Schistosomiasis prevention undermined by water point forecasting : An approach based on data fusion model

Teegwende Zougmore[1], Sadouanouan Malo[2], and Bamba Gueye[3]

[1] Université Alioune Diop de Bambey, Bambey, Sénégal
teeg-wende@uadb.edu.sn
[2] Université Nazi Boni, Bobo-dioulasso, Burkina Faso
sadouanouan@yahoo.fr
[3] Université Cheikh Anta Diop , Dakar, Sénégal
bamba.gueye@ucad.edu.sn

Abstract. In this study, we present an approach for called $FuMalMMo$ for Fusion of Machine learning and Mathematical Models. It is an approach based on data fusion which leverages information coming from different datasources to make a decision. The approach we present follows a "Y" pattern where in the left branch, there is a machine learning model in charge of forecasting the water quality of a water point. In the right branch, there is an epidemiological model responsible for making a forecast of the evolution of the density of parasites and snails causing schistosomiasis. In the middle branch, we rely on the theory of belief functions or evidence theory to combine the forecasts made by the two models in order to infer one day ahead the state of infestation of a water point with a precision of 0.75.

Keywords: Data fusion · Water quality prediction · Mathematical epidemiology of infectious diseases · Schistosomiasis

1 Introduction

Schistosomiasis is an acute and chronic parasitic disease caused by trematodes of the genus Schistosoma. The larvae of the parasite, released by intermediate hosts (snails), enter a person's skin when in contact with infested water. The life cycle of the disease transmission involves humans (final host), parasites and intermediate host (snails). The intermediate hosts (snails) live in water points (lakes, dams, rivers, etc.). The water quality of these water points influences their biological cycle as well as that of the parasites [1–3], and consequently plays on the extinction or the persistence of the disease.

We address the spread of schistosomiasis from a health prevention viewpoint. Indeed, we aim to set up an alert system to prevent when a water point is infested. A water point is infested when it contains infected snails that emit parasites. To achieve this, it is necessary to evaluate earlier the quality of water points and the density evolution of snails and parasites. For the earlier assessment of water

quality, we consider a water quality model which can forecast one day ahead the quality of a water point [4]. And for the assessment of the snails and parasites density evolution, we refer to an epidemiological model. To leverage the two sources of information, we formulate an approach which consists of fusion of the results of two mentioned models.

In [5] data fusion is defined as a combination of information originating from several sources in order to improve decision-making. Several methods of data fusion are encountered in the literature. They come essentially from probability theory, evidence theory and fuzzy set theory [6], [7],[8]. The methods are proposed to take into account the characteristics of the information to be combined. The authors in [9] indicate that these different methods are not to be put in competition and are not contradictory. They argue that the choice of one of these methods must be made by finding the best match between the intended application and the specifics of the method.

The information to be fused in our case study present a form of imperfection which is uncertainty. This uncertainty is due to the fact that none of the forecasting models used as a data source is intended to faithfully reflect reality. This results in forecasts with margins of error. Uncertainty is represented and quantified by probability theory [6, 8]. But its use requires a priori probabilities [8]. These a priori probabilities are difficult to determine in our case study. Evidence theory is a data fusion method that does not require knowledge of a priori probabilities [10]. It extends both set theory and probability theory in the representation of uncertainty [11]. This leads us to employ evidence theory in our study.

The rest of the paper is organized as follows: section 2 presents some basic concepts of evidence theory. In Section 3, we describe the proposed data fusion corresponding to our case study. The section 4 explains how evidence is applied to combine data coming from the data-sources. We present the experimental setup and results respectively in section 5 and 6. Section 7 gives a conclusion and some perspectives to be addressed in future work.

2 Background on evidence theory

The theory of belief functions, also known as the Dempster-Shafer theory or the theory of evidence, was proposed by Dempster and then mathematically formulated by Shafer [8, 9]. It is based on modeling the belief in an event. There are four steps to follow in the application of this theory. The first step is modeling which consists of choosing a mathematical representation for information to be combined. The second step consists of quantifying the information. The third and fourth steps are combination and decision which consist of applying rules to synthesize information and take decision.

2.1 Information modeling

Information to be fused is modeled by mass function or basic belief assignment. By setting $D = \{d_1, d_2, ..., d_n\}$, the frame of discernment where each d_i designates

a hypothesis in favor of which a decision can be made, the mass function can be defined on 2^D with values in $[0,1]$ [6, 12]. For a source S_j, the mass function m_j verifies:

$$\sum_{A \in 2^D} m_j(A) = 1 \tag{1}$$

From a mass function, it is possible to derive other functions [9, 13] such as: (i) the credibility denoted "*bel*" which represents the total mass of belief in A and the plausibility "Pl" which is interpreted as the maximum belief in A. We just present the equation of 'Pl" that we need in decision step. "Pl" is defined as follows :

$$pl_j(A) = \sum_{A \cap B} m_j(B) \quad \forall A \subseteq D \tag{2}$$

Mass function can be discounted to take in account the reliability of sources [13]. It is done by introducing a discounting coefficient $\alpha_j \in [0,1]$ [13]. The mass function for all $A \in 2^D$, $A \neq D$ is thus redefined as:

$$\begin{cases} m_j^{'}(A) = (1-\alpha_j) m_j(A), \ A \subset D \\ m_j^{'}(D) = (1-\alpha_j) m_j(D) + \alpha_j \end{cases} \tag{3}$$

2.2 Estimation

Except when an expert expresses his opinion in the form of a mass function directly, in all other cases, there is no generic method to solve this problem [9]. We expose here a mass function deduced from a probability of realization $s \in [0,1]$ of the hypothesis or set of hypotheses A. Thus, we have:

$$\begin{cases} m_j(A) = s, \ A \subset D \\ m_j(\overline{A}) = 1 - s \\ m_j(B) = 0, \ B \neq A \subset D \end{cases} \tag{4}$$

where $s \in [0,1]$, a real is considered as the probability of occurrence of the event A.

2.3 Combination

We present here the Demptser-shafer's rule which is the rule that we have used. It operates a conjunctive combination followed by a normalization. The normalization consists in distributing the mass of the conflict to all the other elements of 2^D except $\emptyset$. We describe this rule by considering only two masse functions, m_1 and m_2.

We propose to distinguish the combined mass function by the notation m_{comb}. The combination performed by the Demptser-shafer's rule is defined as follows:

$$m_{comb}(A) = (m_1 \oplus m_2)(A) = \frac{1}{1-k} \sum_{B_1 \cap B_2 = A} m_1(B_1) m_2(B_2) \tag{5}$$

with $m_{comb}(\emptyset) = 0$ and k is the normalization term is:

$$k = \sum_{B_1 \cap B_2 = \emptyset} m_1(B_1) m_2(B_2) \tag{6}$$

2.4 Decision

The choice of the final decision or hypothesis can be made according to several criteria [6, 13]. The criterion used to determine the final decision in our context is maximum plausibility. Let Dec denote this decision and pl_{comb} the plausibility deduced from the combined mass m_{comb}. Dec can be defined as follows:

$$Dec = argmax\ pl_{comb}(A) \tag{7}$$

3 Proposed data fusion architecture

The proposed architecture comes in three branches in the form of a "Y" as illustrated on figure 1. The left branch is responsible for providing information

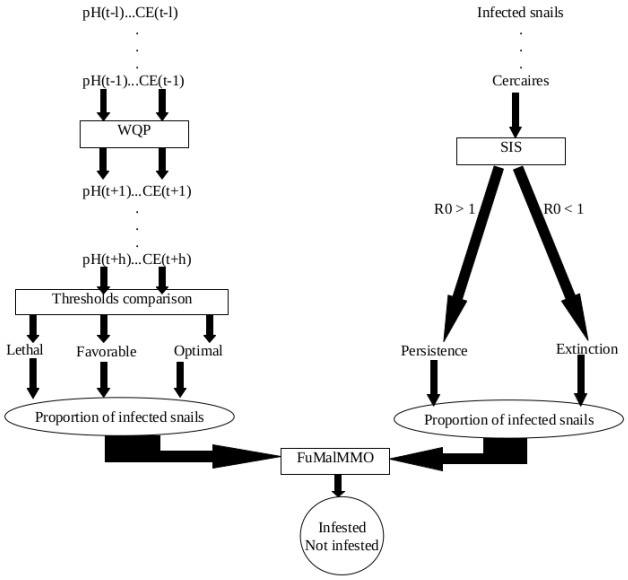


Fig. 1: Data Fusion Architecture

relating to the water quality of a water point. It is based on the water quality forecasting model that we have developed in a previous work [4]. We denote this

model WQP for Water Quality Prediction. The model predicts one day ahead the future values of pH, temperature, dissolved oxygen (DO) and electrical conductivity (EC). Then these predicted values are after compared with thresholds taken from the literature to characterize the water quality as lethal, favorable and optimal. Each category of water quality corresponds to a specific proportion of infected snails. We show in section 4.1 how it is determined.

The right branch is responsible for providing information relating to the evolution of the density of snails and parasites. It is based on an epidemiological model that we have identified in a previous work [14]. We denote this model SIS for Susceptible Infected Susceptible. It is a mathematical model with compartments which takes as input the numbers of human populations, parasites (mirracidia and cercariae) and intermediate hosts (snails) and provides the basic reproduction number R_0. R_0 governs the dynamics of the system. If its value is greater than 1, there is a proliferation of infected snails and the disease persists. If it is less than 1, there is an extinction of the disease. Each condition met by the reproductive number value corresponds to a specific proportion of snails also infected. We show in section 4.1 how it is determined.

The proportions of snails determined according to the output of each model will constitute the data to be combined. The middle branch is responsible for this combination truly speaking itself. This involves performing the four stages of the evidence theory in order to infer the infestation state of a water point. We denote it $FUMalMMO$ for Fusion of Machine learning Model and Mathematical Model.

4 Implementation of evidence theory

4.1 Mass modeling and estimation

The water point will be declared infested or not. This leads us to define a frame of discernment which is made up of two candidate hypotheses : $D = \{d_1, d_2\}$. d_1 is the hypothesis that an observed water point is infested. d_2 is the assumption that it is not. The power set is then $2^D = \{\emptyset, d_1, d_2, d_{1,2}\}$. $d_{1,2} = \{d_1, d_2\}$ and allows translating a part of ignorance on the state of the water point.

Mass function modeling For each possible output of WQP, the defined mass function is m_{wqp}:

$$\{m_{wqp}(d_1), m_{wqp}(d_2), m_{wqp}(d_{1,2})\} \tag{8}$$

With $m_{wqp}(d_1)$ being a numeric value indicating how much WQP believes the water is infested. $m_{wqp}(d_1)$ indicates how much WQP believes that the water point is not infested. $m_{wqp}(d_{1,2})$ allows WQP expressing ignorance as to the probable state of the water point.

And for each output of SIS, the defined mass function is m_{sis}:

$$\{m_{sis}(d_1), m_{sis}(d_2), m_{sis}(d_{1,2})\} \tag{9}$$

$\{m_{sis}(d_1), m_{sis}(d_2), m_{sis}(d_{1,2})\}$ have the same meaning as $\{m_{wqp}(d_1), m_{wqp}(d_2), m_{wqp}(d_{1,2})\}$ but from the point of view of SIS.

$m_{wqp}(\emptyset) = 0$ and $m_{sis}(\emptyset) = 0$ because the discernment framework contains all possible candidate hypotheses. Once the choice of representation has been made, the next step is estimation.

Estimation of mass functions We consider the equation 4 described in section 2.2 to estimate the mass functions. The probability of realization of the hypotheses will correspond to a proportion of infected snails. This proportion depends on the different outputs of WQP and SIS models as described in the architecture in the section 3.

For each output of the WQP model, the mass function is estimated as follows:

$$\begin{cases} m_{wqp}(d_1) = t,\ d_1 \subset D \\ m_{wqp}(d_2) = 1 - t \end{cases} \tag{10}$$

For each output of the SIS model, the mass function is estimated as follows:

$$\begin{cases} m_{sis}(d_1) = t,\ d_1 \subset D \\ m_{sis}(d_2) = 1 - t \end{cases} \tag{11}$$

where t denotes the proportion of infected snails. In the following lines, we indicate how it is determined from each model.

Determination of the proportion of infected snails from WQP It is carried out on the basis of spatio-temporal malacological study. That is to say a study comprising a collection of snails carried out at a given period in different places and followed by an identification of the species then a test for the emission of cercariae. To all this is added an analysis of the physicochemical parameters to establish a correlation with the proportion of infected snails.

The infected snails proportion is determined by the following formula :

$$t^q = \frac{\sum_{i=1}^{n} P_i^q}{n} \tag{12}$$

where t^q indicates the arithmetic mean of infected snails proportion corresponding to water quality q. q being a flag which designates one of the possible output of WQP : lethal, favorable and optimal. P_i^q indicates the infected snails proportion found on a specific place i and corresponding to water quality q.

We selected a study which took place in Tanzania (at Lake Victoria) and which focused on snails belonging to the genus Biomphalaria which is the genus responsible for intestinal schistosomiasis [15]. The study took place from February 2016 to March 2016. Sixteen places were explored. For each category of water quality, we calculate the arithmetic mean of the proportions. Thus, we obtain $t^{optimal} = 63.68\%$ for the optimal category and $t^{favorable} = 32.43\%$ for the favorable category and $t^{lethal} = 0\%$.

Once the proportion has been determined for each category, the resulting mass functions are deduced by the equation 4. We present the results in table 1.

Table 1: Estimated mass functions for each output of WQP

Water quality category	$m_{wqp}(d_1)$	$m_{wqp}(d_1)$	$m_{wqp}(d_{1,2})$
Lethal	0	1	0
Favorable	0.32	0.68	0
Optimal	0.64	0.36	0

The values in the table 1 are interpreted as follows : when water is lethal, a mass function of 1 is assigned to hypothesis d_2. This is interpreted as follows: when the water is lethal, there can be no infestation. When the water is favorable, a mass function of 0.3243 is assigned to the hypothesis d_1 and a mass function of 0.6757 is assigned to the hypothesis d_2. This is interpreted as follows: when the water is favorable, the chances that the point is infested are very low, but it is not excluded that it is. When water is optimal, a mass function of 0.64 is assigned to the hypothesis d_1 and a mass function of 0.36 is assigned to the hypothesis d_2. This is interpreted as follows: when the water is optimal, the chances that the water point is infested are high but nothing excludes that it is not.

Determination of the proportion from SIS The proportion of infected snails is determined based on the value of the basic reproduction number R_0. When :

- $R_0 < 1$, the disease will die out. This means that the system is in a state where the infected compartments do not have enough individuals for the disease to spread. This state is translated by the solution of the system which is $\epsilon^0 = (H0, 0, 0, M, 0, 0)$. Which means that the compartments of susceptible humans and snails have individuals. But there are neither parasites, neither infected individuals. From this we derive a proportion of infected snails $t = 0\%$.
- $R_0 > 1$, the disease will spread. This means that the system is in a state where there are enough individuals in the infected compartments to allow disease transmission. This state is translated by a solution of the system which is $\epsilon^* = (HS^*, HI^*, K^*, MS^*, MI^*, P^*)$. This solution is not an explicit expression like ϵ^0. To determine a proportion of infected snails, it is necessary to:
  - launch a numerical simulation of the system;
  - and calculate a proportion when the variations of the different compartments become insignificant. The calculation is done with the following formula:

    $$t = M_I^*/M^* \tag{13}$$

    with M^* representing the total number of snails observed and M_I^* the number of snails emitting cercariae. For the numerical resolution, we use the simulation data of the model found in [16]. We thus obtain $t = 85\%$.

Here also, once the proportion has been determined for each value of R_0, the resulting mass functions are deduced by the equation 4. We present the results in table 2.

Table 2: Estimated mass functions for each output of SIS

R_0 threshold	$m_{sis}(d_1)$	$m_{sis}(d_1)$	$m_{sis}(d_{1,2})$
$R_0 < 1$	0	1	0
$R_0 > 1$	0.85	0.15	0

The values in the table 1 are interpreted as follows : when $R_0 < 1$, a belief mass of 1 is estimated for the hypothesis d_2. What results in the fact that there cannot be infestation when there is extinction of the disease. When R_0 is greater than 1, a mass of 0.85 is estimated for the hypothesis d_1 and a mass of 0.15 for the hypothesis d_2. This translates into the fact that when the disease persists, there is a very good chance that the water point will be infested. But nothing excludes that this is not the case.

Discounting of mass functions Our sources are forecast models. This indicates that there may be discrepancies between the predicted and expected values. To take into account the errors of these models, we propose to determine an discounting coefficient α by the following formula:

$$\alpha = 1 - R^2 \tag{14}$$

with R^2 denoting the coefficient of determination. We therefore obtain for the source WQP:

$$\alpha_{wqp} = 1 - R^2_{wqp} \tag{15}$$

and for source SIS:

$$\alpha_{sis} = 1 - R^2_{wqp} \tag{16}$$

Once the discounting coefficients are determined, the equation 3 is applied. The masses of belief become thus:

$$\begin{cases} m^{'}_{wqp}(d_1) = (1 - \alpha_{wqp}) m_{wqp}(d_1) \\ m^{'}_{wqp}(d_2) = (1 - \alpha_{wqp}) m_{wqp}(d_2) \\ m^{'}_{wqp}(d_{1,2}) = (1 - \alpha_{wqp}) m_{wqp}(d_{1,2}) + \alpha_{wqp} \end{cases} \tag{17}$$

And

$$\begin{cases} m^{'}_{sis}(d_1) = (1 - \alpha_{sis}) m_{sis}(d_1) \\ m^{'}_{sis}(d_2) = (1 - \alpha_{sis}) m_{sis}(d_2) \\ m^{'}_{sis}(d_{1,2}) = (1 - \alpha_{sis}) m_{sis}(d_{1,2}) + \alpha_{sis} \end{cases} \tag{18}$$

We use the symbol "$^{'}$" to designate the discounted masses.

4.2 Combination of mass functions

The Demptser-shafer combination rule described by the equation 5 is applied to obtain the mass set overall after the merger. The normalization term k obtained is equal to:

$$k = m^{'}_{wqp}(d_1) m^{'}_{sis}(d_2) + m^{'}_{wqp}(d_2) m^{'}_{sis}(d_1) \tag{19}$$

We thus obtain the combined masses of belief:

$$\begin{cases} m_{comb}(d_1) = \frac{1}{1-k}[m^{'}_{wqp}(d_1)m^{'}_{sis}(d_1) + m^{'}_{sis}(d_1)m^{'}_{wqp}(d_{1,2}) + m^{'}_{wqp}(d_1)m^{'}_{sis}(d_{1,2})] \\ m_{comb}(d_2) = \frac{1}{1-k}[m^{'}_{wqp}(d_2)m^{'}_{sis}(d_2) + m^{'}_{sis}(d_2)m^{'}_{wqp}(d_{1,2}) + m_{wqp}(d_2)m_{sis}(d_{1,2})] \\ m_{comb}(d_{1,2}) = \frac{1}{1-k}[m^{'}_{wqp}(d_{1,2})m^{'}_{sis}(d_{1,2})] \end{cases} \quad (20)$$

4.3 Decision of mass function

At this stage, there is a calculation of plausibilities which is carried out with the equation 2. We obtain :

$$\begin{cases} pl_{comb}(d_1) = m_{comb}(d_1) + m_{comb}(d_{1,2}) \\ pl_{comb}(d_2) = m_{comb}(d_2) + m_{comb}(d_{1,2}) \\ pl_{comb}(d_{1,2}) = m_{comb}(d_1) + m_{comb}(d_2) + m_{comb}(d_{1,2}) = 1 \end{cases} \quad (21)$$

Then we retain the hypothesis with the maximum plausibility as the decision to be made.

$$Dec = max(pl_{comb}(d_1), pl_{comb}(d_2)) \quad (22)$$

In the final result, we want the water point to be classified as "infested" or "not infested", i.e. there is no reject. Consequently, the final decision is "infested" if the plausibility of the hypothesis d_1 is greater than that which corresponds to d_2 whatever the plausibility of $d_{1,2}$.

5 Evaluation

5.1 Experimental setup

In this section, we present the required data and the testing process.

Required data. Carrying out the experiment requires data which can be used can be served as inputs for WQP model and SIS model.

A study providing this necessary data was conducted in [17]. It was carried out from January 2016 to May 2017 in Panamasso and focused, among other things, on a parasitological study and a malacological study associated with an analysis of the physicochemical parameters of a water point.

The parasitological study took place precisely in January 2016 and revealed that he human prevalence rate of intestinal schistosomiasis is 27.47% and the total size of the population which is 3065 inhabitants. The malacological study consisted of collecting snails of the genus Biomphalaria (responsible for intestinal bilharziasis) and in taking physicochemical parameters such as PH, temperature (TEMP), electrical conductivity (EC) and dissolved oxygen (DO). The different collections took place in three seasons: winter (June 2016 - November 2016), cold (December 2016 - February 2017) and hot (March 2017 - May 2017). Each collection phase was spread over 10 days.

For our test, we are interested in two seasons, the cold season and the hot season. We summarize the data of these two seasons in the table 3. In addition to the physicochemical parameters, the density of snails collected and the proportion of infected snails are indicated.

Table 3: Average of physicochemical parameters and malacological data

			Biomphalaria	
Season (Collection month)	Parameter	Value	Density (/30 min)	Proportion of infected snails
Cold (December 2016)	PH Temp(°C) CE OD (%)	7 24 125 75	64.62	40.04%
Hot (March-April 2017)	PH Temp(°C) CE OD (%)	6 27 220 35	187.66	59.18%

In addition to these data relating to the specific case of the Panamasso water point, other data relating to "biological" and "contextual" parameters are needed. For the first type, the values are taken from the literature [16]. For the contextual parameters, we used the 2016 statistical yearbook of Burkina Faso [18]. These are the human natural birth rate (46/1000) and the life expectancy (56.7 years) to determine the recruitment rate and the human natural death rate respectively.

Testing process. We proceed by:

- defining some scenarios from the data presented in the table 3 and by simulating the application of certain control methods that break the cycle of transmission of the disease. The different scenarios established are presented in the table 4.

Table 4: Scenarios determined on the basis of physicochemical parameters and malacological and parasitological data.

Scenario	Description	Expected situation
1	None of humans recovers	Infested Water
2	Full recovery of illned humans (100%)	Infested Water
3	Removal of mollusks via molluscide	Uninfested water
4	Environmental Pollution	Uninfested Water

- an execution of WQP and SIS models as follows :
  - the SIS model takes as input the December population densities and the necessary parameters; it then provides a R_0 valid for the period from January 2017 to March 2017.
  - the WQP model takes as input all the physicochemical parameters of the past two days; we start with the last two days of December 2016. Then the last day of December 2016 and the 1st day of January 2017,

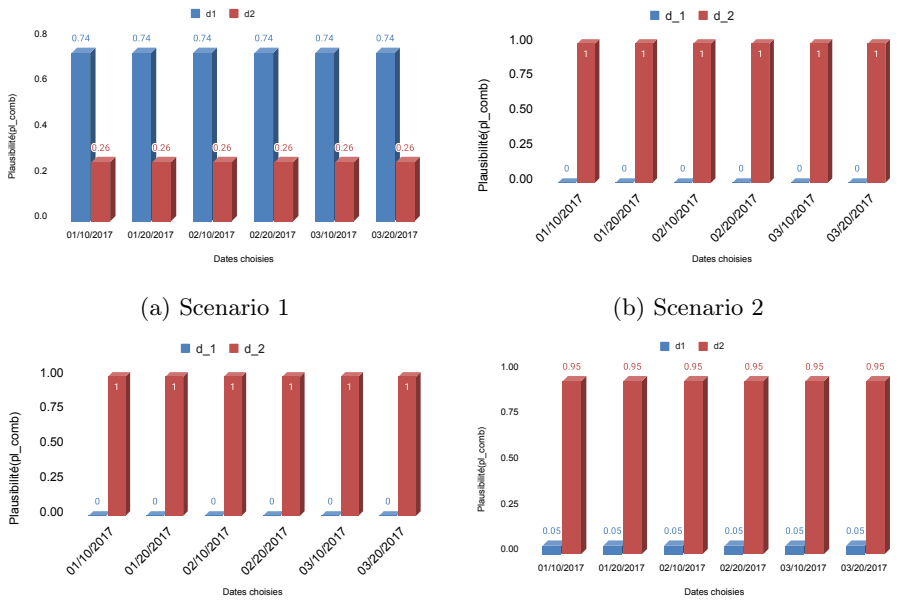
and so on until we cover the entire period January 2017 to March 2017. For each observation of the two last days, it provides the next day's water quality.

– an execution of the $FuMalMMO$ model which consists of applying the evidence theory as described in the section 4 on the forecasts of water quality and the evolution of the density of snails. We then randomly choose six dates in the period from January 2017 to March 2017. And we perform the fusion taking into account the value of R_0 of the period and the water quality of the water point forecasted for each of these dates.

6 Results

6.1 Infestation states observed for each scenario

The basic reproduction numbers obtained for the different scenarios are as follows: $R_0 = 39$ for scenario 1 and scenario 4. $R_0 = 0$ for scenario 2 and scenario 3. We present the results obtained for the different scenarios in table 2a, 2b, 2c and 2d.



(a) Scenario 1

(b) Scenario 2

(c) Scenario 3

(d) Scenario 4

Fig. 2: Plausibilities obtained after combination

For the scenario 1, all the chosen dates indicate a situation of an infested water point. This result is consistent with the expected situation as mentioned

in the table 4. For the scenario 2, the chosen dates indicate a situation of an uninfested water point. These results are contrary to the expected situation mentioned in the table 4. For the scenario 3, all the chosen dates indicate a situation of an uninfested water point. But at this time, the results are consistent with the expected situation. For the scenario 4, we obtain a situation of uninfested water point for the chosen dates. These results are consistent with the expected situation mentioned in the table 4.

These results are obtained from the quality of the water point forecasted on the chosen dates and the value of R_0 of the period from January 2017 to March 2017. In the table 5, 7, 6 and 8, it is presented the different operations performed to infer the infestation state of water on a chosen date. The date 2017-03-10 is taken as an example. The same operations are done for the other chosen dates. The column headers of the tables indicate the different candidate hypotheses. The row headers represent respectively the mass functions assigned by WQP and SIS as well as the combination and decision operations. The contents of the tables are different according to the mass functions assigned following the prediction results of the two models. In this scenario 1, the water quality forecast is favorable and the basic reproduction number R_0 is equal to 39. In this scenario 2, the water quality forecast is favorable and the basic reproduction number R_0 is equal to 0. In this scenario 3, the water quality forecast is favorable and the basic reproduction number $R0$ is equal to 0. And in the scenario 4, the water quality forecast is lethal and the basic reproduction number $R0$ is equal to 39.

Table 5: mass functions obtained on date of 2017-03-10 for scenario 1

	d_1	d_2	$d_{1,2}$
m_{wqp}	0.32	0.67	0.01
m_{sis}	0.85	0.15	0
m_{comb}	0.74	0.26	0
pl_{comb}	0.74	0.26	1

Table 6: mass functions obtained on 2017-03-10 for scenario 3

	d_1	d_2	$d_{1,2}$
m_{wqp}	0.32	0.67	0.01
m_{sis}	0	1	0
m_{comb}	0	1	0
pl_{comb}	0	1	1

Table 7: mass functions obtained on 2017-03-10 for scenario 2

	d_1	d_2	$d_{1,2}$
m_{wqp}	0.32	0.67	0.01
m_{sis}	0	1	0
m_{comb}	0	1	0
pl_{comb}	0	1	1

Table 8: mass functions obtained on 2017-03-10 for scenario 4

	d_1	d_2	$d_{1,2}$
m_{wqp}	0	0.99	0.01
m_{sis}	0.85	0.15	0
m_{comb}	0.05	0.95	0
pl_{comb}	0.05	0.95	1

6.2 Overall performance of the $FuMalMMO$ model

The $FuMalMMO$ model relies on two forecasting models to infer the infestation status of a water point. It thus behaves like a binary classification model. We evaluate here its performance in relation to the different scenarios defined in order to assess its ability to infer in advance the state of infestation of a water

point. To do this, we use different measures that are used to quantitatively assess the results obtained.

The water point studied is either in an infested state or not. We call the infested state positive and the uninfested state negative. A correctly classified state is considered true positive or true negative. On the contrary, a misclassified condition is either a false positive or a false negative.

True positives and false positives resulting from our method are compared against expected situations. To assess performance, we adopt the following metrics: accuracy, precision and recall. We do not go further on how these metrics are calculated. One can refer [19] for more information.

The figure 3 tells us about the overall performance bale according to these three metrics. We also calculated these metrics for each of the models involved in the fusion taken separately. To do this, it was considered that each model taken individually should decide on its own, the state of the water point based on the hypothesis having received the greatest plausibility among those it has issued.

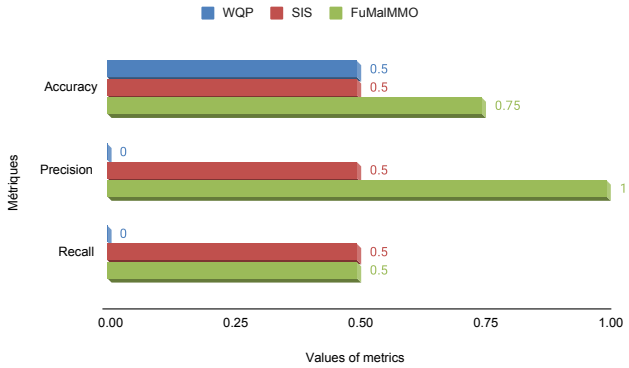


Fig. 3: Comparison of FuMalMMo against WQP and SIS

The x-axis of the graph in figure 3 represents the values of the metrics. On the y-axis, we have the different metrics used for the evaluation of the models.

With the data fusion approach of the two WQP and SIS models, we end up with a $FuMalMMO$ model which is able to classify in advance with great confidence the cases of infestation (Precision $= 1$). On the other hand, when it comes to classifying all the positive cases in advance, its performance is average (recall=0.5). But we note that it is able to classify in advance all types of cases with good accuracy (0.75). By comparing the values of the metrics of $FuMalMMO$ with those obtained from WQP and SIS, we can note that each model taken separately has modest performances compared to those of the fusion model which combines the two.

7 Conclusion

Data fusion makes it possible to leverage several pieces of information in order to make a better decision. In this paper, we have shown how to use data fusion

to assess earlier if a water point harbors infected snails. The fusion method used is the evidence theory or belief functions theory.

Using this theory, we achieved the fusion of information from two complementary models. One is a machine learning model forecasting water quality and the other is an epidemiological model forecasting the density evolution of snails and parasites.

With a recall of 0.5, an accuracy of 0.75 and a precision of 1, the fusion approach leads to an efficient model to warn in one day ahead that a water point is likely to be infested with parasites causing schistosomiasis.

In this work, the mass functions assigned by the water quality prediction model are fixed. We think it would be interesting to consider determining them dynamically. That is to say, instead of categorizing the water quality into lethal, favorable and optimal and then determining the belief masses afterward, we will try to determine these mass functions via a regression model.

References

1. Emile Abdel Malek. Factors conditioning the habitat of bilharziasis intermediate hosts of the family Planorbidae. *Bulletin of the World Health Organization*, 18(5-6):785–818, 1958.
2. Gbocho Yapo Félicien, Diakité Nana Rose, Akotto Odi Faustin, and N'Goran Kouakou Eliézer. Dynamique des populations de mollusques hôtes intermédiaires deschistosoma haematobium et schistosoma mansoni dans le lac du barrage de taabo (sud côte d'ivoire). *Journal of Animal &Plant Sciences*, 25(3):3939–3953, 2015.
3. Influence des paramètres physico-chimiques sur la répartition spatiale des mollusques hôtes intermédiaires des schistosomes humains dans le delta du fleuve sénégal. 29.
4. Teegwende Zougmore, Bamba Gueye, and Sadouanouan Malo. An ai-based approach to the prediction of water points quality indicators for schistosomiasis prevention. In *2022 IEEE Multi-conference on Natural and Engineering Sciences for Sahel's Sustainable Development (MNE3SD)*, pages 1–6, 2022.
5. BLOCH (I.); MAITRE (H.). *01 - Fusion de données en traitement d'images: modèles d'information et décisions*, volume 11. GRETSI, Saint Martin d'Hères, France, 1994.
6. Isabelle Bloch. Fusion d'informations numériques: panorama méthodologique. *Journées Nationales de la Recherche en Robotique*, 2005:79–88, 2005.
7. Bahador Khaleghi, Alaa Khamis, Fakhreddine O. Karray, and Saiedeh N. Razavi. Multisensor data fusion: A review of the state-of-the-art. *Information Fusion*, 14(1):28–44, January 2013.
8. Chaza Chahine. *Fusion d'informations par la théorie de l'évidence pour la segmentation d'images*. PhD thesis, Université Paris-Est; École Doctorale des Sciences et de Technologie (Beyrouth), 2016.
9. Eric Lefèvre. *Fonctions de croyance: de la théorie à la pratique*. PhD thesis, 2012.
10. Mokhtar Bouain. *Système embarqué de fusion multi-capteurs pour la détection et le suivi d'obstacles statiques et dynamiques*. PhD thesis, Université de Valenciennes et du Hainaut-Cambresis, 2019.

11. Thierry Denœux. Théorie des fonctions de croyance: application en reconnaissance de formes et en fusion d'informations, 2010.
12. Arnaud Martin. La fusion d'informations. *Polycopié de cours ENSIETA-Réf*, 1484:117, 2005.
13. Mihai Cristian Florea, Anne-Laure Jousselme, Éloi Bossé, and Dominic Grenier. Robust combination rules for evidence theory. *Information Fusion*, 10(2):183–197, 2009.
14. Teegwende Zougmore, Sadouanouan Malo, Bamba Gueye, and Stanislas Ouaro. Toward a data fusion based framework to predict schistosomiasis infection. In *2020 IEEE 2nd International Conference on Smart Cities and Communities (SCCIC)*, pages 1–8, 2020.
15. Antje Fuss, Humphrey Deogratias Mazigo, and Andreas Mueller. Malacological survey to identify transmission sites for intestinal schistosomiasis on ijinga island, mwanza, north-western tanzania. *Acta tropica*, 203:105289, 2020.
16. Bakary Traoré, Ousmane Koutou, and Boureima Sangaré. Global dynamics of a seasonal mathematical model of schistosomiasis transmission with general incidence function. *Journal of Biological Systems*, 27(01):19–49, February 2019. Publisher: World Scientific Publishing Co.
17. Frédéric Ragnanguénéwindé COMPAORE. *Analyse spatio-temporelle des facteurs environnementaux et socio-sanitaires favorables à la persistance des maladies liées à l'eau : cas de la schistosomiase au Burkina Faso*. PhD thesis, Iinstitut International d'Ingénierie de l'Eau de L'Environnement, July 2018.
18. FASO BURKINA. Institut national de la statistique et de la demographie. *Annuaire statistique de l'économie et des finances*, 2016.
19. Bradley J Erickson and Felipe Kitamura. Magician's corner: 9. performance metrics for machine learning models, 2021.