican reservoirs due to all months being present, 12 months.

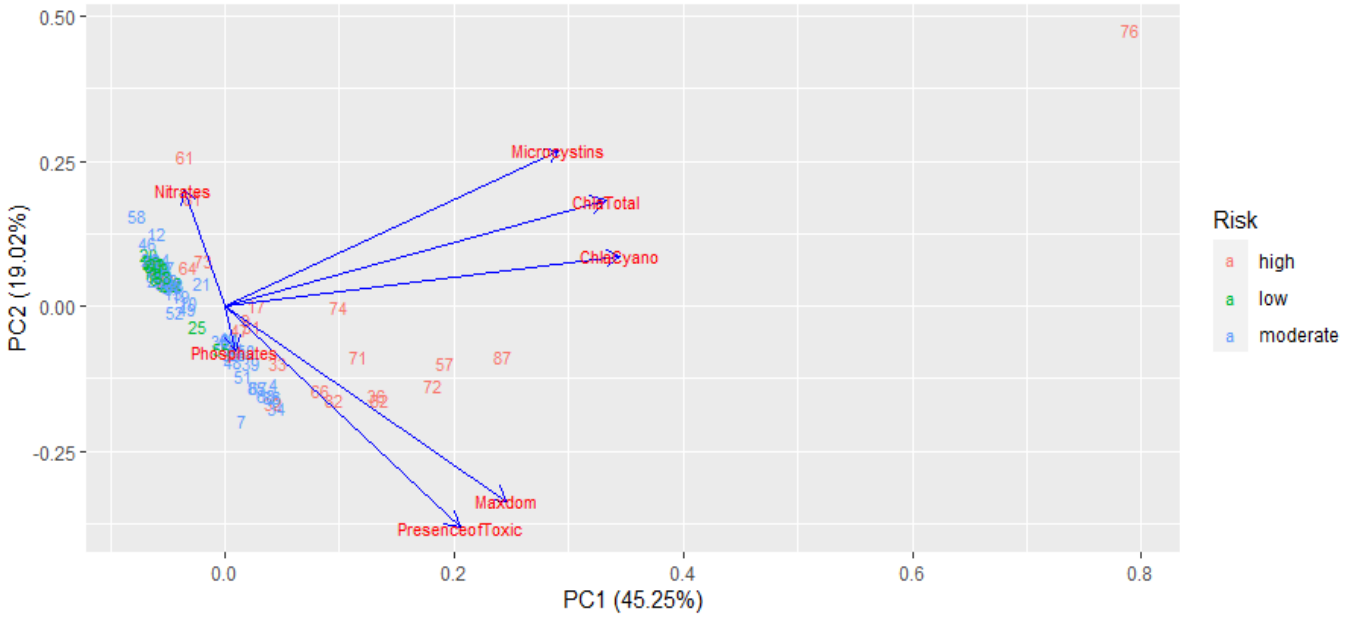
Country BN	Node name	Units	Node States					
			1	2	3	4	5	6
Spain res.	Nitrates (NO_3^-)	mg/L	<0.9	0.9-1.6	1.6-3.5	3.5-4.9	4.9-7.4	>7.4
Mexico res			<0.15	0.15-0.47	0.47-0.65	0.65-0.85	0.85-1.15	>1.15
Spain river			<0.14	0.14-0.22	0.22-0.35	0.35-0.63	0.63-1.9	>1.9
Spain res	Phosphates (PO_4^{3-})	mg/L	<0.1	0.1-0.4	0.4-0.6	0.6-1.3	1.3-2.9	>2.9
Mexico res			<2.2	2.2-4.7	4.7-7.5	7.5-9.5	9.5-11.5	>11.5
Spain river			<0.01	0.01-0.03	0.03-0.04	0.04-0.06	0.06-0.07	>0.07
Spain res	Chl*a*Total	µg/L	<5	5-10	10-20	20-44	44-100	>100
Spain res	Chl*a*Cyano	µg/L	<4.8	4.8-15	15-25	25-39	39-67	>67
Spain res	Maxdom	%	<7	7-27	27-50	50-70	70-84	>84
Mexico res	Conductivity	mS/cm	<0.3	0.4-0.8	0.8-1.3	1.3-2.1	2.1-3.5	>3.5
Spain river		µS/cm	<105	105 - 237	237 - 472	472 - 811	811- 1158	>1158
Mexico res	pH		<7.5	7.5-8.3	8.3-9.7	9.7-10.4	>10.4	
Spain river			<5.7	5.7-6.3	6.3-7	7-7.5	7.5 - 8	>8
Spain river	CV	m/s	<0.2	0.2-0.3	0.3-0.57	0.57-0.85	0.85-1.2	>1.2
Spain river	Biofilm size	%	<20	20-50	50-57.5	57.5-75	75-92	>92
Spain river	Depth	cm	<5.7	5.7-11.6	11.6-15.7	15.7-22.5	22.5-29.8	>29.8
Spain river	Light		Direct sun	Almost shadow	Shadow	Lateral sun		
Mexico res	Season		Spring	Summer	Autumn	Winter		
Spain river			Spring	Summer	Autumn			
Mexico res	Temperature	ºC	<16	16.5-17.8	17.8-19.6	19.6-22.4	>22.4	
Mexico res	Dissolved Oxygen	mg/L	<4.3	4.3-6.4	6.4-9	9-11.2	11.2-14.6	>14.6
Spain res	Microcystins	(µg/L)	<0.01	0.01 - 0.4	0.4-1.0	1.0-10	10-20	>20
Mexico res			<1	1.0-4.0	4.0 - 6.8	6.8-10	10-12	
Spain res	Presence of Toxic Genus		pres.	absence				
Spain river			pres.	absence				
Spain river	Anatoxin-a	µg ANA/mg dw	>1	1.0-3.0	3.0-6.0	6.0-17.0	17.0-27.8	>27.8
Spain res	Risk		High	Moderate	Low			

***Annex 2** Papers where data was found for Mexican BN*

Lentic bodies of water and papers where raw data can be found

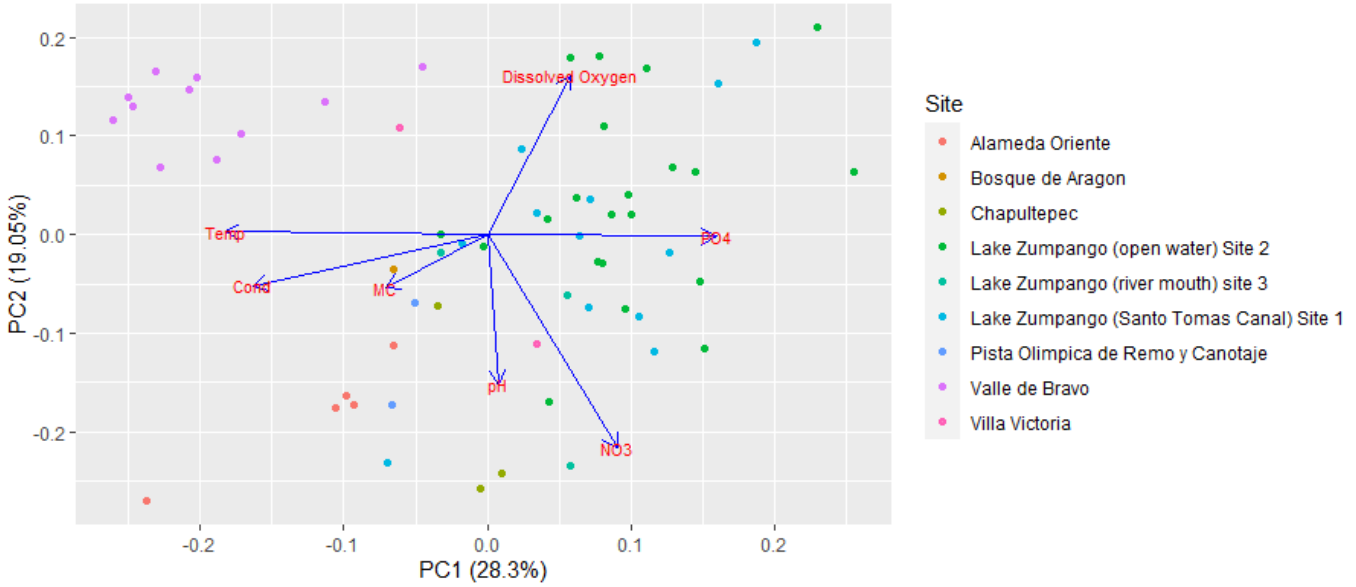
Lentic Water Body	Reference
Los Berros	Mercado-Borrayo, 2007
Villa Victoria	Arzate-Cardenas, 2008
Valle de Bravo	
Chapultepec	
Alameda Oriente	
Pista Olimpica de Remo y Canotaje	
Bosque de Aragon	
Lake Zumpango	Figueroa-Sánchez, Nandini and Sarma, 2020

Annex 3 *Spanish Lentic Systems PCA*



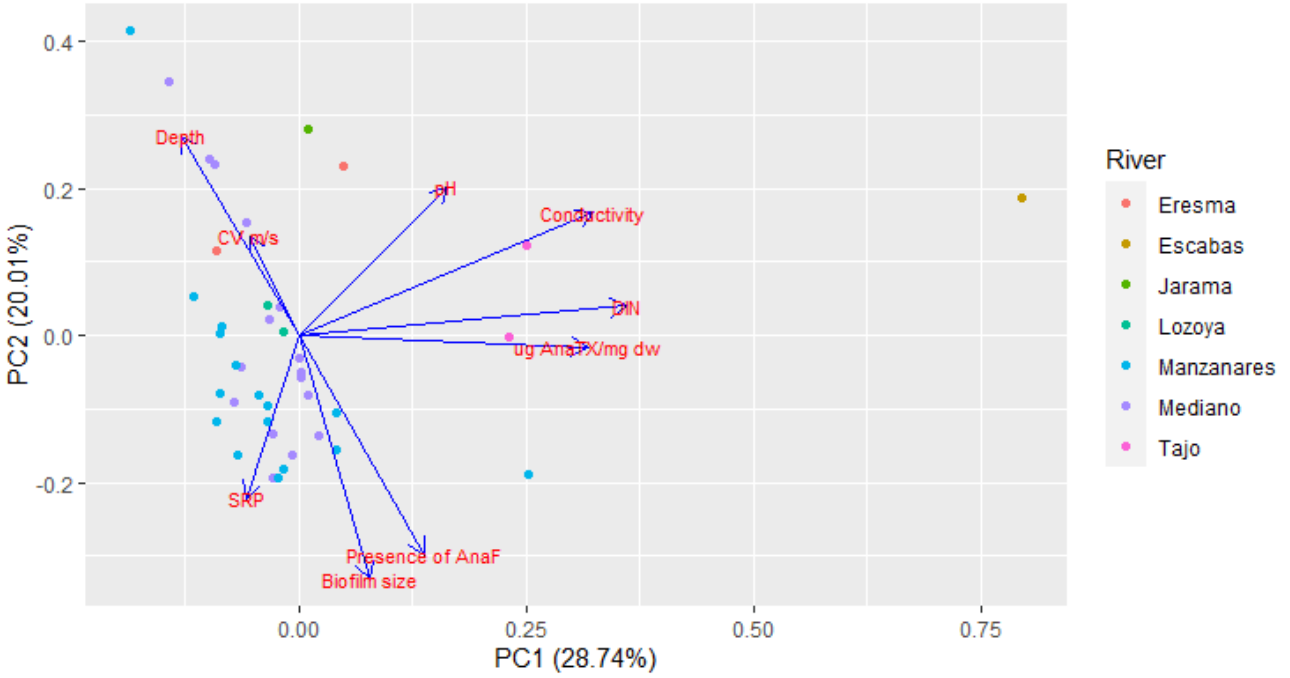
PC1 vs PC2 loadings and scores for Spanish lentic systems. Numbers represent the corresponding sampling site. Percentages on axes represent the amount of variance explained by each principal component.

Annex 4 *Mexican Lentic Systems PCA*



PC1 vs PC2 loadings and scores for Mexican lentic systems. Percentages on axes represent the amount of variance explained by each principal component. Abbreviations are as follows: Temp. = temperature, Cond. = conductivity, MC = microcystin, NO3 = NO_3^-, and PO4 = PO_4^{3-}

Annex 5 *Spanish Lotic Systems PCA*



PC1 vs PC2 loadings and scores for Spanish lotic systems. Percentages on axes represent the amount of variance explained by each principal component. Abbreviations are as follows: SRP = Soluble reactive phosphorus, DIN = Dissolved inorganic nitrogen, CV = Current velocity