**Abstract:**

Agricultural activities are a source of surface runoff of pathogens, pesticides, and nutrients into groundwater, resulting in discharge and contaminating surface water bodies. Lack of financial resources makes risk assessment through specific analysis challenging for drinking water suppliers. Inability to identify agricultural lands with a high-risk level and implement action measures might lead to public health issues. As a result, it is essential to identify hazards and conduct risk assessments even with limited data. This study proposes a risk assessment model for agricultural activities based on only available data and integrating various types of knowledge, such as expert and literature-based knowledge. To accomplish so, we built a Bayesian network with continuous and discrete inputs capturing raw water quality and land use upstream of the drinking water intakes. This probabilistic model integrated the vulnerability of the drinking water intake, the exposure to the threat, and the threat caused by agricultural activities, including animal and crop production inventoried in drainage basins. The probabilistic dependencies between nodes that structure this model are mostly established by setting a novel technique that involves adopting a mixed aggregation method. The latter is traditionally used in ecological assessments following a deterministic approach. After validation, this probabilistic model was used for four water intakes in a heavily urbanized watershed with agricultural activities in the south of Quebec, Canada. The findings imply that this methodology may detect hotspots and can help stakeholders direct their efforts and investments to the most appropriate target in the first instance.
work included in this submission and the role the funder(s) played. This includes grants and any commercial funding of the work or authors.

This statement will be typeset if the manuscript is accepted for publication.

*Please review the submission guidelines and the instructions link below for detailed requirements and guidance.*

**Competing Interests**

On behalf of all authors, disclose any competing interests that could be perceived to bias this work.

This statement will be typeset if the manuscript is accepted for publication.

*Please review the instructions link below and PLOS Water's competing interests policy to determine what information must be disclosed at submission.*

**Data Availability**

Before publication, Authors are required to make fully available and without restriction all data underlying their findings. Please see our PLOS Data Policy page for detailed information on this policy.

A Data Availability Statement, detailing where the data can be accessed, is required at first submission. Insert your Data Availability Statement in the box below.

Data are available at Open Science Framework: https://osf.io/zyekt/?view_only=01527a99a3374a2ab7879e42ff01052b
Identifier: DOI 10.17605/OSF.IO/ZYEKT
<table>
<thead>
<tr>
<th>Availability Statement.</th>
</tr>
</thead>
<tbody>
<tr>
<td>This statement will be typeset if the manuscript is accepted for publication.</td>
</tr>
</tbody>
</table>

*PLOS allows rare exemptions to address legal and ethical concerns. If you have legal or ethical restrictions, please use the box below to detail these in full sentences for the Journal team to consider.*
Full title: Risk assessment of drinking water intake contamination from agricultural activities using a Bayesian network

Short title: Bayesian network and agricultural risk assessment

Raja Kammoun\(^1\)*, Natasha McQuaid\(^1\), Vincent Lessard\(^2\), Eyerusalem Adhanom Goitom\(^1\), Michèle Prévost\(^1\), Françoise Bichai\(^1\), Sarah Dorner\(^1\)

\(^1\) Department of Civil, Geological and Mining Engineering, Polytechnique Montréal, Montreal, Quebec, Canada

\(^2\) Conseil des bassins versants des Mille Îles, Saint-Eustache, Quebec, Canada

* Corresponding author

E-mail: raja.kammoun@polymtl.ca (RK)

Abstract

Agricultural activities are a source of surface runoff of pathogens, pesticides, and nutrients into groundwater, resulting in discharge and contaminating surface water bodies. Lack of financial resources makes risk assessment through specific analysis challenging for drinking water suppliers. Inability to identify agricultural lands with a high-risk level and implement action measures might lead to public health issues. As a result, it is essential to identify hazards and conduct risk assessments even with limited data.

This study proposes a risk assessment model for agricultural activities based on only available data and integrating various types of knowledge, such as expert and literature-based knowledge. To accomplish so, we built a Bayesian network with continuous and discrete inputs capturing raw water quality and land use upstream of the drinking water intakes. This probabilistic model integrated the vulnerability of the drinking water intake, the exposure to the threat, and the threat caused by agricultural activities, including animal and crop production inventoried in drainage basins. The probabilistic dependencies between nodes that structure this model are mostly established by setting a novel technique that involves adopting a mixed...
aggregation method. The latter is traditionally used in ecological assessments following a deterministic approach. After validation, this probabilistic model was used for four water intakes in a heavily urbanized watershed with agricultural activities in the south of Quebec, Canada. The findings imply that this methodology may detect hotspots and can help stakeholders direct their efforts and investments to the most appropriate target in the first instance.

1 Introduction

Agricultural practices are important sources of chemical and microbial pollution in surface waters (1-5), that may pose a threat to public health. Runoff from agricultural fields leads to the transport of pesticides and nutrients into receiving waters (6). The potential for contamination of surface waters depends on the type, quantity, adsorption, and absorption capacity of applied pesticides, nitrogen and phosphorus (7-9). In addition, soil characteristics (10), rainfall rates (11-13), and agricultural management practices (14-16) are essential factors that determine the susceptibility of runoff to contamination. An increase in nutrient levels from animal and crop production can contribute to the proliferation of cyanobacteria in drinking water sources (17-20). Some cyanobacteria blooms can release toxins that pose challenges in terms of drinking water safety (21-26). Animal excrements and manure are also a source of microbial contamination in surface waters (27-30). Factors influencing the survival and transport of pathogens in water resources include: water temperature, pH, biotic interactions, manure types and characteristics, and farm management practices (31).

Risk assessments can raise awareness of the threats posed by agricultural contamination to drinking water sources. Several methods to assess the risks associated with agricultural activities are available. In recent years, many risk assessment techniques have used deterministic modeling approaches focused on ecological, ecotoxicological, and human health issues (3, 32-37). Deterministic modeling approaches are costly, require extensive calibration and need a large amount of site-specific data on water quality and hydrodynamics. The central problem with this approach is that it cannot be implemented by all drinking water suppliers due to a lack of financial resources and data. On the other hand, the deterministic approach
could lead to an erroneous risk management decision, as it evaluates the impact of only one risk scenario at a time. Some of these limitations can be addressed by using a probabilistic approach that considers many scenarios while assessing risks, and the influence of uncertainty in the different input variables and circumstances on the risk magnitude (38).

Bayesian networks (BN) are a valuable tool to conduct a probabilistic risk evaluation under uncertainty and integrate different types of knowledge (39). BNs have been extensively applied in different fields in different ways, such as in the fields of ecological risk assessment (40-43), human health risk assessment (40, 44), climate change risk assessment (45), disaster risk assessment (46, 47), water quality and treatment (48, 49), best management practices of non-point source pollution (50), and water resource management (51). To the best of our knowledge, these approaches have not been applied to vulnerability assessments of drinking water sources in agricultural watersheds.

The aim of this study was to develop a probabilistic model that decision-makers could use to swiftly identify animal and crop production areas that generate higher risks at drinking water intakes (DWIs). Our specific objectives were to: (1) develop a BN that combines the three components of risk assessment: hazard, exposure, and vulnerability; (2) test the applicability of existing methods used to define the conditional independence relationship among the related variables in a BN; (3) develop a novel method to define an efficient Bayes independence in case of water resource risk assessment; (4) assess the probability distribution of the overall hazard of agricultural activities to pinpoint the drainage basins that require more detailed investigation for source water protection.

2 Materials and methods

2.1 Study area

The study was conducted for four drinking water intakes, serving a population of approx. 440,000, drawing water from a river in Southern Quebec, Canada (Fig 1). Drinking water intakes 1 (DWI_1) and 4 (DWI_4) are located in the most upstream and downstream sections of the river, respectively. DWI_2 and DWI_3
are located in the middle. The 42-km long river drains a lake fed by the Ottawa River, as described in Jalliffier-Verne, Leconte (52). The Ottawa River is under the influence of the largest (146,300 km²) watershed in Eastern Canada, and its average discharge is around 1,950 m³/s (53). This watershed is dominated by forests (75%) of which 40% is dense mixed wood. The agricultural activities represent only 6% of its total surface. However, the adjacent watersheds that have a local influence on the river that supplies the DWIs of this study cover an area of 1,008 km². This area is zoned as follows: 30% of agricultural land, 28.69% of urban areas, and 20.63% of forest land use.

Fig 1. Study site

2.2 DWI protection zone delineation

The DWI protection zones include the land area that contributes water to the drinking water supply. Delineation of protection zones is a critical step to assess non-point source pollution risks to drinking water sources, reduce the risk of contamination, and support drinking water safety.

In Quebec, provincial regulation requires the delineating of surface water protection zones using a fixed distance approach (54). This approach only comprises a strip of land on both sides of a river and does not include land surfaces whose relief leads to the drainage of land, lakes, and streams upstream to the river under study. For more information on regulatory methods for delineating surface water protection areas, see the supplementary material (S1 Text). To control the diffuse pollution, it is necessary to consider the drained non-point source pollution lands upstream of the tributary outlets included in the protection zones of the DWI. Therefore, this study develops an improved methodology of delineation based on levels 1 and 2 sub-watersheds. Level 1 sub-watersheds are the large watersheds whose outlet is the study site river. Level 2 sub-watersheds are those that drain into level 1 sub-watersheds and have the river in level 1 sub-watersheds as their outlet.

The geographic information system software (ArcGIS) was used to delineate the protection zones using the following data layers: location of DWIs, topological hydrographic network lines, and high-water mark. The
Immediate protection zone was delimited within 50 m downstream and 500 m upstream of the water intake. This area includes any surface water, portions of tributaries, and a 10-m strip of land measured from the high-water mark. Surface water and portions of tributaries were also included in the intermediate protection zone, although only a 120-m strip of land measured from the high-water mark. This protection zone was delimited within 50 m downstream and 10 km from the upstream of the DWI (54). This delineation was used to identify all level 1 and 2 sub-watersheds with outlets within the boundaries of these protection zones and to include them in the study site. Every sub-watershed selected is a drainage basin (DB) (Figure A in S1 Fig). By definition, a drainage basin is a territory including a set of parcels used for one or more anthropogenic activities presenting a source of diffuse pollution and whose drained water is discharged into a watercourse through an outlet. Each drainage basin presents a potential source of microbial and/or chemical contamination to the studied DWIs.

2.3 Bayesian Network Framework

A probabilistic and spatial risk assessment of agricultural activities impacting DWIs was performed at a regional scale using a BN and ArcGIS. The risk assessment framework was built to integrate the three components of risks: hazard, vulnerability, and exposure (Fig 2).

A BN is a hierarchical representation and a probabilistic graphical model representing a relationship between a range of variables. Every single variable is represented as a node, and relationships between variables are presented by arcs defined by conditional dependencies (55). This directed acyclic graph allows the modeling of various uncertain events, facts, and systems (56). More definitions of BN can be found in other studies (38, 47, 50, 57-60). Bayesian networks are also known as Conditional Probability Networks, Casual Probability Networks, probabilistic networks, belief networks, Bayesian belief networks, or Bayesian reliability networks.

In this study, the BN was developed using the software Netica 6.09 (Norsys Software Corp), applying the Bayes’ theorem (61). This software is an appropriate tool for creating, changing, and saving networks, as reviewed by Uusitalo (39).
Fig 2. Risk assessment process.

2.3.1 Defining the nodes

The first step to structure the BN model was defining a set of random variables (nodes) and their states, which can be discrete or continuous. Discrete variables have a limited set of fixed values or states that they can take. Continuous variables are classified as real numbers since they may have any value within a certain range (62). Continuous variables were discretized by breaking them up into intervals by assigning thresholds to define each interval. Input nodes are labeled “parent” nodes, and “child” nodes have input connections from one or more parent nodes (40, 58).

The determination of the critical nodes constituting the BN was carried out based on the problem and the objective of the study, which considers the DWI as the main subject of the risk assessment from agricultural activities. These nodes were chosen to represent all components of the risk assessment model based on the interaction of vulnerability to microbiological contamination (VMC), vulnerability to chemical contamination (VCC), vulnerability to water scarcity (VWS), exposure (EXP), and probability of hazard (PH). In the Working Group II (WGII) sixth assessment report from Intergovernmental Panel On Climate Change (63), the hazard is described in terms of the potential occurrence of the chemical and microbial contaminants that may cause degradation of water resource. The exposure depends on the presence of DWI in locations that could be adversely affected. The vulnerability is evaluated as the propensity or predisposition of the DWI to be adversely affected.

In this study, the intrinsic vulnerability of the intake was assessed. This vulnerability was defined by the sensitivity of the raw water at the intake. Three key variables were used to assess the vulnerability to microbiological contamination of water intakes: (i) the geometric mean of Escherichia coli concentrations was used to present the central tendency of the microbial contamination indicator while being robust to extreme values; (ii) the 99th percentile of E. coli concentrations was used to account for the peaks of microbial contamination that are usually responsible for waterborne diseases (64); and (iii) the seasonal variability of microbial contamination that was calculated based on a consecutive disparity index that
estimates the variability of \textit{E. coli} concentrations while considering the order of their chronological distribution, and without being independent of the data set’s average (65). All variables were calculated for the four seasons of the year (66) (Winter, Spring, Summer, and Autumn) to study microbial contamination distribution and evaluate the sensitivity to seasonal variations.

Assessment of the vulnerability to chemical contamination was carried out based on the total phosphorus in raw water and the utilization of the land by activities that may release chemical contaminants in the immediate and intermediate DWIs’ protection zones. Raw water quality data available over a period of five years (January 2013 to December 2017) was used to assess the microbiological and chemical vulnerability of DWIs as required by Water Withdrawal and Protection Regulation (54).

The results of the vulnerability to water scarcity from the Leveque (67) study were used. This vulnerability was assessed based on (i) the water levels in the studied river and (ii) the water demand assigned to each catchment. The exposure was estimated based on the location of the outlet of each drainage basin in relation to the intake and its protection areas. This node allows assessing the non-point source pollution exposure pathways to the raw water supply.

The probability of hazard was estimated based on the proportion of the drainage basins occupied by crop and animal production activities. Crop classes and animal classes were used as input variables to estimate surface water contamination levels by pesticides and microbial contaminants. The final node presents the overall risk probability distribution across three possible states (low, medium, and high) of agricultural activities occupying a watershed located upstream of a DWI.

The input variables, their states, and the discretization methods are described in Table 1. Once the nodes were set up, links (dependencies) between them were defined using conditional probability tables (CPTs) (Section 2.3.2). Finally, the values for each input parameter were computed using the data associated with the study site to produce a probability distribution for each state associated with the nodes.
The determination of the remaining nodes, their states, and directed links was accomplished based on scientific knowledge (the literature), the Quebec regulations, and the availability of data in provincial and federal databases. The choice of nodes was discussed with experts and stakeholders to provide a suitable BN, rectify, refine, and fill the potential gaps associated with the variables and their status.
Table 1: Summary of input variables’ properties.

<table>
<thead>
<tr>
<th>Input variable</th>
<th>Description</th>
<th>Type of discretization</th>
<th>State discretization methodology</th>
<th>States (from lowest to highest level)</th>
<th>Data source</th>
</tr>
</thead>
<tbody>
<tr>
<td>DBA</td>
<td>Drainage basin (DB) area</td>
<td>Continuous</td>
<td>Statutes were defined based on percentiles of the area of 4723 sub-watersheds (level 2) in Quebec.</td>
<td>0 to 5 Km² / 5 to 12 km² / 12 to 45 km² / &gt;= 45 km²</td>
<td>(68)</td>
</tr>
<tr>
<td>PPCO</td>
<td>Plant production cover</td>
<td>Continuous</td>
<td>Statutes were defined based on Water Withdrawal and Protection Regulation (WWPR) (schedule IV) (54): land use-based method for assessing water vulnerability.</td>
<td>0 to 20% / 20 to 50% / &gt;= 50%</td>
<td>(69)</td>
</tr>
<tr>
<td>APCO</td>
<td>Animal production cover</td>
<td>Continuous</td>
<td>Environmental impacts depend on the crop class. The statutes were specified based on the data layer available in the ministry database (Ministère de l’environnement et de la lutte contre les changements climatiques (MELCC)), and their classification was made based on the following references: Beaudin, Deslandes (70), Michaud, Deslandes (71), Gasser, Bolinder (72), Giroux (73), and Giroux (74).</td>
<td>0 to 20% / 20 to 50% / &gt;= 50%</td>
<td>Data are provided by the Ministère des Affaires municipales et de l’Habitation of Quebec (MAMH).</td>
</tr>
<tr>
<td>CPC</td>
<td>Crop production classes</td>
<td>Discrete</td>
<td>No crop production / Non-cultivated agricultural area / Undefined culture / Specialized culture / Perennial culture / Annual crop</td>
<td>No animal production / Class VII² / Classes V(3) - VI(4) / Classes II(5) - III(6) - IV(7) / Class I(8)</td>
<td>(75)</td>
</tr>
<tr>
<td>APC</td>
<td>Animal production classes</td>
<td>Discrete</td>
<td>Animal manure is a source of microbial contamination for surface waters and contains pathogens. These pathogens are specific to the types of animals from which manure is derived. The statutes were specified based on the data layer available in the MAMH database, and their classification was made based on the following references: Soupir, Mostaghimi</td>
<td>No animal production / Class VII² / Classes V(3) - VI(4) / Classes II(5) - III(6) - IV(7) / Class I(8)</td>
<td>Data are provided by MAMH.</td>
</tr>
<tr>
<td>Name</td>
<td>Description</td>
<td>Type of discretization</td>
<td>State discretization methodology</td>
<td>States (from lowest to highest level)</td>
<td>Data source</td>
</tr>
<tr>
<td>----------</td>
<td>--------------------------------------------------</td>
<td>------------------------</td>
<td>--------------------------------------------------------------------------------------------------</td>
<td>----------------------------------------------------------------------------------------------------</td>
<td>------------------------------------------------------------------------------------------------</td>
</tr>
<tr>
<td>PZ</td>
<td>Protection zone</td>
<td>Discrete</td>
<td>Statutes were defined based on WWPR (Art. 70, 72, and 74) (54): protection zone delineation.</td>
<td>Extended protection zone / Intermediate protection zone / Immediate protection zone</td>
<td>Data are provided by the watershed organization.</td>
</tr>
<tr>
<td>GMEC_X</td>
<td>Geometric mean of (E. ) coli concentrations</td>
<td>Continuous</td>
<td>Statutes were defined based on the regulation of the drinking water quality (80), and the</td>
<td>0 to 15 CFU/100 ml 15 to 150 CFU/100 ml &gt;= 150 CFU/100 ml</td>
<td>Raw water quality data are provided by drinking water treatment plants managers.</td>
</tr>
<tr>
<td>PKEC_X</td>
<td>99th percentile of (E. ) coli concentrations</td>
<td>Continuous</td>
<td>Statutes were defined based on the development of an index of bacteriological and physicochemical quality (IQBP) (81).</td>
<td>0 to 200 CFU/100 ml 200 to 2000 CFU/100 ml &gt;= 2000 CFU/100 ml</td>
<td></td>
</tr>
<tr>
<td>SVMC_X</td>
<td>Seasonal variability of microbial contamination</td>
<td>Continuous</td>
<td>Statutes were defined based on the results of the index of seasonal variability to microbial contamination developed by Kammoun, Leveque (65).</td>
<td>0 to 1 / 1 to 1.5 / &gt;= 1.5</td>
<td></td>
</tr>
<tr>
<td>VWS</td>
<td>Vulnerability to water scarcity</td>
<td>Discrete</td>
<td>Statutes were defined based on the water scarcity vulnerability index developed by Leveque (67).</td>
<td>Low / Medium / High</td>
<td>(67)</td>
</tr>
<tr>
<td>PKTP</td>
<td>99th percentile of total phosphorus concentrations</td>
<td>Continuous</td>
<td>Statutes were defined based on the WWPR (54), and the development of an index of bacteriological and physicochemical quality (IQBP) (81).</td>
<td>0 to 50 µg/l 50 to 100 µg/l / &gt;= 100 µg/l</td>
<td>Raw water quality data are provided by drinking water treatment plants managers.</td>
</tr>
<tr>
<td>MTP</td>
<td>Median of total phosphorus concentrations</td>
<td>Continuous</td>
<td>Statutes were defined based on the basis of the WWPR (54).</td>
<td>0 to 30 µg/l 30 to 50 µg/l &gt;= 50 µg/l</td>
<td></td>
</tr>
<tr>
<td>AIMPZ</td>
<td>Area of immediate protection zone</td>
<td>Continuous</td>
<td>Statutes were defined on the basis of the WWPR (54).</td>
<td>0 to 1 km(^2) / 1 to 5 km(^2) /</td>
<td>(82-87)</td>
</tr>
<tr>
<td>Name</td>
<td>Description</td>
<td>Type of discretization</td>
<td>State discretization methodology</td>
<td>Data source</td>
<td></td>
</tr>
<tr>
<td>------------</td>
<td>-----------------------------------------------------------------------------</td>
<td>------------------------</td>
<td>----------------------------------</td>
<td>----------------------------------------------------------------</td>
<td></td>
</tr>
<tr>
<td>AINPZ</td>
<td>Area of intermediate protection zone</td>
<td>Continuous</td>
<td>0 to 10 km² / 10 to 30 km² / &gt;= 30 km²</td>
<td></td>
<td></td>
</tr>
<tr>
<td>PIMPA / PINPA</td>
<td>Percentage of the immediate/ intermediate protection zone used for activities that may release chemical contaminants</td>
<td>Discrete</td>
<td>No / 0 to 20% / 20 to 50% / &gt;= 50%</td>
<td>Land use data are provided by MAMH.</td>
<td></td>
</tr>
</tbody>
</table>

(1) X refers to season: Winter/ Spring/ Summer/ Autumn

(2) Class VII: other types of animal production/ experimental farm/ other agricultural activities; (3) Class VI: apiculture; (4) Class V: poultry and egg production; (5) Class IV: sheep and goat farming; (6) Class III: pig farming; (7) Class II: equine farming; (8) Class I: beef farming/ dairy cattle/ grazing land and pasture
2.3.2 Filling the CPTs

Filling the CPTs of the Bayesian network is the most challenging step in setting up the framework. To accomplish this step, different known methods were tested: (i) expert judgment (38, 58, 59), (ii) the interpolation method (88), and (iii) the Cain method (89, 90). Methods that did not generate reliable outcomes in terms of probabilities were rejected under the expert decision.

It was challenging to estimate the quantity of chemical and microbiological contaminants released during agricultural production activities, as well as their diffusion and dispersion into the drinking water intakes. Thus, employing expert judgment and scientific knowledge from the literature were the best methods for determining the conditional distribution of the quantities of pollutants that may be released in the immediate and intermediate protection zones. The exposure to pollutants released by upstream agricultural activities was also assessed using this method. In these situations, completing the CPTs was based on experts’ judgments and their interpretation by the extensive research team (rather than through formal interviews).

This is due to the nature of the project: experts’ judgments were interpreted throughout four years of work on a collaborative project between six municipal partners, a watershed organization, and a research team at Polytechnique Montreal. Thus, the probabilities for each status of each child node were elucidated through discussion meetings with researchers, professionals, and municipal water stakeholders and were enhanced by scientific knowledge.

However, it is challenging to ensure that the definition of probabilities distributions of several tables is coherent with each other rather than using simply the judgment of experts. This is mostly because experts lack the rigid regularity of machines since they are not machines (Ban et al., 2014; Das, 2004). During a long process, boredom and tiredness are enough to make sure that the criteria used to figure out the distributions are not always used the same way.

One solution would be to standardize the method of obtaining probabilistic information from the experts by providing a much smaller selection of probabilities distributions. Therefore, the interpolation method was used to get the remaining probability distributions in a reasonable time (88). This method requires
assigning a weight to each parent node to quantify the relative strength of its influence over the child node. This step was done based on expert judgment. The sum of the weights must be equal to one. Then, conditional probability distribution (CPD) elicitation was done for three cases: when all parent nodes are in extreme states (high and low states) and medium states. The determination of the probability distribution for the remaining combinations of the various parents’ states was done by interpolation. This method was allied to all CPTs of the Bayesian network. However, it was deemed to be not adequate for our study because an examination of the probability distributions revealed that the resulting probabilities were inconsistent. It was difficult to maintain consistency between the probability distributions, especially in cases where a parent node's state was highly different from the one used for interpolation. In this instance, this method does not account for the influence of variation bias in the interpolation. Therefore, interpolation method was not pursued further.

The Cain method was also tested. It involves determining CPDs for cases where the relationships between the parent nodes are known; these obtained conditional probability distributions are termed anchor CPDs. Anchors were defined according to the number of parents and the number of statuses associated with each parent. The distribution of missing conditional probabilities was derived by interpolating those corresponding to the anchor CPDs and using interpolation factors. The latter were calculated using the CPDs at anchors. Interpolation factors of a parent node were calculated for each change from one state to another to quantify the positive to negative state reduction. The processes of defining anchors and calculating interpolation factors are explained in more details in Cain (89) and Mkrtchyan, Podofillini (90).

Only two CPTs, relating to crop and animal activity levels, were completed using the Cain approach because this method is applicable for CPTs in which the effect of the parent state on the child state has a constant direction: the increase in the area occupied by animal and crop production activities results in an increase in the level of agricultural activities in the drainage basin. The results of filling in the probability tables for the other child nodes were judged to be unacceptable since the interpolation is dependent on the number of favourable and unfavourable states for the child state of interest. As a result, there will be the same
probabilities distributions for different child nodes even though the parent nodes have varying degrees of influence on them (91).

Therefore, a novel approach to filling CPTs was developed that combines a mixed aggregation method (92) with expert judgment. The mixed aggregation was used to aggregate indicators in the ecological assessments field using a deterministic approach (92). It was adapted to applied in the case of a probabilistic approach and for water resource assessment. First, the different state of each parent node was standardized to a common scale from 0 (best possible state) to 1 (worst possible state). Second, the assignment of weights \( W \) for each parent node \( P \) was performed to reflect its importance relative to the child node. Third, causal dependencies were performed with aggregation. In this step, additive and maximal aggregation were combined (Eq. (1), (2), (3), (4)). Additive aggregation \( f_{\text{add}} \) compensates for the poor status of a parent node by other nodes with a good status. On the other hand, maximal aggregation \( f_{\text{max}} \) leads to pessimistic results since only the parent node with the worst status is considered. The combination of the additive and maximal aggregation allows for some compensation while penalizing the very poor state to not neglect an impact that requires more attention to protect water resources.

\[
\begin{align*}
    f_{\text{add-max}}(P_1, P_2, ..., P_n) &= \alpha \cdot f_{\text{add}}(P_1, P_2, ..., P_n) + (1 - \alpha) \cdot f_{\text{max}}(P_1, P_2, ..., P_n) \quad (1) \\
    f_{\text{add}}(P_1, P_2, ..., P_n) &= \sum_{i=1}^{n} W_i \cdot P_i = W_1 \cdot P_1 + W_2 \cdot P_2 + \cdots + W_n \cdot P_n \quad (2) \\
    f_{\text{max}}(P_1, P_2, ..., P_n) &= \max(P_1, P_2, ..., P_n) \quad (3) \\
    \sum_{i=1}^{n} W_i &= 1 \text{ and } \alpha + (1 - \alpha) = 1 \quad (4)
\end{align*}
\]

\( \alpha \) and \( 1-\alpha \) were chosen according to the relative importance to be given to each aggregation method. The more alpha tends to 0, the more the maximum aggregation is emphasized. Otherwise, additive aggregation tends to be prioritized.

The most favorable weight \( W \) set of parent nodes and aggregation method was determined by performing a calibration that requires knowledge and experience and by building on previous work (82-87) done to assess the potential risk of agricultural activities on water resources using a deterministic approach. This
step was performed by analyzing the outputs resulting from each possible combination of $\propto$ and $W_n$. After validating the most appropriate variables and depending on the number of parent nodes, a 2D or 3D graph was generated to extract the CPDs. These data were used to fill the CPTs. The weights selection and graph generation process were codified in R (V.11.456, RStudio, Inc.). Table in S1 Table presents an example of CPT associated with the “Global risk level (GRL)” node that was determined by combining three parent nodes: probability of hazard (PH), exposure (EXP), and DWI vulnerability level (DWIVL). All probability distributions were calculated based on the 3D graph (Figure B in S1 Fig) generated by combining all possible cases related to the parent nodes.

Table 2 shows the methods applied to fill 18 CPTs. This table also lists the methods that were tested out but were rejected because of unreliable outcomes as explained above.

Table 2: Overview of the CPTs filling methods.

<table>
<thead>
<tr>
<th>Child node</th>
<th>Name</th>
<th>Description</th>
<th>Parent node</th>
<th>Weight of parent node</th>
<th>$\alpha$</th>
<th>Nominated CPTs filling method</th>
<th>Tested and rejected CPTs filling methods</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>CAL</td>
<td>Crop activities level</td>
<td>PPCO DBA</td>
<td>NA(*)</td>
<td>NA</td>
<td>Cain method</td>
<td>Expert judgment - Interpolation method - Mixed aggregation method combined to expert judgment</td>
</tr>
<tr>
<td></td>
<td>AAL</td>
<td>Animal activities level</td>
<td>APCO DBA</td>
<td>NA</td>
<td>NA</td>
<td>Expert judgment</td>
<td>- Cain method - Interpolation method - Mixed aggregation method combined to expert judgment</td>
</tr>
<tr>
<td></td>
<td>CRIM</td>
<td>Contaminants that could be released in the immediate protection zone</td>
<td>PIMPA AIMPZ</td>
<td>NA</td>
<td>NA</td>
<td>Expert judgment</td>
<td></td>
</tr>
<tr>
<td></td>
<td>CRIN</td>
<td>Contaminants that could be released in the intermediate protection zone</td>
<td>PINPA AINPZ</td>
<td>NA</td>
<td>NA</td>
<td>Expert judgment</td>
<td></td>
</tr>
<tr>
<td></td>
<td>EXP</td>
<td>Exposure</td>
<td>PZ</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>CAH</td>
<td>Crops activities hazard</td>
<td>CAL CPC</td>
<td>0.9 0.1</td>
<td>1</td>
<td>Mixed aggregation method</td>
<td>Interpolation method - Mixed aggregation method combined to expert judgment</td>
</tr>
<tr>
<td></td>
<td>APH</td>
<td>Animal production hazard</td>
<td>AAL APC</td>
<td>0.9 0.1</td>
<td>1</td>
<td>Mixed aggregation method</td>
<td>Interpolation method - Mixed aggregation method combined to expert judgment</td>
</tr>
<tr>
<td></td>
<td>PH</td>
<td>Probability of hazard</td>
<td>APH CAH</td>
<td>0.6 0.4</td>
<td>1</td>
<td>Expert judgment</td>
<td></td>
</tr>
</tbody>
</table>
### Child node

<table>
<thead>
<tr>
<th>Name</th>
<th>Description</th>
<th>Parent node</th>
<th>Weight of parent node</th>
<th>α</th>
<th>Nominated CPTs filling method</th>
<th>Tested and rejected CPTs filling methods</th>
</tr>
</thead>
<tbody>
<tr>
<td>MQRW _(x)**</td>
<td>Microbiological quality of raw water</td>
<td>PKEC_(x) GMEC_(x) SVMC_(x)</td>
<td>0.5 0.4 0.1</td>
<td>0.8</td>
<td></td>
<td></td>
</tr>
<tr>
<td>VMC</td>
<td>Vulnerability to microbiological contamination</td>
<td>MQRW_W MQRW_SP MQRW_S MQRW_A</td>
<td>0.25 0.25 0.25 0.25</td>
<td>0.5</td>
<td></td>
<td></td>
</tr>
<tr>
<td>RWQLTP</td>
<td>Raw water quality level for total phosphorus</td>
<td>PKTP MTP</td>
<td>0.2 0.8</td>
<td>0.7</td>
<td></td>
<td></td>
</tr>
<tr>
<td>CVLU</td>
<td>Chemical vulnerability level based on land use</td>
<td>CRIM CRIN</td>
<td>0.7 0.3</td>
<td>0</td>
<td></td>
<td></td>
</tr>
<tr>
<td>VCC</td>
<td>Vulnerability to chemical contamination</td>
<td>RWQLTP CVLU</td>
<td>0.8 0.2</td>
<td>0.7</td>
<td></td>
<td></td>
</tr>
<tr>
<td>DWIVL</td>
<td>DWI vulnerability level</td>
<td>VMC VWS VCC</td>
<td>0.4 0.4 0.2</td>
<td>0.4</td>
<td></td>
<td></td>
</tr>
<tr>
<td>GRL</td>
<td>Global risk level</td>
<td>PH DWIVL EXP</td>
<td>0.8 0.1 0.1</td>
<td>1</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

(*) Not applicable

(**) x refers to season: winter/ Spring/ Summer/ Autumn

#### 2.3.3 Sensitivity analysis

A sensitivity analysis was performed to quantify the magnitude of the sensitivity of the changes that occur in the endpoint probabilities when the parent nodes are changed and to identify which of the variables are most relevant (40, 60). This analysis is necessary to understand where more information is needed to reduce uncertainty.

This sensitivity analysis was executed by calculating the mutual information \( I \) between two discrete variables \( X \) and \( Y \) (93):

\[
I(X,Y) = H(X) - H(X|Y) = \sum_{x\in X} \sum_{y\in Y} p(x,y) \log_2 \frac{p(x,y)}{p(x)p(y)}
\]

(5)
where the sum is over all states \( x \) and \( y \) of variables \( X \) and \( Y \), respectively, \( H(X) \) is the entropy reduction of \( X \) before any new finding, \( H(X|Y) \) is the conditional entropy of \( X \) at given \( Y \). In the case of computing the \( I \) between continuous random variables, the double sum has to be replaced by a double integral.

These measures of entropy reduction and mutual information between various variables and a given endpoint node were conducted with the “Sensitivity to finding” command in the Netica software.

3 Results

3.1 Bayesian network model

The developed BN model integrates 43 nodes (Fig 3) and 1326 conditional probabilities. The nodes constituting this BN are described in Table 1 and Table 2. The probabilities associated with each node whose sum is equal to 100% are shown on the right side of the status window of the detailed BN (Figure C in S1 Fig). Grey nodes present the input variables.

Fig 3. Bayesian network structure for assessing drinking water intake contamination risk from agricultural activities. The parent nodes in gray are associated with input variables that define the properties of a specific drainage basin and variables related to a selected intake.

3.2 DWIs vulnerability

As illustrated in Fig 4, more than 60% of DWIs were extremely vulnerable. This is most likely because these intakes are in a highly urbanised watershed with agricultural activity (section 2.1). The vulnerability of source water to microbiological contamination was noted at the four intakes. The seasonal variation in microbial contamination of raw water was investigated (Figure D in S1 Fig), and our results revealed that all intakes were associated with medium to high probability distributions for all seasons. The probabilities describing poor water quality could be due to the transport of microbial contaminants during the late winter and early spring snowmelt period, but also during episodes of heavy rainfall in the summer, and autumn as demonstrated by Sylvestre, Burnet (29).
DWDI_3 and DWDI_4 were highly vulnerable to chemical contamination (82.3% and 72.2% probability in the high state, respectively) (Fig 4) as a result of total phosphorus leading to lower quality of the raw water. In addition, a significant area of the intermediate zone of DWDI_3 was occupied by anthropogenic activities that may release chemical contaminants (Figure E in S1 Fig). DWDI_1 and DWDI_2 were moderately vulnerable to chemical contamination (55.2% and 65% probability in the medium state, respectively). They were also influenced by the land use upstream of the intakes, as illustrated in Figure E in S1 Fig.

Fig 4. The probability distribution of different categories of vulnerability (global, microbiological contamination, chemical contamination, and scarcity) for all drinking water intakes (DWIs).

3.3 Probability of the hazard

The protection area and drainage basin delineation results showed that the study site includes 25 DBs that discharge within the boundaries of the intermediate protection zones of the four DWIs. The immediate protection areas of all intakes and seven drainage basins (DB_1, DB_3, DB_5, DB_16, DB_17, DB_21, and DB_22) were not occupied by agricultural activities, so they do not constitute a risk from agriculture to water quality at DWIs. The remaining drainage basins and their occupancy by animal and crop production activities are shown in the supplementary information (Figure F to Figure N in S1 Fig). The minimum areas of crop and animal production activities were 0.08 km$^2$ (in DB_20) and 0.03 km$^2$ (in DB_12 and DB_18), respectively, while the DB_8 has the greatest area covered by these activities. The crop activities were spread over 3488.26 km$^2$, and the animal activities were occupied by 3177.22 km$^2$ of this DB. The total area occupying all the DBs upstream of all the DWIs were 8278 km$^2$ in crop production activities and 7830 km$^2$ in animal activities.

Fig 5 shows that the probability distribution of the overall hazard (PH), the crop activity hazard (CAH), and the animal production hazard (APH) for each DB. The probability of hazard was at high risk in DB_8, DB_10 and DB_23 (81.7%, 96.6% and 81.6% probability in the high state, respectively). This outcome is due to the area occupied by agricultural activities and the type of production that presents the greatest risk to water quality: class I for animal production and annual crop for vegetable production. The probability of
hazard was moderate in DB_6, where 51.1% was in the medium state. For DB_2, DB_7, DB_15, and DB_24, the probability of hazard was ranged between low and medium levels with close percentages (for example, 48.7% and 49.6% probability in low and medium states, respectively for DB_2). The remaining DBs were all characterised by low hazard probabilities level, which averaged 63.6%.

Fig 5. Bar plots for BN sub-model assessing the probability distribution of overall hazard, crop activity hazard, and animal production hazard within the drainage basins.

3.4 Global risk

The overall risk from crop and animal activities was relatively high in three drainage basins (DB_8, DB_10, DB_23) whose outlets were located upstream of these DWIs: DWI_2, DWI_3, and DWI_4 (Fig 6). Their probabilities of being in high status vary between 81.9% and 95.5%. Another five DBs (DB_2, DB_6, DB_7, DB_15, and DB_24) represented a moderate overall risk with probabilities ranging from 55.6% to 60.9% (Fig 6). Eight DBs (DB_4, DB_11, DB_12, DB_13, DB_14, DB_18, DB_19, and DB_25) were associated with a low global risk with probabilities that are almost equal between the low and medium status, and with a variance in probabilities that does not exceed 10%. DB_9 and DB_20 were associated with low global risks with probabilities ranging from 54.8 to 67.4% in the low status.

The crops activities hazard (CAH), animal production hazard (APH), exposure (EXP), DWI vulnerability level (DWIVL), vulnerability to microbiological contamination (VMC), and vulnerability to chemical contamination (VCC) are the nodes of interest in this study, so they were selected as findings’ nodes. Subsequently, the sensitivity for the endpoint "global risk level (GRL)" in every DB occupied by agricultural activities was calculated and analyzed as explained in section 2.3.3. Results of this analysis (Fig 7) showed that the developed model was most sensitive to the crops activities hazard (CAH) (from 17.8 to 29.5% variance reduction) and the animal production hazard (APH) (from 6.54 to 25.6%) for all DWIs. This model was also moderately influenced by the exposure node (from 2.5 to 4.13%), especially for the cases of the drainage basins that were slightly occupied by a single type of production activity, either crop or animal, such as BD_9 and BD_20 that are located upstream DWI_2, DWI_3 and DWI_4.
Fig 6. Global risk distribution for all drainage basins upstream four DWIs.

Fig 7. Results of sensitivity of the « Global risk level » variable.

4 Discussion

The probability of DWIs contamination by agricultural activities was assessed using a probabilistic approach in this study. The proposed Bayesian model generates a probabilistic estimation of the DWIs vulnerability level, crop and animal production activities hazard, the exposure, and the overall risk associated with each drainage basin occupied by agricultural activities. This model is a potential tool for source water managers to easily apply since it is based on readily available input data available and raw water analyses commonly required by the regulation on the quality of drinking water (80). This regulation specifies drinking water quality standards for all water systems that are designated for human use.

One of the strengths of this model is the ability to embed different types of data in the same system. For example, to fill the input variables, water quality and land use data were used, and the outputs of other studies (vulnerability to water scarcity) can be included. The discretization of all nodes constituting this model allows for more efficient and effective communication with stakeholders regarding the understanding, the interpretation of risks, and the preparation of action plans since the terminology, statuses are based on regulatory criteria and lexicons already used in the provincial databases. The graphical model can be used as a communication tool between scientists, municipal officials, government professionals, and members of a watershed organization. It could also be a simple approach for farmer organizations that support surface water pollution reduction to convey potential risks to their members. The acyclic graph facilitates the understanding and the collaboration among multidisciplinary members, thus increasing the impact in the decision-making process, as discussed in Stritih, Rabe (94) and Laurila-Pant, Mantyniemi (95).

In testing several methods to fill CPTs, it was concluded that the chosen methodology can significantly influence the final output of the risk assessment. The additive aggregation was applied through the mixed
aggregation by choosing $\alpha = 1$ to fill CPTs related to child nodes of crops activities hazard (CAH), animal production hazard (APH), probability of hazard (PH), and global risk level (GRL). This method is suitable for cases that do not require emphasizing the pessimism bias. For instance, a high level of crop activity (CAL node) may not represent a high hazard (CAH node) in the case where most of the area is covered (CPC node) by undefined crops or non-cultivated land. In another case, it is needed to emphasize the worst-case bias by setting $\alpha$ to 0, which applied to the case of assessing vulnerability to chemical contaminants based on land use (CVLU node) in immediate (CRIM node) and intermediate protection zone (CRIN node).

This means that the maximum aggregation is applied to penalize the worst of inputs to protect the water resource against the highest pressure. However, for the rest of the cases, combining both types of aggregation (additive and maximal) was done due to the evaluation of complementary variables, such as the case of chemical and microbiological water quality assessment based on raw water quality analysis. The mixed aggregation method allowed to include the inter-correlation between the parent nodes, which rendered the risk assessment more unbiased and objective than other methods. Our case study also revealed that the weight assigned to each parent node might reflect its relative significance and reliance on its respective child node.

As shown in Fig 6 and maps of agricultural activities (from Figure F to Figure N in S1 Fig), medium level risk probability dominates for DBs that were slightly occupied by agricultural activities. This was due to the high DWIs vulnerability, and the location of the outlets of the DBs at a distance that does not exceed 10 km upstream from these intakes. This implies that even though the hazard probability is not so significant, the vulnerability of the intake and the exposure of the threat could be the factors contributing to a higher risk level (medium to high). Therefore, all nodes in the model are essential for assessing the risk of agricultural activities. As a result, across a basic network graph, this Bayesian model gives a clearer understanding of the relative relationship and interaction between all variables (96, 97). As anticipated, results proved that the BN could be applied to water resource risk assessment with the same efficiency as environmental risk assessment (38). The developed model provides a transparent and systematic way to
characterize risk through uncertainty. It reveals all states with their estimated probabilities. The Bayesian network has the advantage of being used rapidly to investigate outputs under various scenarios, as mentioned by Dorner, Shi (50).

The sensitivity analysis results showed that it is interesting to focus more on the CAH, APH, and EXP to improve the global risk estimation. For this purpose, the use of data regarding pesticides (types, quantities, method of application and frequency), animal production, and characteristics of the receiving environment (soil type, hydrodynamics of the river, etc.) are beneficial to have a more efficient assessment. These data are not available in provincial databases. To bridge this data gap, targeted analysis are thus needed. Such models are suitable for determining the catchment areas that could potentially result in the highest risk to water intakes. This simple model is handy for authorities and stakeholders to estimate the level of risk on a regional and provincial scale with commonly available data. Therefore, this model is considered a jump-start for stakeholders to channel their efforts and investments properly to the most appropriate target for contamination mitigation efforts. This risk assessment model may benefit from an upgrade over time by including variables that define the modeling of microbial pathogens and chemical contaminants after rain discharge. As shown by Haack, Duris (79), Soupir, Mostaghimi (76), and Jaffrezic, Jarde (77), water bodies were the most sensitive to livestock-derived fecal and chemical marker inputs after rainfall, and bacteria concentrations frequently exceed standards for primary contact.

Discussion of the risk analysis outcomes with stakeholders requires recognizing uncertainties resulting from the lack of data, inaccuracies in laboratory measurements for raw water quality monitoring, gaps in the knowledge, and inexactness of geographic data. Nevertheless, as Uusitalo (39) pointed out, the advantage of the BN is that it incorporates uncertainty explicitly and naturally through the probability distribution.

As expected, this model has a limitation concerning its validation since it is not readily possible to perform a quantitative validation. This limitation is in accordance with the interpretations discussed in reviews of Kaikkonen, Parviainen (38) and Phan, Smart (51). For this reason, in this study, model validation was performed by examining the outcomes of many scenarios in comparison to the regulatory vulnerability
analyses. Results were consistent; however, the BN approach provides more context for risk levels and an
approach to provide an aggregated and global assessment of DWI vulnerability.

These modelling approaches are easily adapted to other drainage basins with risks to their water intakes.
On a larger scale (national and international scale), the model could be adjusted by adding nodes specific
to the new data and those that characterize the study site and the regulations of the aimed country. At this
scale, the involvement of experts is a must for the filling of the new CPTs and the validation of the model.

5 Conclusions

This study used the Bayesian network (BN) to assess the risk of agricultural activities for four DWIs that
draw their water across a river in the south of Quebec, Canada. The findings have led us to conclude that:

- The Bayesian model allowed for the combination and manipulation of data in various formats
  (numerical, statistical, and geographical data) into a consistent methodology for water resource risk
  assessment for drinking water supplies.

- The interactions between all risk components (the vulnerability of drinking water intake, the
  exposure to the threat, and the probability of hazard) can be considered under uncertainty into a
  single model. Stakeholders and agricultural organizations may simply utilize this probabilistic
  reasoning tool to grasp the causal relationships between variables, allowing them to identify the
  key factors that demand special attention due to their large impacts on the outcomes.

- By referencing expert knowledge, the built model may fill data gaps needed for assessing the
  vulnerability of drinking water intake. The explicit identification of data gaps and sensitivity
  analyses facilitates the prioritization of future data collection needs. Furthermore, the model can
  serve as a basis for integrating local knowledge from various stakeholders in the watershed.

- The mixed aggregation method (func-max) combined with expert judgment is a simple and
  straightforward way for populating conditional probability tables, and it is suitable for considering
  parent-child dependencies. The correlation between the parent nodes is also considered, making
the evaluation accurate based on expert judgment. The weights allocated to the additive ($f_{add}$) and
maximum ($f_{max}$) aggregation allow for a more focused examination of the influence factor in the
field of water resource protection. This is a crucial characteristic when the effect of parent nodes
functioning concurrently is considered. This method is useful for filling CPTs semi-automatically
to help experts calculate reasonable probabilities when data are not available.

- The model developed in this paper is the first element of a risk assessment cycle. This Bayesian
model proved successful in identifying drainage basins occupied by agricultural activities that may
pose a medium to high risk of contamination of drinking water intakes. These findings can assist
managers and authorities in developing targeted sampling programs for microbiological and
chemical testing of surface water while reducing costs and prioritizing drainage basins associated
with a high level of risk. The analysis results should constitute the input variables of this Bayesian
network while incorporating new nodes and updating better pinpoint the source of risk. In a
nutshell, this Bayesian model provides an easy-to-use tool for developing evolutionary
management strategies.

From a water resource protection perspective, in an urban and agricultural watershed, this research is
expected to serve as a basis for future studies on assessing microbial risk from combined sewer overflows
(CSOs) using a probabilistic approach.

6 Acknowledgements

We thank the partner municipalities and the watershed organization team responsible for the study area for
contributing to the data collection and vulnerability assessments.

7 References

1. Berens MJ, Capel PD, Arnold WA. Neonicotinoid Insecticides in Surface Water, Groundwater,


l’Environnement et de la Lutte contre les changements climatiques, Direction du suivi de l’état de l’environnement; 2015.


Supporting information

S1 Text. Protection zone delineation.

S1 Fig. Supplementary results.

S1 Table. Conditional probability distribution.
Fig 2

DESIGNING THE BAYESIAN NETWORK (BN)

- Initial construction of the Bayesian network (BN)
  - Defining nodes (variables)
  - Defining node states (discrete, Boolean, continuous)
  - Discretize continuous variables
  - Defining links
  - Filling conditional probability tables (CPTs)

Validating and evaluating the network
- Scenarios testing
- Sensitivity analysis

APPLICATION OF THE BAYESIAN NETWORK MODEL TO THE CASE STUDY

- Data collection and processing in geographic information system
- Protection zones and drainage basins (DBs) delineation
- Risk assessment in Netica

Risk components
- Hazard
- Exposure
- Vulnerability
- Risk

Risk subcomponents
- (indicators, factors, parameters...
Click here to access/download 

**Supporting Information**

*S1_Table. Conditional probability distribution.docx*
Click here to access/download

Supporting Information

S1_Fig. Supplementary results.docx