American Journal of Biological and Environmental Statistics
2022; 8(4): 102-111
http://www.sciencepublishinggroup.com/j/ajbes
doi: 10.11648/j.ajbes.20220804.12
ISSN: 2471-9765 (Print); ISSN: 2471-979X (Online)

Development of a Groundwater Quality Prediction Model for the M'pody Village of Anyama

Meless Djedjro Franck-Renaud[1, 3, *], Gbagbo Tchape Aubin[1], Kpaibe Sawa Andre Philippe[1, 2], Yapo Toussaint Wolfgang[1], Kouassi-Agbessi Therese Brah[1], Amin N'cho Christophe[1, 2]

[1]National Institute of Public Hygiene, Abidjan, Ivory Coast

[2]Department of Analytical Chemistry - Bromatology, General Chemistry, Mineral Chemistry, Pharmaceutical and Biological Sciences Training and Research Unit, Felix Houphouet-Boigny University, Abidjan, Ivory Coast

[3]Department of Physics, Biophysics, Mathematics, Statistics and Computer Science, Pharmaceutical and Biological Sciences Training and Research Unit, University Felix Houphouet-Boigny, Abidjan, Ivory Coast

Email address:

melessrenaud@gmail.com (Meless Djedjro Franck-Renaud), christopheamin@gmail.com (Amin N'cho Christophe), aubintg2007@gmail.com (Gbagbo Tchape Aubin), agbessitherese@gmail.com (Kouassi-Agbessi Therese Brah), twolfgang2y@gmail.com (Yapo Toussaint Wolfgang), Andre.kpaibe@gmail.com (Kpaibe Sawa Andre Philippe)

[*]Corresponding author

To cite this article:

Meless Djedjro Franck-Renaud, Gbagbo Tchape Aubin, Kpaibe Sawa Andre Philippe, Yapo Toussaint Wolfgang, Kouassi-Agbessi Therese Brah, Amin N'cho Christophe. Development of a Groundwater Quality Prediction Model for the M'pody Village of Anyama. *American Journal of Biological and Environmental Statistics*. Vol. 8, No. 4, 2022, pp. 102-111. doi: 10.11648/j.ajbes.20220804.12

Received: September 22, 2022; **Accepted**: November 7, 2022; **Published**: November 16, 2022

Abstract: *Context:* In the village of M'pody in the Anyama district, located about 60 kilometers from the town of Anyama, a diarrhea epidemic was detected in January 2020 and affected 69 people, mostly children aged 0 to 5 years. According to the affected population, these cases of diarrhea were related to the consumption of water from the improved village water system, which had not been maintained for nearly three years. The objective of this work was to develop a bacteriological characterization model of the water table in the village of M'pody (Ivory coast) based on physicochemical parameters and meteorology in order to estimate the concentration of indicator germs of fecal pollution (Escherichia coli) by well. *Methods:* The methodology consisted of four water sampling campaigns per well during the year's four seasons on all 72 wells in this region, for a total of 288 visits. Conventional physico-chemical parameters were determined using electrochemical and spectrophotometric methods. Bacteriological parameters were determined by the membrane filtration technique. A sanitary inspection was also carried out. The development of the prediction model for the Escherichia coli indicator was performed using a linear mixed model. The performance of our model was evaluated by bootstrap and k-fold cross-validation techniques. *Results:* The mixed linear model with random intercept (log transformation) chosen following the spaghetti plot and likelihood ratio test gave the following results: The predictive model explained 30,24% of the variance in Escherichia coli concentrations (log transformation). It is based on 9 variables. Validation of the model performance by bootstrap gave us a very low relative bias < 5%, average prediction errors (RMSE) and absolute prediction errors per K-fold lower than 2,5. *Conclusion:* The development of the statistical model for predicting concentrations of fecal pollution indicator bacteria in wells was made possible by the existence of reliable databases. These databases made it possible to use 9 explanatory variables in a scientific approach to explaining the variable explained Escherichia coli. The validation of the predictive performances by K-fold and bootstrap showed that the model predictions are accurate and the bootstrap estimates of the parameters are unbiased. This implemented model could be used in the event of a declaration of waterborne diseases in this locality before the results of the microbiological analysis are returned.

Keywords: Mixed Linear Model, Bootstrap, Principal Component Analysis

1. Introduction

Water is an essential resource for basic human needs and the environment [1]. Indeed, it is used in various fields such as maritime transport, agriculture (soil irrigation), industry

(cooling of thermoelectric power stations), aquatic recreation and especially the collective or individual supply of drinking water, Usable for domestic and hygiene purposes [2]. Among the sources of drinking water supply, groundwater is traditionally the preferred water source for drinking water, as it is more protected from pollution than surface water [3]. In Côte d'Ivoire, in both rural and urban areas, some populations use drinking water that generally comes from groundwater, particularly from traditional wells [4]. However, these aquifers can be polluted by human activities and be responsible for waterborne diseases [2, 5].

Studies by Coulibaly et al., Fofana, Ahoussi et al., and Yapo et al. in the urban area of Côte d'Ivoire [6-9] Had shown anthropogenic chemical and bacteriological pollution of domestic wells in precarious neighbourhoods of Abidjan.

In the M'pody village in the Anyama district, about 60 kilometres from the city of Anyama, an epidemic of diarrhea was detected in January 2020 and affected 69 people, most of them children from 0 to 5 years old. According to the population concerned, these cases of diarrhea are linked to the consumption of water from the village water supply system (HVA) that has not been maintained for nearly 3 years [10]. In Côte d'Ivoire, the National Institute of Public Hygiene is the structure authorized to control the quality of drinking water through the hygiene laboratory. This control involves microbiological and physicochemical characterization. The results of the analysis of the majority of the physicochemical parameters are available in a very short period of time whereas those of microbiology require a longer time often more than 24 hours [11]. Given these delays, water will have already been consumed by the population when the test results become available. Moreover, since most standards of potability are based on faecal coliforms in particular Escherichia coli (*E. coli*), we thought it interesting to develop tools for detecting the health risk associated with this parameter before microbiological confirmation 24 to 48 hours. This tool could be used for use in case of a water-borne disease report in this locality and will allow for early intervention before the results are available: for example, the closure of wells, the use of calcium or sodium hypochlorite or instruction to boil water before consumption.

2. Materials and Methods

2.1. Study Setting

The locality of M'pody belongs to the sub-prefecture of Anyama, a suburb of Abidjan located in southern Côte d'Ivoire (Figure 1). Anyama covered an area of 114 km^2 and had an estimated population of 148,962 according to the 2014 General Census of Population and Housing (RGPH) [12]. The inhabitants of the village of M'pody obtained their water supply from wells and improved village hydraulics (HVA) [10]. The village had 2731 inhabitants [12].

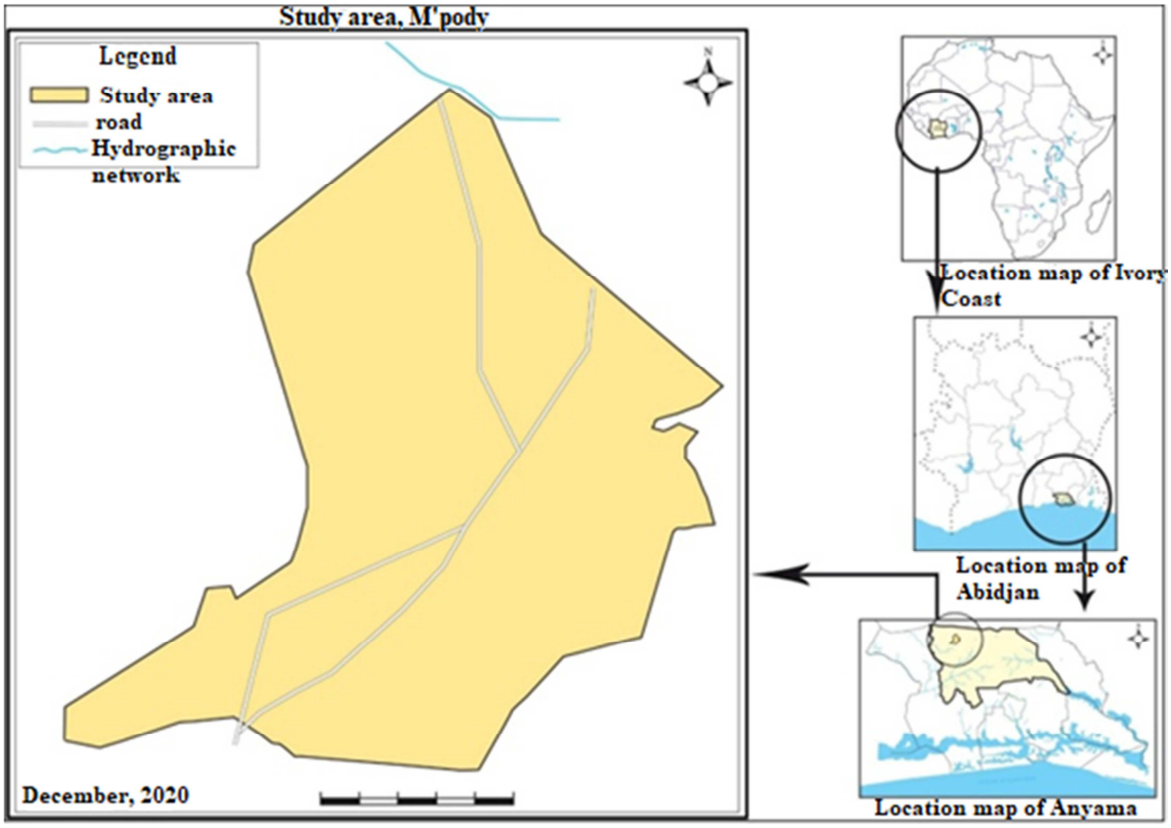

Figure 1. *Study Area Presentation.*

2.2. Schema of Study

It is an ecological study of retrospective cohort type carried out on all the 72 wells that this locality counts during the four seasons of the year 2020 (big dry season (February) and rainy (June), small dry season (August) and rainy (October).

2.3. Material

In this study we used:

1) Physicochemical and microbiological data measured on all 72 wells of the village of M'pody during the four campaigns by the water laboratory of the National Institute of Public Hygiene of Abidjan (INHP);
2) Meteorological data for the city of Anyama for the year 2020 recorded by the Société d'exploitation et de développement aéroportuaire, aéronautique et météorologique [13];
3) Health inspection data collected at the end of all campaigns.

2.4. Methods

2.4.1. Sample Collection, Transportation and Retention

Four samples were taken from each of the 72 wells. These samples were taken for each well on a specific date for each season. The samples were stored in a cooler containing cold accumulators, protected from light, at a temperature between 4°C and 8°C and transported to the laboratory in accordance with the cold chain.

2.4.2. Measurement of Variables

The classical physico-chemical parameters were determined using electrochemical and spectrophotometric methods. The microbiological analysis was carried out by the technique of filtration on membrane then plating on specific medium. Concerning the visual inspection, a questionnaire was administered to the heads of household by the inspectors for the analysis of potential sources of pollution. This questionnaire consisted of 30 dichotomized or nominal qualitative variables that highlighted risk factors for well contamination (presence of pollution sources, distance wells septic tanks (m), distance of latrine wells (m), presence of agriculture/human waste/breeding sites around wells, presence of household activities around wells, etc.) and protective factors (presence of curbstones, presence of protective fences, presence of concrete slabs, internal control and frequency of well maintenance,).

2.4.3. Variables

The dependent variable taken into account for the modelling is the concentration of faecal coliform (*E. coli*). The independent or explanatory variables are 21 chemical parameters, 3 physical parameters, 9 meteorological parameters and 30 explanatory health inspection variables.

2.5. Statistical Analysis

The development of this model was carried out in four stages:

1) Reduction of the number of variables.
2) Choice of explanatory variables to feed the model.
3) Implementation of the linear mixed model.
4) Validation of model performance.

2.5.1. Reducing the Number of Variables

For the quantitative variables we used a Principal Component Analysis. Since each variable was measured in its unit, we first created a matrix of reduced centered data:

1) Determine the correlation matrix to study the pairwise correlation between the independent quantitative variables and then the standardized distance matrix.
2) Check if the matrix was not singular by calculating the determinant which must be different from 0.
3) Check also that the matrix was not an identity matrix by Bartlett's sphericity test.
4) Determine the number of components to extract according to the Kaiser rule or the eigenvalue scree.
5) Project the wells in the selected factorial designs.
6) Project the explanatory variables on the selected components (correlation circle).
7) Project the explained variable (*E. coli*) onto the correlation circle.
8) Regress the *E. coli* concentration on the selected components.
9) The components with a $p < 0.05$ were used for the modeling.

For the qualitative variables, we used clustering. This one helped us to reduce the number of categorical variables that will enter the final model.

2.5.2. Variable Selection

The variables retained from the PCA were those that contributed to the formation of components with a $p < 0.05$. A bivariate analysis was carried out on these variables by setting a p-value threshold of 20% for the choice of quantitative explanatory variables. For the qualitative variables, the choice of those included in the model was made following the clustering by determining the variables that distinguished the clusters. The Akaike criterion in the step-by-step descending procedure of the variables resulting from the PCA and the clustering allowed us to make the final choice of the explanatory variables. The literature review completed the choice of explanatory variables.

2.5.3. Linear Mixed Effects Models

We made a spaghetti plot to choose the model with either random slope or random intercept or random slope and intercept. Transform the *E. coli* concentration variable into a logarithm if possible. The model could be written as follows if it is a random intercept and slope model:

$$Y_{i,t}=(\alpha_0+ \alpha_{0i}) +(\beta_1+ \beta_{1i})*ti+\xi_{i,\,t} \quad (1)$$

Where,

$Y_{i,t}$: Concentration In *E. coli*, From well i to Season ti;

α_0 and β_1: Fixed effects;

α_{0i} and β_{1i}: Random errors, $\alpha 0i$ N (0, σ^2_0), β_{1i} N (0, σ^2_1), cov (α_{0i}, β_{1i}) = σ^2_{01};

$\xi_{i,t}$: Random errors, $\xi_{i,t}$ N (0, σ^2_ξ).

2.5.4. Validation of Model Performance

To evaluate the performance of the model, we used bootstrap and K-fold cross-validation techniques.

(i). Bootstrap Validation

The specific steps in the Bootstrap validation of the model are as follows. [14]:

1) Generate a bootstrap sample by resampling from the data and/or the estimated model;
2) Obtain estimates for all model parameters for the bootstrap sample;
3) Repeat steps B (1000) times to obtain the bootstrap distribution of parameter estimates and calculate the mean, standard deviation and 95% CI of this distribution.

Let $\widehat{\theta_b^*}$ be the estimated parameter for the ith bootstrap sample. Given a data set, the expected value of the bootstrap estimator on the bootstrap distribution is calculated as the average of the parameter estimates from the B (1000) bootstrap samples

$$\widehat{\theta_b}=\frac{1}{B}\sum_{b=1}^{B}(\widehat{\theta_b^*}) \quad (2)$$

The bootstrap standard deviation is obtained as the sample standard deviation of the parameter $\widehat{\theta_b^*}$

$$\widehat{SE_b}=\sqrt{\frac{1}{B-1}\sum_{b=1}^{B}(\widehat{\theta_b^*}-\widehat{\theta_b})^2} \quad (3)$$

Another approach is to use a normal to construct a bootstrap CI, using the estimate of the SEb

$$\widehat{\theta_b}-\widehat{SE_b}*Z_{1-\alpha/2}<\theta<\widehat{\theta_b}+\widehat{SE_b}*Z_{1-\alpha/2} \quad (4)$$

$Z_{1-\alpha/2}$ is the quantile of the standard normal distribution.

The relative bias of the asymptotic estimate will be obtained by comparing the asymptotic estimate $\widehat{\theta_B}$ and the real value θ0 as a result.

$$\text{RBiais}(\theta_0)=\frac{1}{B}\sum_{B=1}^{B}\left(\frac{\widehat{\theta_b}-\theta 0}{\theta 0}*100\right) \quad (5)$$

(ii). K Fold Validation

The k-fold validation consists of dividing the original sample into k (5) samples (or "blocks"), and then selecting one of the 5 samples as the validation set while the other k-1 samples constitute the training set. After learning, a validation performance can be calculated. Then we repeat the operation by selecting another validation sample among the predefined blocks. At the end of the procedure we obtain 5 performance scores, one per block. A number of model fit metrics will be calculated such as R^2, RMSE, MAE.

The coefficient of determination (R^2) is the percentage of the total variation of the response variable explained by the regression line. Its formula is:

$$R^2=1-\frac{SSE}{SSJ} \quad (6)$$

With

$$SSE=\sum_{je=1}^{n}(y_{je}-\widehat{y_{je}})^2 \quad (7)$$

is the sum of the squared differences between the predicted and observed value,

$$SSJ=\sum_{je=1}^{n}(y_{je}-\overline{y_{je}})^2 \quad (8)$$

is the sum of the squared differences between the mean and observed.

The root mean square error (RMSE) is the average prediction error (square root of the mean square error).

$$RMSE=\sqrt{\frac{\sum_{je=1}^{n}(y_{je}-\widehat{y_{je}})^2}{n}} \quad (9)$$

The mean absolute error (MAE) is the average absolute prediction error. It is less sensitive to outliers.

$$MAE=\frac{\sum_{je=1}^{n}|y_{je}-\widehat{y_{je}}|^2}{n} \quad (10)$$

The software we used to analyze our data was R software version 4.1.1.1 and excel software version 2017. The significance level was 5%.

3. Results

3.1. Reducing the Number of Variables

3.1.1. Principal Component Analysis

The existing relationship between all the variables taken in pairs and the correlation coefficients between these different variables were given by the correlation matrix (Figure 2). The correlation matrix indicates negative correlations between precipitation on the day of sampling and the concentrations of Cl, NH_4, NO_3, temperature and conductivity; and also between wind speed on the day of sampling and the variables of minimum and maximum temperature on the day of sampling and sodium concentration. Positive correlations between turbidity and the variables colour, TAC, DHT, HCO_3, SO_4; and between colour and the presence of organic matter are also noteworthy. The matrix is not singular with a determinant equal to 0.0000006. Bartlett's sphericity test showed that the matrix is not an identity matrix as $p < 0.05$. The choice of the number of components was made according to the Kaiser criterion and allowed to keep components whose eigenvalues are higher than 1, hence the first 10 components out of 30 for the 2 explained variables.

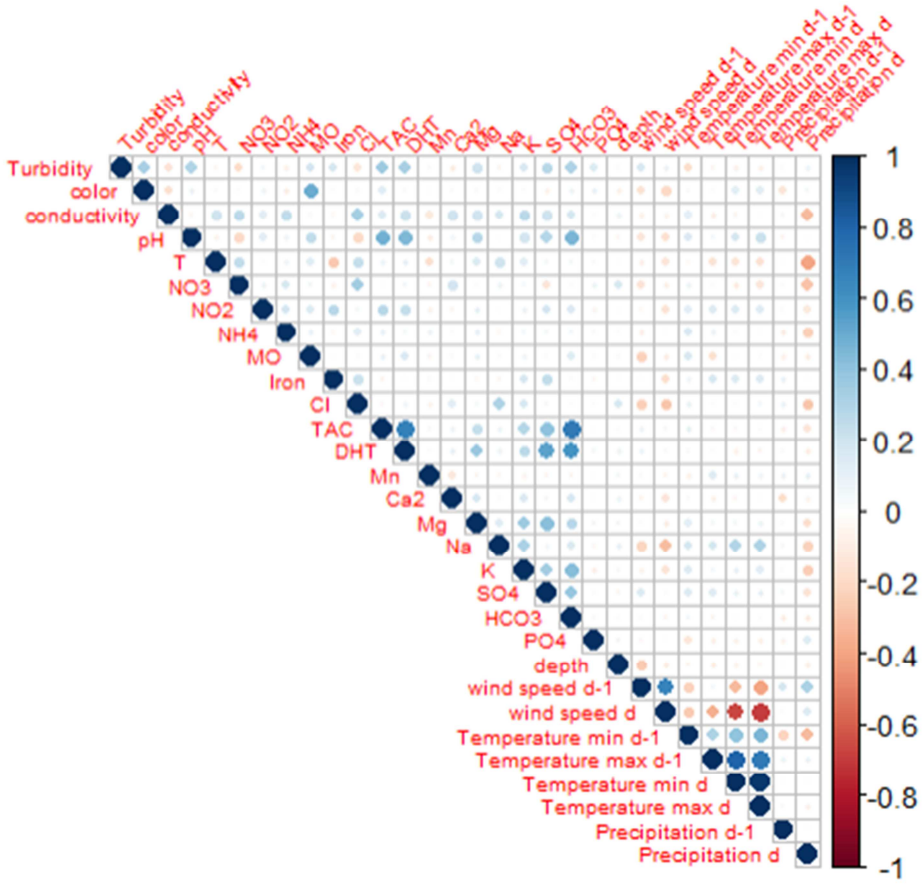

Figure 2. *Correlations between the physico-chemical parameters of well waters.*

The projection of the explanatory and explained variables onto the first components are presented on the correlation circles. *E. coli* concentration appears to be positively correlated with turbidity, wind speed on the day and the day before sampling and temperature (Figure 3).

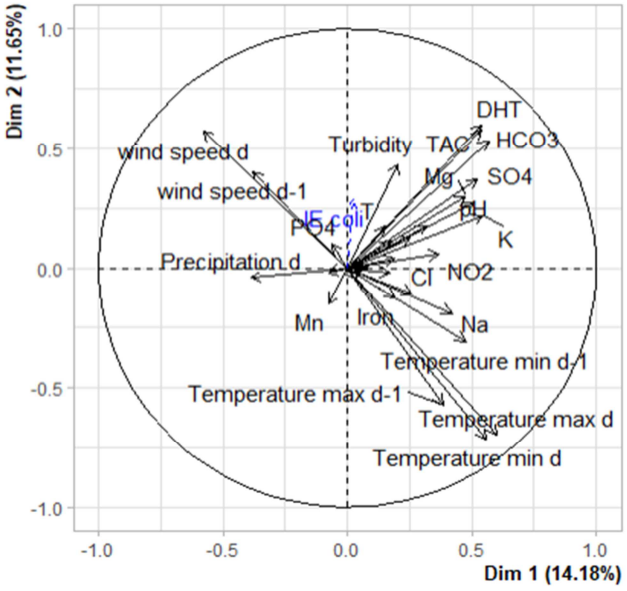

Figure 3. *Analysis in variable space with* $\log_{E.\ coli}$ *(factorial 1).*

Linear regression on the components resulted in only components 2, 3 and 9 of the 10 components being retained for log10 (*E. coli*).

3.1.2. Clustering

12 variables were able to statistically significantly distinguish clusters on all health inspection variables at α=5% risk. These are:

Latrine less than 15 m away, Unauthorized dumping upstream of the well, Protective fence, Nature of pollution, Type of nuisance, Type of disinfectant, Closed at the time of the survey, Distance of septic tank from the well, Maintenance of the well, Household activity around the well, Periodicity of maintenance of the well, Water point or source of pollution.

3.2. Selection of Variables

The selection of the variables began with the quantitative variables by retaining the variables that contributed to the formation of the axes that were significantly associated with the explained variables. The variables Turbidity, pH, wind speed of the day before and the day of sampling, minimum and maximum temperature of the day before and the day of sampling, conductivity, temperature, nitrate, chlorine, calcium, sodium, iron, manganese, phosphate, Total Alkalinity, bicarbonate, precipitation of the day before and the day of sampling were the variables that best contributed to the dimensions 2, 3 and 9, for which in bivariate analysis

there was a p < 0,2. For the categorical variables we used the variables that were significantly associated with cluster differentiation. These are the 12 variables mentioned above at the clustering level. The Akaike criterion allowed us to retain 9 variables associated with the variable $\log_{E.\ coli}$.

3.3. Linear Mixed Effects Models

A spaghetti plot shown in Figure 4 and likelihood ratio test with estimates based on the restricted maximum likelihood approach (p> 0.05) selected a random intercept linear model.

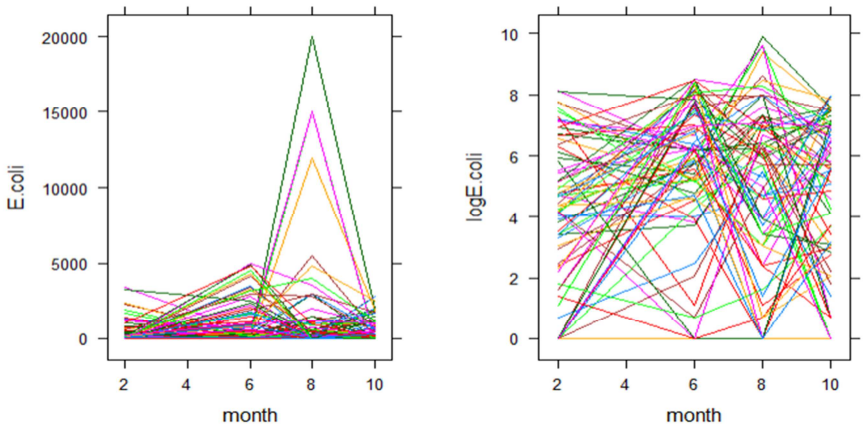

Figure 4. *Evolution of E. coli and* $\log_{E.\ coli}$ *concentrations for each well over time.*

The predictive model explained 30.24% of the variance in *E. coli* concentrations (log transformation). It is based on turbidity, pH, iron, chlorine, sodium, minimum and maximum temperature of the day of sampling, rainfall preceding the day of sampling, presence of latrines within 15 m of the well. The inter-individual variability of the *E. coli* concentration at time 0 is $\sigma\ b\ 0 = 0{,}85$. The selected model is written as follows:

$$\text{Log}\ (E.\ coli{+}1)ij = (0.65 + \alpha 0i) + 0.02{*}\ (\text{Turbidityij}){+}\ 0.90{*}\ (\text{pHij}){+}\ 0.07{*}\ (\text{Ironij})$$

$$-\ 0.04{*}\ (\text{Clij}){+}\ 0.04{*}(\text{Naij}){-}\ 0.12{*}(\text{Temperature min dij}){+}\ 0.05{*}\ (\text{Temperature max dij}) + 0.6\ {*}(\text{Precipitation d-1.ij})$$

$$+\ 0.25{*}(\text{Latrine within 15 mij}){+}\xi ij$$

$$\forall\ i \in [[1;\ 72]],\ \forall tij \in [[2;\ 6;\ 8;\ 10]] \tag{11}$$

3.4. Validation of Model Performance

The validation of the performance requires the verification of the conditions for the application of a linear mixed model, which are the normality of the random and residual errors and the homoscedasticity. The log of *E. coli* concentrations appears to evolve linearly with time (Figure 5). There is marked homoscedasticity as the variance of the concentration log appeared not to increase too much with time. Plotting the residuals against the values predicted by the model is useful for detecting heteroscedasticity. One can conclude that there is homoscedasticity. The qqplot shows points that undulate around a straight line: the assumption of normality of the residuals seems reasonable here.

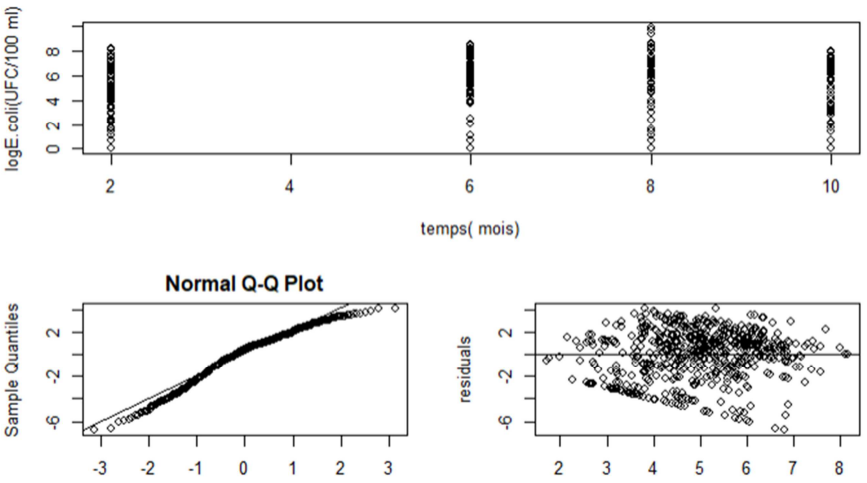

Figure 5. *Verification of residue assumptions* $\log_{E.\ coli}$ *model validity.*

The assumption of normality of the random effect on interception seems reasonable (Figure 6).

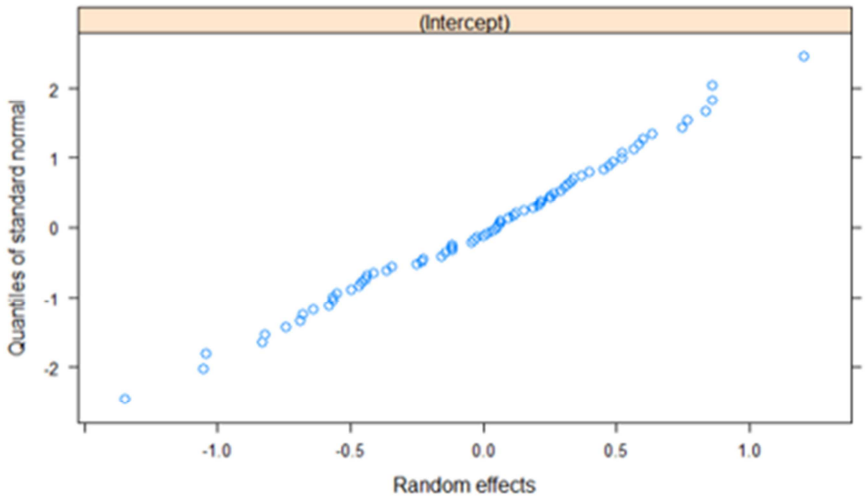

Figure 6. *Testing the validity of random effects assumptions.*

3.4.1. Validation Bootstrap

Table 1 below shows the parameters of the model on the empirical data, the replication means, the standard errors and the relative biases for each parameter of the explanatory variables.

Table 1. *Estimation of model performance by bootstrap validation.*

Parameters	Observed	Replication averages	Standard errors	Relative Bias
Intercept	0,6549	0,6870	1,4928	4,9
Turbidity	0,0208	0,0209	0,0089	0,3384
PH	0,9022	0,8940	0,2763	0,9146
Iron	0,0705	0,0700	0,0301	0,6582
Cl	-0,0460	-0,0462	0,0113	0,3998
Na	0,0436	0,0438	0,0239	0,4506
Temperature min d	-0,1193	-0,1188	0,0504	0,4247
Temperature max d	0,0480	0,0479	0,0411	0,0659
Precipitation of d-1	0,6033	0,6042	0,2493	0,1356
Latrine within 15 m	0,2517	0,2519	0,3251	0,0901

3.4.2. K Fold Validation

The mean conditional R^2 is 0,31, the mean prediction error (RMSE) is 1,95 and the mean absolute prediction error (MAE) is 1,6 (Figure 7).

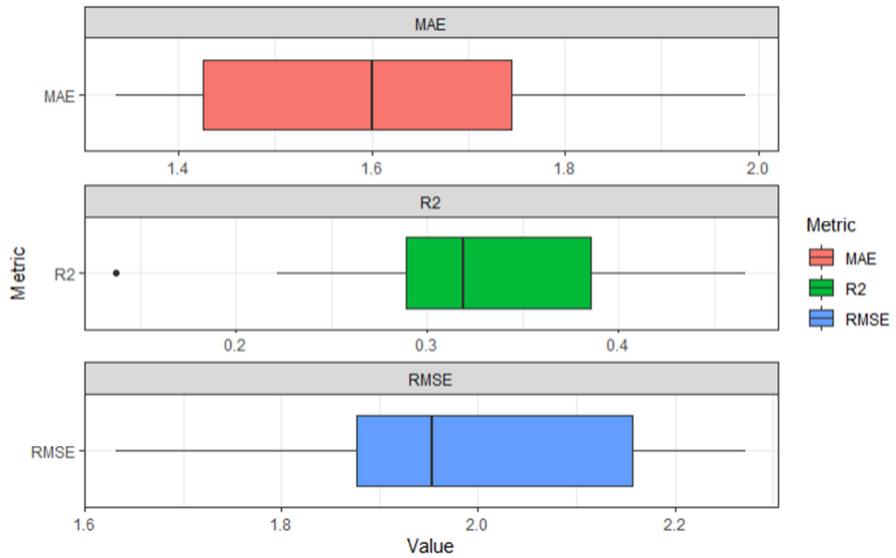

Figure 7. *Performance validation of the $\log_{E.\ coli}$ model by k-fold validation.*

4. Discussion

An epidemic of diarrhea was declared in the village of M'pody following the consumption of well water. In order to set up an early intervention system in the event of a declaration of waterborne disease in this locality, we thought of setting up bacteriological characterisation models of the water tables of the village of M'pody (Ivory Coast) based on physicochemical parameters and meteorology in order to estimate the concentration of indicator germs of faecal pollution by well.

In our study, several combinations of variables were tested to reproduce the observed *E. coli* concentrations. The final choice was 9 variables. The variance explained by taking into account only the fixed effects (marginal R^2) is 20.94%, while the variance explained by the fixed effects and the group effects (conditional R^2) is 30.25%. The variables turbidity and precipitation are positively correlated with *E. coli* concentration. These variables are more likely to be found in predictive models for fecal coliforms. This was the case in the study by Hebert et al, 2004 on the modelling of the bacteriological quality of a potential swimming site in Beauport Bay (QUEBEC), 2003 [15]. The precipitation recorded 24 hours prior to sampling has a significant impact on the microbiological quality of the water. Especially since more than 50% of the wells in this area have sanitary facilities upstream of the wells. These upstream facilities can be a source of bacterial transport during rainfall. The use of turbidity as a predictor is due to the fact that turbidity is believed to be caused by the presence of suspended solids. The study by AD N'Diaye et al showed a positive and highly significant correlation between the presence of faecal coliforms and turbidity [16]. Iron is an essential nutrient for bacteria. It acts both as a macroelement and as a cationic trace element. It plays an important role in bacterial metabolism as an electron donor and is involved in many enzymatic reactions, notably in superoxide dismutase, which is involved in oxygen production. Iron is also involved in the synthesis of nucleic acids in bacteria, viruses and parasites [17]. Temperature and pH are two important elements in bacterial growth. *E. coli* multiply at temperatures between 7 °C and 50 °C, the optimum temperature being 37 °C. Some strains grow in acidic foods, up to a pH of 4,4 [18]. The explanatory variable chlorine is negatively correlated with the concentrations of fecal coliforms and enterococci. The presence of chlorine in the well indicates non-compliance. Not using this variable in this model would bias the estimate of bacterial concentration as it is used by the population. The best way to measure the predictive power of a model is to test the model on an independent data set that is not used for parameter estimation. However, independent data sets are often unavailable, difficult to collect, and expensive [19]. One way to solve the problem is through cross-validation (CV). CV is a method of resampling data by dividing a data set into two: a training data set and a test data set. The training data set is used to fit a model, and the test data set is used to evaluate the predictive performance of the fitted model through prediction errors. This process is repeated several times and the CV estimate of the error is the average prediction error over the test data sets [20]. In our study the predictive performance of our model by k-fold validation is given as follows, the average prediction error (RMSE) varied from 1,9 to 2,15 with an average of 1,95. The absolute prediction error ranged from 1,45 to 1,75 with a mean (MAE) of 1,6 and the mean R^2 is 0,31. The basic idea when measuring these performance parameters was to see how bad or wrong the model predictions were compared to the actual observed values. Thus, a high RMSE is "bad" and a low RMSE is "good". 4 situations can occur.

1) Low RMSE, high R^2 (best case).
2) Low RMSE, low R^2.
3) High RMSE, high R^2.
4) High RMSE, low R^2 (worst case) [21].

The model is in the second situation with low RMSE and R^2. The prediction of the model is accurate. In other words, the residuals are close to zero. Cross-validation protects against over-fitting by selecting a model that captures the overall patterns of a data set and avoiding models that exploit local features of a data set. Bootstrap parameter estimates were considered unbiased when the relative bias was less than 5%, moderately biased when the relative bias was between 5% and 10%, and highly biased if it was greater than 10% [22]. In our study, the bias of each estimated parameter is less than 5% and therefore unbiased. The validation of our model shows that the residual bootstrap performs better than the case bootstrap, as the estimation of the uncertainty of the parameters of these models by the case bootstrap led to a moderate or high bias (Table 2).

Table 2. *Relative bias of REML parameter estimates for the residual and case bootstrap.*

Parameters	Relative bias	
	Residual Bootstrap	**Bootstrap cases**
Intercept	4,90	86,19
Turbidity	0,34	18,66
PH	0,91	12,35
Iron	0,66	14,21
Cl	0,40	0,18
Na	0,45	31,04
Temperature min d	0,42	11,29
Temperature max d	0,07	28,26
Precipitation d-1	0,14	3,01
Latrine within 15 m	0,09	20,66

This study contrasts with the study by Hoai Thu Thai et al. [22] when comparing bootstrap approaches for estimating parameter uncertainties in linear mixed-effects models, where case bootstraps are preferred over residual bootstraps. However, the residual bootstrap is appropriate when most real data sets tend to show non-uniformity in sampling designs, as in our study. Our study encountered limitations with low average conditional R^2s which could mean the non-accounting of some predictors, the lack of independent data for the validation of predictive performances.

5. Conclusion

The development of a statistical model for predicting the concentration was made possible by the existence of reliable databases on physical-chemical microbiological parameters, sanitary inspection and meteorology. Based on these databases, the explained variable (*E. coli*) could be explained by nine explanatory variables. The validation of the predictive performances by K fold and bootstrap showed that the predictions of our models are accurate and the bootstrap estimates of the parameters are unbiased. The implemented models could be used in case of a declaration of waterborne diseases in this locality before the results of the microbiological analysis are returned.

6. Recommendation

This model should be used by the National Institute of Public Hygiene for early decision making in the event of a declaration of waterborne disease in this locality. The implementation of this model could serve as a locomotive for the development of models in localities with high well water consumption.

Acknowledgements

We would like to thank in particular M. Laurent Lehot and Professor Roch Giorgi of the University of Aix Marseille. We also thank Professor Amin N'cho Christophe and Tchape Aubin of the National Institute of Public Hygiene.

References

[1] Otchoumou KFE, Bachir SM, Etienne AG, Issiaka S. Contribution of remote sensing and GIS in the identification of groundwater resources in the Daoukro region (Central-Eastern Ivory Coast). Int J Innov Appl Stud 2012; 1: 35–53.

[2] Festy B, Hartemann P, Ledrans M, Levallois P, Payment P, Tricard D. Water Quality. Environ Santé Publique-Fond Prat 2003: 333–68.

[3] Guergazi S, Achour S. Physico-chemical characteristics of the town of Biskra's feedwater. Practice of chlorination. Larhyss J 2005; 4: 119–27.

[4] Adelaïde OYMJ, Yaya OL, Bernard YO, Veronique M. Impacts of Environmental Management on the Quality of Traditional Well Water in the Soubré Region (South-West of Côte d'Ivoire). J Water Resour Prot. 27 déc 2017; 9 (13): 1634-44Ernest AK, Nagnin S, Gbombélé S, Théophile L, Solange OM, Pacôme ZS. Groundwater pollution in Africans biggest towns: case of the town of Abidjan (Côte d'Ivoire). Eur J Sci Res 2008; 20: 302–16.

[5] Yapo O, Mambo V, Seka A, Ohou MJA, Konan F, Gouzile V, et al. Evaluation of the quality of domestic well water in the deprived areas of four municipalities of Abidjan (Ivory Coast): Koumassi, Marcory, Port-Bouet and Treichville. Int J Biol Chem Sci 2010; 4.

[6] Lacina C, Dramane D, Adama C, Germain G. Use of water resources, sanitation and health risks in the precarious neighbourhoods of the commune of Port-Bouët (Abidjan; Ivory Coast). VertigO- Electronic Rev En Sci Environ 2004; 5.

[7] Fofana F. Assessment and mapping of the vulnerability of the Abidjan groundwater to pollution using Drastic and God methods. Memo DEA Univ Abobo-Adjamé Côte D'Ivoire 2005: 72.

[8] Ahoussi KE, Koffi YB, Kouassi AM, Soro G, Biémi J. 2013. Étude hydrochimique et microbiologique des eaux de source de l'ouest montagneux de la Côte d'Ivoire : Cas du village de Mangouin-Yrongouin (sous-préfecture de Biankouman). Journal of Applied Biosciences, 63: 4703– 471 n.d.

[9] Jourda JP, Kouamé KJ, Saley MB, Kouadio BH, Oga YS, Deh S. Contamination of the Abidjan aquifer by sewage: An assessment of extent and strategies for protection. Groundw. Pollut. Afr., CRC Press; 2006, pp. 305–14.

[10] KOACI. Ivory Coast: 69 cases of "detected" diarrhea in Anyama, no loss of life. KOACI n.d. https://www.koaci.com/article/2020/01/24/cote-divoire/sante/cote-divoire-69-cas-de-diarrhee-detectes-a-anyama-aucune-perte-en-vie-humaine_138802.html (accessed November 12, 2021).

[11] Rompré A, Servais P, Baudart J, De-roubin M-R, Laurent P. Detection and enumeration of coliforms in drinking water: current methods and emerging approaches. J Microbiol Methods. 2002; 49 (1): 31 54. n.d.

[12] DGPR (GENERAL CENSUS OF POPULATION AND HABITAT). (2014): Socio-demographic data. Permanent Technical Secretariat of the RGPH Technical Committee, p 26 n.d.

[13] Nomades DC. 2020 Anyama weather. Hist Météo n.d. https://www.historique-meteo.net/afrique/cote-d-ivoire/anyama/2020/ (accessed May 19, 2022).

[14] Palm R. Use of bootstrap for statistical issues related to parameter estimation. Biotechnol Agron Society Environ 2002; 6: 143–53.

[15] Modelling the bacteriological quality of a potential bathing site in Beauport Bay, QUEBEC, ETE, 2003 - Google Search n.d. accessed May 16, 2022.

[16] N'Diaye AD, Sid MO, Kankou AO, Namr KI. Assessment of faecal coliform contents by coupling with physic o-chemical parameters in ACP: case of the WWTP effluents of the Sebkha, Nouakchott market perimeter. Cameroon J Exp Biol 2011; 7: 35–40.

[17] Iron and infectious diseases François Bricaire, Paris - Google Search n.d. https://www.google.com/search?q=Fer+et+maladies+infectieux+Fran%C3%A7ois+Bricaire%2C+Paris&oq=Fer+et+maladies+infe

[18] Escherichia coli (*E. coli*) s.d. https://www.who.int/fr/news-room/fact-sheets/detail/e-coli (consulté le 16 mai 2022).

[19] Snee RD. Validation of regression models: methods and examples. Technometrics 1977; 19: 415-28.

[20] Yang Y, Huang S. Suitability of five cross validation methods for performance evaluation of nonlinear mixed-effects forest models – a case study. For Int J For Res 2014; 87: 654–62.

[21] Wheeler W. Evaluating linear regression models using RMSE and R2. wwblog 2021. https://medium.com/wwblog/evaluating-regression-models-using-rmse-and-r%C2%B2-42f77400efee (consulté le 17 mai 2022).

[22] Thai H-T, Mentré F, Holford NH, Veyrat Follet C, Comets E. A comparison of bootstrap approaches for estimating uncertainty of parameters in linear mixed effects models. Pharm Stat 2013; 12: 129-40.