Variables selection by the LASSO method. Application to malaria data of Tori-Bossito (Benin)

Bienvenue Kouwaye, Noël Fonton, Fabrice Rossi, Gilles Cottrell, Norbert Mahouton Hounkonnou

Variables selection by the LASSO method. Application to malaria data of Tori-Bossito (Benin)

B. T. KOUWAYE [a], N. H. FONTON [b], F. ROSSI [c], G. COTTRELL [d] AND N. M. HOUNKONNOU [e]

[a] *University of Abomey-Calavi, International Chair in Mathematical Physics and Applications (ICMPA-UNESCO Chair), 072 BP 50 Cotonou, Republic of Benin*
E-mail: kouwaye2000@yahoo.fr

[b] *University of Abomey-Calavi, International Chair in Mathematical Physics and Applications (ICMPA-UNESCO Chair), 072 BP 50 Cotonou, Republic of Benin*
E-mail : hnfonton@gmail.com

[c] *Université Paris 1 Panthéon Sorbonne, SAMM*
E-mail : Fabrice.Rossi@univ-paris1.fr

[d] *Institut de Recherche pour le développement (IRD/UMR 216) Paris (France)*
E-mail : gilles.cottrell@ird.fr

[e] *University of Abomey-Calavi, International Chair in Mathematical Physics and Applications (ICMPA-UNESCO Chair), 072 BP 50 Cotonou, Republic of Benin*
E-mail : hounkonnou@yahoo.fr

This work deals with prediction of anopheles number using environmental and climate variables. The variables selection is performed by GLMM (Generalized linear mixed model) combined with the Lasso method and simple cross validation. Selected variables are debiased while the prediction is generated by simple GLMM. Finally, the results reveal to be qualitatively better, at selection, the prediction point of view than those obtained by the reference method.

keywords: Lasso, cross validation, variables selection, prediction.

Introduction

The usual method for variables selection and model construction in life science as epidemiology and medicine are conventional methods like linear models, generalized models, mixed models and generalized mixed models. All of these models have shown over time deficiencies. The Lasso method proposed by Tibshirani [1] is a regularized estimation approach for regression models using an L_1-norm and constraining the regression coefficients. The Lasso is attractive method because it simultaneously performs variables selection and shrinkage. Tibshirani (1996) originally proposed quadratic programming to solve the optimization of the Lasso problem. In this paper, we developed the GLMM-LASSO method. This method combines GLMM (Generalized linear mixed model), the LASSO method and a simple cross validation. It running combining the full gradient ascent algorithm, the Fisher scoring, the Laplace approximation and the EM algorithm. The variables selection is performed by GLMM combined with the Lasso method and simple cross validation. Selected variables are debiased while the prediction is generated by simple GLMM. This paper is structured as follows. In section 1, we present the method of work, in the section 2, we present a few details of theorical results on the GLMMs algorithm, the gradient ascent, the Fisher scoring , the Laplace approximation. In section 3, the cross validation procedure is explained, in section 4 we present the main results, and in section 5 the Conclusion.

1 Materials and methods

In this section, we briefly recall the description of the study area, the mosquito collection and identification as well as the data and related variables. For more details, see [2].

Study area

The study was conducted in the district of Tori-Bossito (Republic of Benin), from July 2007 to July 2009. Tori-Bossito is on the coastal plain of Southern Benin, 40 kilometers north-east of Cotonou. This area has a subtropical climate and during the study the rainy season lasted from May to October. Average monthly temperatures varied between 27°C and 31°C. The original equatorial forest has been cleared and the vegetation is characterized by bushes with sparse trees, a few oil palm plantations and farms. The study area contained nine villages (Avamé centre, Gbédjougo, Houngo, Anavié, Dohinoko, Gbétaga, Tori Cada Centre, Zébè and Zoungoudo). Tori Bossito was recently classified as mesoendemic with a clinical malaria incidence of about 1.5 episodes per child per year [10]. Pyrethroid-resistant malaria vectors are present [3].

Mosquito collection and identfication

Entomological surveys based on human landing catches (HLC) were performed in the nine villages every six weeks for two years (July 2007 to July 2009). Mosquitoes were collected at four catch houses in each village over three successive nights (four indoors and four outdoors, i.e. a total of 216 nights every six weeks in the nine villages). Five catch sites had to be changed in the course of the study (2 in Gbedjougo, 1 in Avamè, 1 in Cada, 1 in Dohinoko) and a total of 19 data collections were performed in the field from July 2007 to July 2009. In total, data from 41 catch sites are available. Each collector caught of predictionall mosquitoes landing on the lower legs and feet between 10 pm and 6 am. All mosquitoes were held in bags labeled with the time of collection. The following morning, mosquitoes were identified on the basis of morphological criteria. All An. gambiae complex and An. funestus mosquitoes were stored in individual tubes with silica gel and preserved at 220°C. P. falciparum infection rates were then determined on the head and thorax of individual anopheline specimens by CSP-ELISA [11].

Environnement and behavioral data

Rainfall was recorded twice a day with a pluviometer in each village. In and around each catch site, the following information was systematically collected: (1) type of soil (dry lateritic or humid hydromorphic)assessed using a soil map of the area (map IGN Benin at 1/200 000 e , sheets NB-31-XIV and NB-31-XV, 1968) that was georeferenced and input into a GIS; (2) presence of areas where building constructions are ongoing with tools or holes representing potential breeding habitats for anopheles; (3) presence of abandoned objects (or ustensils) susceptible to be used as oviposition sites for female mosquitoes; (4) a watercourse nearby; (5) number of windows and doors; (6) type of roof (straw or metal); (7) number of inhabitants; (8) ownership of a bed-net or (9) insect repellent; and (10) normalized difference vegetation index (NDVI) which was estimated for 100 meters around the catch site with a SPOT 5 High Resolution (10 m colors) satellite image (Image Spot5, CNES, 2003, distribution SpotImage S.A) with assessment of the chlorophyll density of each pixel of the image. Due to logistical problems, rainfall measurements are only available after the second entomological survey. Consequently, we excluded the first and second surveys (performed in July and August 2007 respectively) from the statistical analyses.

Variables

The dependent variable was the number of Anopheles collected in a house over the three nights of each catch, and the explanatory variables were the environmental factors, i.e. the mean rainfall between two catches (classified according to quartile), the number of rainy days in the ten days before the catch (3 classes [01], [24], >4 days), the season during which the catch was carried out (4 classes: end of the dry season from February to April; beginning of the rainy season from May to July; end of the rainy season from August to October; beginning of the dry season from November to January), the type of

soil 100 meters around the house (dry or humid), the presence of constructions within 100 meters of the house (yes/no), the presence of abandoned tools within 100 meters of the house (yes/no), the presence of a watercourse within 500 meters of the house (yes/no), NDVI 100 meters around the house (classified according to quartile), the type of roof (straw or sheet metal), the number of windows (classified according to quartile), the ownership of bed nets (yes/no), the use of insect repellent (yes/no) and the number of inhabitants in the house (classified according to quartile).

Table 1: **Description of variables**.

	Nature	**Number of modalities**	**Modalities**
Repellent	Nominal	2	Yes/ No
Bed-net	Nominal	2	Yes/ No
Type of roof	Nominal	2	Tole/ Paille
Ustensils	Nominal	2	Yes/ No
Presence of constructions	Nominal	2	Yes/ No
Type of soil	Nominal	2	Humid/ Dry
Water course	Nominal	2	Yes/ No
Majority Class	Nominal	3	1/4/7
Season	Nominal	4	1/2/3/4
Village	Nominal	9	
House	Nominal	41	
Rainy days before mission	Numeric	Discrete	$0/2/\cdots/9$
Rainy days during mission	Numeric	Discrete	$0/1/\cdots/3$
Fragmentation Index	Numeric	Discrete	$26/\cdots/71$
Openings	Numeric	Discrete	$1/\cdots/5$
Number of inhabitants	Numeric	Discrete	$1/\cdots/8$
Mean rainfall	Numeric	Continue	$0/\cdots/82$
Vegetation	Numeric	Continue	$115.2/\cdots/$ 159.5
Total Mosquitoes	Numeric	Discrete	$0/\cdots/481$
Total Anopheles	Numeric	Discrete	$0/\cdots/87$
Anopheles infected	Numeric	Discrete	$0/\cdots/9$

2 GLMM-Lasso method

2.1 Generalized linear mixed models: GLMMs

The Poisson distribution of parameter ν for a random variable Y is defined as :

$$P(Y = k) = e^{-\nu}\ \frac{\nu^k}{k!};\ \mathrm{k} \in \mathbb{N} \tag{2.1}$$

Its variance variance $Var(Y)$ and its expectation $E(Y)$ are $Var(Y) = E(Y) = \nu$. The model under matrix shape is:

$$g[E(Y|\, u,\, \beta)] = X\beta + Zu \tag{2.2}$$

where (Y| u, β) follows a Poisson distribution of parameter E(Y| u, β); g is the link function :

$$g(x) = \ln(x)\,;\ x > 0. \tag{2.3}$$

Assume that :

$$E(Y \mid u, \beta) = \mu\,;\ g(\mu) = \eta \text{ and } u \sim N(0, \Gamma) \tag{2.4}$$

n is the number of observations,

X is the $n \times p$ dimension matrix of covariables,

β is the p $\times 1$ dimension of fixed parameter vector,

u is the $q \times 1$ dimension matrix of random effects,

Z is the $n \times$ q dimension design matrix of covariables,

θ is the vector of variance component of u, its dimension depend on the number of random intercept and slope,

Y_i is the ith observations of Y,

Γ is the $q \times q$ dimension matrix of variance-covariance of random effects, it is only function of θ.

The data used in this work are hierarchical at three levels (catch, site and village).

For the k^{th} catch in the jth site in the ith village, the explanatory model is defined as :

$$g[E(Y_{ijk}/a_i\,, b_j, \beta)] = X_{ijk}\beta + a_i + b_{ji} \tag{2.5}$$

where a_i and b_{ji} follow as gaussian distribution defined as :

$$a_i \sim N(0, {\sigma_a}^2) \text{ and } b_{ji} \sim N(0, {\sigma_b}^2) \tag{2.6}$$

a_i is the random intercept at the i^{th} village, b_{ji} is the random intercept at j^{th} site level (house of collection) of the i^{th} village. For the determination of the parameters of the model, we use the algorithm PIRLS (Penalized Iterative Reweighted Least Squares) and the Laplace approximation [5, 6].

If L($\beta, \theta|$ Y) is the Likelihood of the model, then :

$$L(\beta, \theta|Y) = f(Y|\beta, \theta) = \int_u P(Y|\beta, u).f(u|\Gamma)\, du \tag{2.7}$$

where

L(β,θ/Y) is the likelihood function of (β,θ) giving the observations Y_i;

$f(Y|\beta, \theta)$ is the marginal density of Y giving β et θ;

$P(Y|\beta, u)$ is the mass probability function of Y, giving β and u;

$f(u|\Gamma)$ is the gaussian density function of u.

The parameters (β,θ) are solutions of the equations :

$$(\hat{\beta}; \hat{\theta}) = Arg \max_{(\beta,\theta)} L(\beta; \theta)) \tag{2.8}$$

The estimator $\hat{u}$ of u is defined as:

$$\hat{u}(\beta, \theta) = Arg \max_u [P(Y|\,\beta, u).f(u|\,\Gamma)] \tag{2.9}$$

The function logarithm is continuous and strictly increased then :

$$\hat{u}(\beta, \theta) = Arg \max_u [\ln(P(Y|\,\beta, u).f(u|\,\Gamma))] \tag{2.10}$$

and

$$\hat{u}(\beta, \theta) = Arg \max_u [\ln P(Y|\,\beta, u) + \ln f(u|\,\Gamma)] \tag{2.11}$$

The estimation of $\hat{u}$ by this method is not evident because Γ is unknown. In reality, the estimation of the parameters of the model is computed by an iterative process using two algorithms at each iteration; algorithm PIRLS (Penalized Iterative Reweighted Least squares) and Laplace approximation [5, 6].

2.2 GLMM-Lasso method

The regularization in GLMMs consists to penalize the likelihood in (Eq 2.7) by adding a penalty term $\lambda \sum_{i=1}^{p} |\beta_i|$. Then the log-likelihood penalized is :

$$l_{pen}(\beta, \theta|Y) = l_{GLMM}(\beta, \theta|Y) - \lambda \sum_{i=1}^{p} |\beta_i| \tag{2.12}$$

We assume that the matrix Γ is known and the penalty problem is reduce to :

$$(\hat{\beta}; \hat{\theta}) = Arg \max_{(\beta,\theta)} \left[l_{GLMM}(\beta, \theta|Y) - \lambda \sum_{i=1}^{p} |\beta_i| \right] \tag{2.13}$$

The algorithm used in this study is the full gradient algorithm based on the algorithm of Goeman (2010). This method can be adapted to the case which some parameters are not penalized. In this case the the penalty term in (Eq 2.12) will be replaced by $\sum_{i=1}^{p} \lambda_i |\beta_i|$ where $\lambda_i = 0$ for unpenalized parameters. The algorithm glmmLasso is similar to the algorithm of Goeman (2010), it can automatically switch to Fisher scoring when the tendency is a slow convergence to the optimum typical for gradient ascent. But this algorithm need an additional step to estimate the variance-covariance component Γ of random effects.

Description of the algorithm PIRLS

This algorithm run with fixed β. At the r^{th} iteration, the relation within coefficients is defined as:

$$\eta^{(r)} = X\beta + Zu^{(r)} \tag{2.14}$$

The following quantities are evaluated

$$\mu^{(r)} = g^{-1}(\eta^{(r)}),\ \frac{d\eta}{d\mu} = G^{(r)} \text{ and } W^{(r)}, \tag{2.15}$$

$W^{(r)}$ is the matrix of weight automatically generated using the data.
Suppose $z = g(E(Y \mid \beta, \Gamma))$. The Taylor development at order 1 of z at μ gives :

$$z^{(r)} \approx \eta^{(r)} + G^{(r)})(Y - \mu^{(r)}) \tag{2.16}$$

The vector u is solution of the equation:

$$(Z^T W^{(r)} Z)u^{(r+1)} = Z^{(T)} W^{(r)} z^{(r)} \tag{2.17}$$

This equation is solved by the PLS (Penalized least square) using Γ^{-1} as penalty. The solution at each iteration is defined as :

$$(Z^T W^{(r)} Z + \Gamma^{-1})u^{(r+1)} = Z^{(T)} W^{(r)} z^{(r)} \tag{2.18}$$

Γ is symmetric and positive defined matrix. Its decomposition is :

$$\Gamma^{-1} = \Delta^{-1}\Delta \tag{2.19}$$

Suppose

$$Z^* = Z\Delta^{-1} \text{ and } u^* = \Delta u \tag{2.20}$$

then

$$(Z^{*T} W^{(r)} Z^* + I)u^{(r+1)} = Z^{*T} W^{(r)} z^{(r)} \tag{2.21}$$

and

$$u^{(r+1)} = (Z^{*T} W^{(r)} Z^* + I)^{-1}(Z^{*T} W^{(r)} z^{(r)}). \tag{2.22}$$

The values , $u^{*(1)},\ u^{*(2)}, \ldots, u^{*(n)}$, converge to $\hat{u}^*(\beta, \theta)$ if the quantity $||\eta^{(r+1)} - \eta^{(r)}||/||\eta^{(r)}||$ is under a fixed threshold. At each iteration, the matrix of variances-covariances of u is approximated by :

$$Var(u^{(r)}) \approx D^{(r)} = \hat{\Gamma}^{(r)} = (Z^{*T} W^{(r)} Z^* + I) \tag{2.23}$$

and

$$\hat{\beta} = Arg \max_{\beta} \left[d(Y|\, \beta, \hat{u}) - (\hat{u}^*)^T (\hat{u}^*) + 2 \ln(\det(\hat{\Gamma})) \right] \tag{2.24}$$

where $d(Y|\, \beta, \hat{u})$ is the deviance of the model.

2.3 Cross validation

The great challenge in the Lasso method is the choice of the regularization parameter λ. Generally, the cross validation is used to solve the problem. The method developed in this paper is similar to k-fold cross validation. It uses random effect according to the nature of the data. The set of observations is divided in k subset. Practical experience suggests that k = $\min(\sqrt{n}, 10)$ [8]. Let n be the number of observations and n_k the number of observations in the k^{th} fold. Because of random effect and the number of villages,

$$k * n_k = n, \;\; n_k \;\epsilon\; \mathbb{N} \text{ and } \; k > 9 \tag{2.25}$$

The learning is performed on $(k-1)$ fold E_A learning set and the test on the last fold E_T the test set. The relation between the number of subsets, the number of observations in one subset and n can be written by :

$$\sum_{q=1}^{k} n_q = n, \; n_k \approx \frac{n}{q} \;\text{ and }\; n_q \in \mathbf{N} \tag{2.26}$$

For each value of lambda, the cross validation algorithm runs like :

- the set of all observations is divided in k-subsets,
- a GLMM-LASSO model is compute on E_A,
- a prediction of the number of anopheles on E_T is computed,
- a quadratic error of prediction is calculated using the predictions and the observations of the test subset,
- the mean of quadratic error of prediction is computed using the quadratic error of prediction.

2.4 Parametric Risk function

The natural loss function used in optimization is the quadratic loss. We introduce a new loss function and obtain a the risk for each observation by:

$$Err_{pred_i} = \frac{|Y_{obs_i} - \hat{Y}_i|}{Y_{obs_i} + \phi}; \;\; \phi \,\epsilon\, \mathbb{R}_+^* \tag{2.27}$$

Then

$$Err_{pred_i} = f(\lambda, \phi), \;\; \lambda \,\epsilon\; \mathbb{R}, \;\; \text{et } \phi \,\epsilon\, \mathbb{R}_+^* \tag{2.28}$$

The estimate $\hat{\phi}$ of the new parameter ϕ introduced in the term Err_{pred} is evaluated through the distribution of the prediction error. The optimal value of λ is evaluated the score of cross validation is defined as : for each λ_i, the score of cross validation is defined as :

$$Score(\lambda_i) = \frac{1}{k} \sum_{r=1}^{k} \frac{1}{n_k} \sum_{q=1}^{n_k} \frac{|Y_q^{[r]} - \hat{Y}_q^{[r]}(\lambda_i)|}{Y_q^{[r]} + \hat{\phi}} \tag{2.29}$$

$Y^{[r]}$ is the vector of observations in the test set number r,
$\hat{Y}^{[r]}$ is the vector of predicted value for the test set number r using the model built on the learning set (all observations without the test set number r).
For all positive value of λ, the score exist and is finite.

$$\forall \lambda \in \mathbb{R}, \; \lambda \geqslant 0, \; Score(\lambda) < \infty \tag{2.30}$$

Equation (2.30) shows that the algorithm of the score of cross validation converge and there exist a value of λ that minimize the function $Score(.)$. If λ_0 is this value then :

$$\lambda_0 = Arg \min_{\lambda_i} [Score(\lambda_i)] \tag{2.31}$$

The cross validation algorithm possesses the optimal properties and converge. One can determine the value of λ_0.
We first use a large and fine grid of the values of the regularizing parameters to show the behavior of the predictors (Climatic and environmental) variables.

3 Results

3.1 Prediction error and optimal scaling parameter

The different values of the scaling parameter are : $\phi \in \{0.1, 0.2, 0.3, 0.4, 0.5, 0.5, 0.6, 0.7, 0.8, 0.9, 1, 2, 3\}$

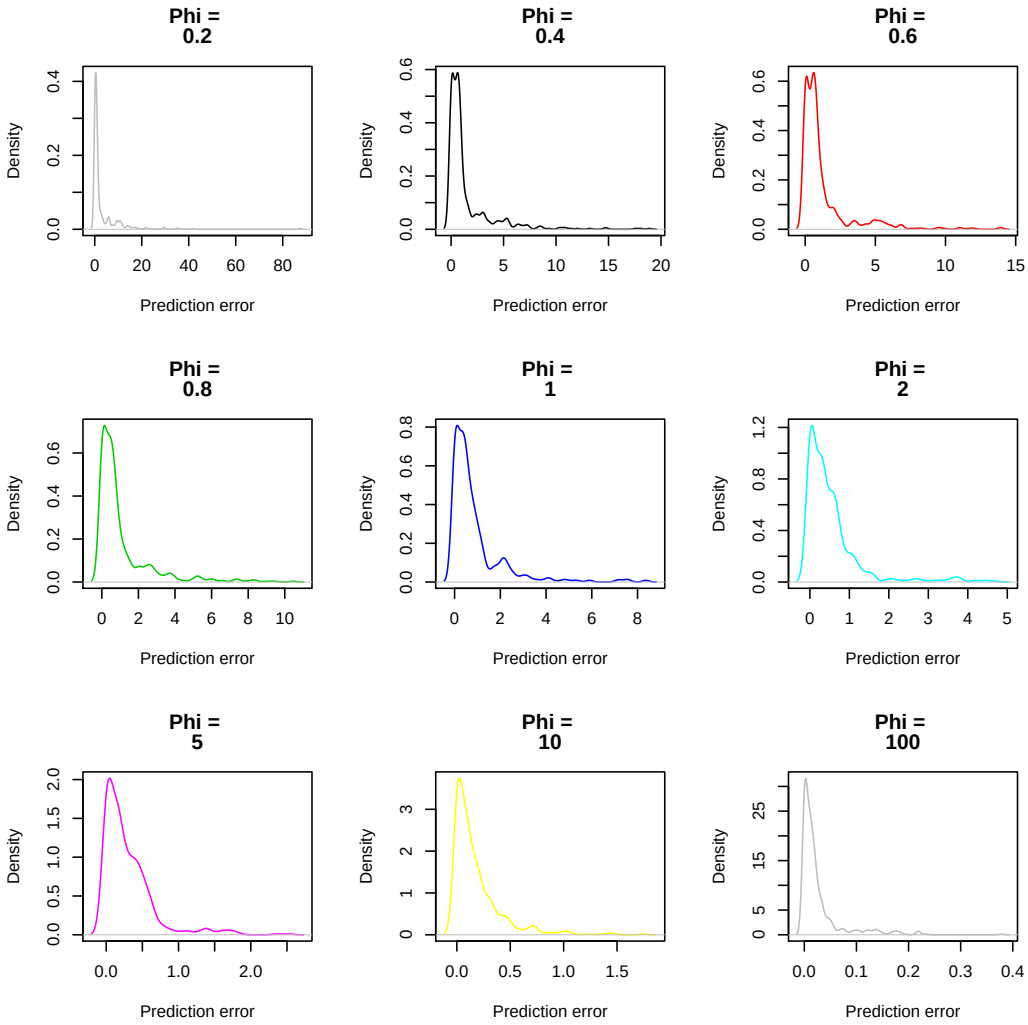

Figure 1: Distribution of prediction error according to the scaling parameter ϕ

The optimal value of ϕ is obtained when the information is distorted as little as possible. Results show that $\phi = 1$

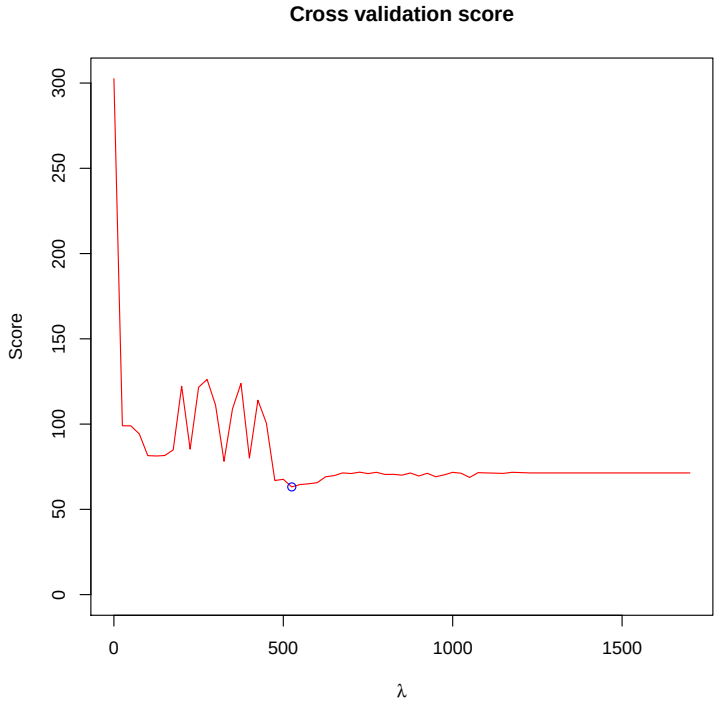

Figure 2: Score of cross validation

3.2 Cross validation and Optimal lambda

The score is not convex and its optimum converge to a finite value when λ is increasing to infinity.
The blue point on the curve is the minimum error of prediction for the GLMM-LASSO. This minimum is obtained at $\lambda = 500$.

3.3 Variables selected for the predictive model

For the optimal value of the regularization parameter $\lambda = 500$, the variables selected are :

- **Reference model** : Season, the number of rainy days during the three days of one survey, mean rainfall between 2 survey, number of rainy days in the 10 days before the survey, the use of repellent, NDVI, the interaction between season and NDVI [2].
- **LASSO method** : Season, the number of rainy days during the three days of one survey, mean rainfall between 2 survey, number of rainy days in the 10 days before the survey and NDVI.

Table 2: **Summary of the models for the prediction**

Method	**Lambda**	**Mean**	**Std**	**Quadratic risk**	**absolute risk**
Observations	-	3.746732	8.46	-	-
GLM Reference	-	3.76108	5.26	52.23	3.46
LASSO	525	3.743461	4.95	54.12	3.44

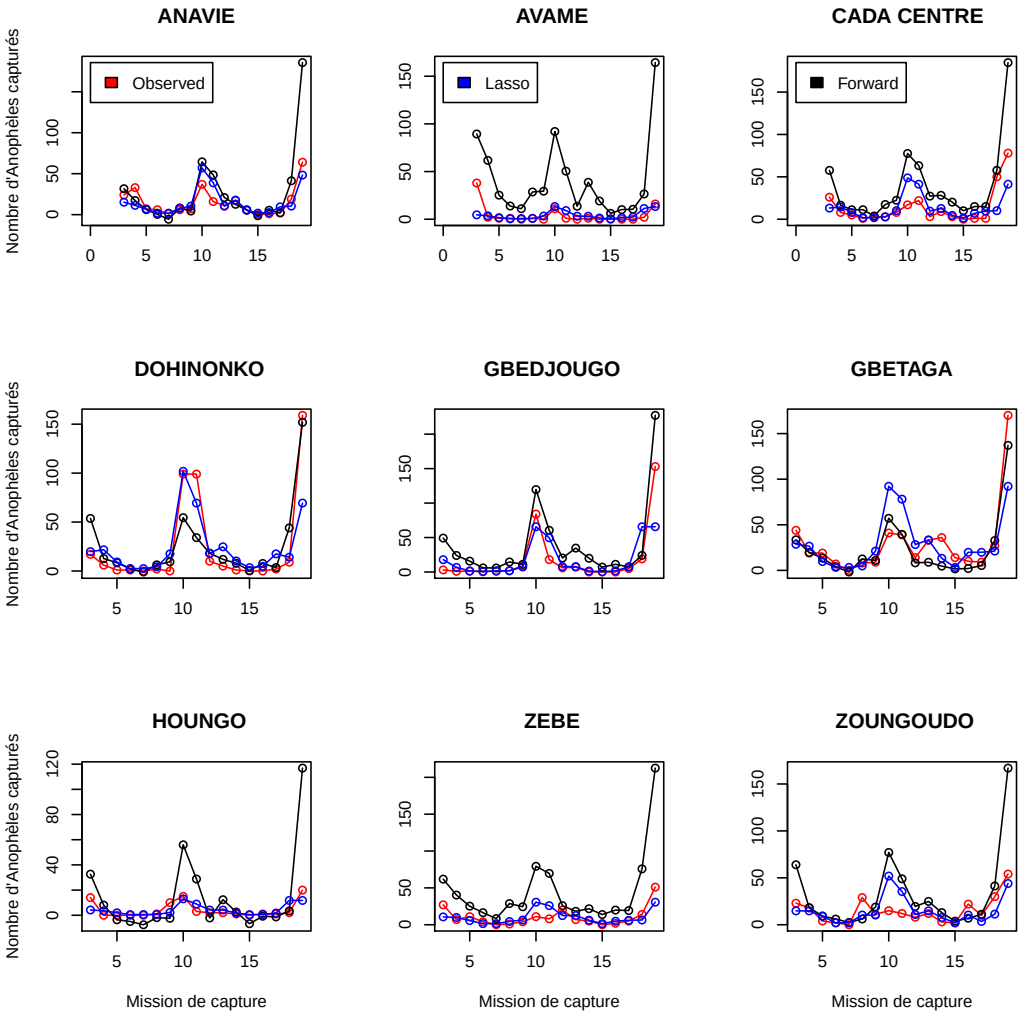

Figure 3: Comparison of the observations, the fitted values of the reference model and the Lasso method in the nine villages of study area

4 Conclusion

In this work, we implemented an algorithm for the prediction of malaria risk using environmental and climate variables. We performed the variables selection with the help of GLMM-LASSO a method combining GLMM, Lasso and a cross validation. The selected variables were debiased and the prediction was achieved by simple GLMM. The results obtained by such a procedure are clearly better improved compared to those obtained by the reference method. The improvement concerns all properties such as the quality of the selection and prediction.

Acknowledgments

We thank all the member of the laboratories: IRD/UMR216/CERPAGE (Cotonou), LERSAB (Abomey-Calavi), SAMM (Paris-France) and AUF (Agence Universitaire de la Francophonie)

References

[1] R. Tibshirani, Regression shrinkage and selection via the lasso, *Journal of the Royal Statistical Society. Series B (Methodological)* **58**, 267-288 (1996).

[2] G. Cottrell, B. Kouwayè, C. Pierrat, A. le Port, A. Bourama, N. Fonton, M. N. Hounkonnou, A. Massougbodji, V. Corbel and A. Garcia, Modeling the Influence of Local Environmental Factors on Malaria Transmission in Benin and Its Implications for Cohort Study, *PlosOne* **7**, 8 (2012).

[3] G. B. Damien, A. Djenontin, V. Corbel, C. Rogier, S. B. Bangana and al, Malaria and infection disease in an erea with pyrethroid-resitant vectors in southern Benin, *Malaria Journal* **9**, 380 (2010).

[4] S. Arlot and F. Bach, Cours d'apprentissage Statistique, Methode à noyaux, *Misc: Notes de cours prises par L. Montuelle et L. Le Goff.*

[5] D. Bates, Mixed models in R using the lme4 package Part 5 : Generalized linear mixed models *Departement of Statistics, University of Winconsin-Madison,* `Douglas.Bates@R-project.org` (2012).

[6] D. Bates, Fitting mixed-effect models using using the lme4 package in R , *nternational Meetting of the Psychometric Society,Departement of Statistics, University of Winconsin-Madison,* `Douglas.Bates@R-project.org`, *(2012)*

[7] G. Tutz and A. Groll, Variables selection for Generalized linear mixed models by $\mathbf{L}_1$-Penality estimation, *Departement of Statistics, University of Munich*, **Technical report**, (2011).

[8] J. De Bradanter, K. Pelckmans, J. A. K. Suykens, J. Vandewalle and B. De Moor, *Robust cross validation score function with application to weigthed least squares support vector machine function estimation* (katholieke Universiteit Leuven, departement of electrical engineering, ESAT-SISTA, 2003).

[9] H. B. Geyer, W. D. Heiss and F. G. Scholtz, Regularization and variable selection via the elastic net, *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **67 (2)** 301–320 (2005)

[10] R. Wirtz, AZ. avala F., Y. Charoenvit and *al*, Comparative testing of monoclonal antibodies against Plasmodium falciparum sporozoites for ELISA development, *Malaria journal*, **9 (1)** 380, BioMed Central Ltd (2010)

[11] D. Georgia B, A. Djenontin, V. Corbel and *al*, Malaria infection and disease in an area with pyrethroid-resistant vectors in southern Benin, *Bull World Health Organ*, **65** 39-45, (1987)