Gametocytes infectiousness to mosquitoes: variable selection using random forests, and zero inflated models

Robin Genuer, Isabelle Morlais, Wilson Toussile

INRIA

INSTITUT NATIONAL DE RECHERCHE EN INFORMATIQUE ET EN AUTOMATIQUE

Gametocytes infectiousness to mosquitoes: variable selection using random forests, and zero inflated models

Robin Genuer — Isabelle Morlais — Wilson Toussile

N° 7497

Décembre 2010

Thème COG

ISSN 0249-6399 ISRN INRIA/RR--7497--FR+ENG

INSTITUT NATIONAL
DE RECHERCHE
EN INFORMATIQUE
ET EN AUTOMATIQUE

INRIA

centre de recherche
SACLAY - ÎLE-DE-FRANCE

Gametocytes infectiousness to mosquitoes: variable selection using random forests, and zero inflated models

Robin Genuer * † , Isabelle Morlais ‡ , Wilson Toussile *

Thème COG — Systèmes cognitifs
Équipes-Projets SELECT

Rapport de recherche n° 7497 — Décembre 2010 — 23 pages

Abstract:

Malaria control strategies aiming at reducing disease transmission intensity may impact both oocyst intensity and infection prevalence in the mosquito vector. Thus far, mathematical models failed to identify a clear relationship between *Plasmodium* gametocytes and their infectiousness to mosquitoes. Natural isolates of gametocytes are genetically diverse and biologically complex. Infectiousness to mosquitoes relies on multiple parameters such as density, sex-ratio, maturity, parasite genotypes and host immune factors. In this article, we investigated how density and genetic diversity of gametocytes impact on the success of transmission through the mosquito vector. We analyzed data for which the number of variables plus attendant interactions is at least of order of the sample size, precluding usage of classical models such as general linear models. We then applied a variable selection procedure based on the random forests score of variable importance. The selected variables were assessed in the zero inflated negative binomial model which accommodates both over-dispersion and the sources of non infected mosquitoes. We found that the most important variables related to infection prevalence and parasite intensity are gametocyte density and multiplicity of infection.

Key-words: PLASMODIUM, MOSQUITOES, VARIABLE SELECTION, RANDOM FORESTS, ZERO INFLATED MODELS.

* Université Paris-Sud, Laboratoire de Mathématique, UMR 8628, Orsay cedex F-91405
† Inria Saclay Ile-de-France
‡ UR016, Institut de Recherche Pour le Développement, 911 Avenue Agropolis, PO Box 64501, F-34394 Montpellier Cedex 5

Centre de recherche INRIA Saclay – Île-de-France
Parc Orsay Université
4, rue Jacques Monod, 91893 ORSAY Cedex
Téléphone : +33 1 72 92 59 00

Capacité d'infection des gamétocytes aux moustiques : sélection de variables basée sur les forêts aléatoires, et modèles modifiés en zéro

Résumé :

De nouvelles stratégies de réduction de la transmission du paludisme nécessite la compréhension des facteurs pouvant influencer l'intensité d'oocystes et la prévalence d'infection chez le moustique vecteur. Jusqu'à maintenant, les modèles mathématiques ne sont pas parvenus à identifier une relation claire entre les gamétocytes de *Plasmodium* et leur capacité à infecter les moustiques. La capacité vectorielle du moustique peut dépendre de multiple facteurs tels que la densité, le sexe-ratio et la maturité du parasite, ainsi que des facteurs immunitaires du moustique. Dans ce papier, nous évaluons l'influence de la densité et de la diversité génétique du parasite sur le succès de sa transmission à travers le moustique vecteur. Nous disposons de données décrites par diverses variables dont le nombre est de l'ordre de la taille de l'échantillon, ce qui constitue un obstacle à l'usage de modèles classiques de régression tel que le modèle linéaire généralisé. Nous considérons alors l'importance des variables des forêts aléatoires pour sélectionner les variables les plus influentes. Les variables sélectionnées sont ensuite évaluées par le modèle binomial négatif modifié en zéro, qui permet de tenir compte à la fois de la sur-dispersion et des sources possibles des moustiques non-infectés. Nous trouvons que les variables les plus importantes reliées à la prévalence d'infection et l'intensité parasitaire sont la densité de gamétocytes et la multiplicité de l'infection.

Mots-clés : Plasmodium, moustiques, sélection de variables, forêts aléatoires, modèles modifiés en zéro.

1 Introduction

Malaria still represents a major health problem in more than one hundred tropical countries. The disease is caused by the parasite *Plasmodium* and its transmission occurs through the bite of an infective *Anopheles* female mosquito. In the last decades, insecticide and drug resistance has seriously hampered its control and alternative measures are urgently needed. Because *Plasmodium* transmission relies on the success of its development within the mosquito vector, called the sporogonic development, new strategies to fight malaria aim at controlling *Plasmodium* during the mosquito life cycle. Within the mosquito vector, malaria parasites undergo several life-stages and their successful development from one transition stage to an other will determine the outcome of infection. When ingested with the blood meal, male and female gametocytes fuse to form a zygote that differentiates into a mobile ookinete. The ookinete then traverses the midgut epithelium and encysts as an oocyst along the basal lamina. The oocyst, after several days of maturation, will release large number of sporozoites into the hemocoel. Sporozoites that will reach salivary glands will then be transmitted to a new host at a subsequent blood meal. *Plasmodium* parasites encounter severe losses during these successive phases and factors controlling parasite densities are not yet completely understood. Blood digestion processes and mosquito immune responses account for parasite decrease, but also the complex interplay between vector and parasite genotypes (Vaughan, 2007; Jaramillo-Gutierrez et al., 2009).

Transmission of *Plasmodium falciparum* sexual stages, the gametocytes, to the mosquito mainly depends on their maturity and density in the human host at the time of the mosquito bite. Even if it has been demonstrated that high gametocyte densities do not guarantee high mosquito infection, a greater infection of mosquitoes is generally observed with higher gametocyte densities (Hogh et al., 1998; Drakeley et al., 1999; Targett et al., 2001; Boudin et al., 2004; Paul et al., 2007; Nwakanma et al., 2008). Gametocyte densities vary greatly between human hosts, due to host acquired immunity, genetic factors of the parasite strain and other environmental parameters (blood quality, fever, anemia, anti-malarial drug uptake). In malaria endemic areas, human hosts are typically infected with multiple genotypes of parasites (Day et al., 1992; Babiker et al., 1999; Anderson et al., 2000; Nwakanma et al., 2008) and within-host competition of parasite genotypes is likely to drive transmission success. Indeed, from experiments using *Plasmodium* animal models, it has been shown that different genotypes of parasites in mixed infections have distinct ability to transmit, the more virulent strain having a competitive advantage (de Roode et al., 2005; Bell et al., 2006; Wargo et al., 2007). If different models have been proposed to correlate the gametocyte density to the transmission success of wild isolates of *Plasmodium falciparum* (Pichon et al., 2000; Boudin et al., 2005; Paul et al., 2007), to date no study related the outcome of infection to parasite complexity within the gametocyte population. Understanding relationships between co-infecting genotypes and how they influence the disease transmission is however of great importance as these might help to predict the spread of resistant strains of parasites and guide strategies for malaria control.

In this paper, we investigate how density and genetic diversity of gametocytes impact on infectiousness to mosquitoes. We analyze mosquito infection

data consisted of oocyst counts with corresponding gametocyte data: densities and genotypes at 7 microsatellite loci. Data were obtained from experiments of membrane feeding of a local colony of *Anopheles gambiae* mosquitoes on blood from volunteers naturally infected by *Plasmodium falciparum* isolates from Cameroon. Gametocyte genotypes are occurrences of several unordered categorical variables, each having numerous levels. Therefore the number of variables plus attendant interactions is at least of order of the sample size. We considered as response variables: the intensity of infection as measured by the mean of oocyst counts in infected mosquitoes, and the infection prevalence defined by the proportion of mosquitoes that became infected. The high number of variables in our data set will obviously lead to over-fitting of many familiar regression techniques such as general linear model (GLM). In addition, we deal with unordered categorical variables with several levels and potentially accompanying interactions. Therefore, following Segal et al. (2001), we use regression trees techniques.

We address the problem of selecting the most influent variables related to the response variable by applying a variable selection procedure, which comes from Genuer et al. (2010), and is based on variable importance from random forests (Breiman, 2001). The resulting method is completely non-parametric and thus can be used on data with a large number of variables of various types. Moreover, it solves the two following constraints about variable selection: 1) to find all variables highly related to the response variable; and 2) to find a small number of variables sufficient for a good prediction of the response variable. The selected variables are then assessed in a modeling for oocyst count which takes into account the complexity of the experiment we deal with. The key point of our modeling is the introduction of a new unobserved variable that enables to distinguish two possible sources of non infected mosquitoes. Indeed, the heterogeneity in the quantity and quality of gametocytes in blood-meal (Vaughan, 2007), and natural variation in mosquito susceptibility (Riehle et al., 2006) are well known phenomena. We then suggest here that mosquitoes with no oocyst can be non infected either because they did not ingest enough gametocytes with the blood-meal, or because they were refractory to the ingested parasites. We fitted a Zero-Inflated (ZI) model, which is a two components mixture model combining a point mass at zero with a proper count model. Since we deal with count data, the typical candidate models were Zero-Inflated Poisson (ZIP) and Zero-Inflated Negative Binomial (ZINB); ZINB having a slight advantage because it captures over-dispersion which is likely to appear in such data.

As a result, we found that the gametocyte density and the multiplicity of infection were the most influent variables for both infection prevalence and parasite intensity. High gametocyte density and low multiplicity of infection resulted in high parasite intensity, whereas high infection prevalence came from high gametocyte density and high multiplicity of infection.

The rest of the paper is organized as follows. Section 2 presents the data to be analyzed in Subsection 2.1, the principle of variable selection based on variable importance from random forests in Subsection 2.2, and the modeling of oocyst count in Subsection 2.3. Section 3 is devoted to the application of these methods on our data. Finally a discussion is given in Section 4.

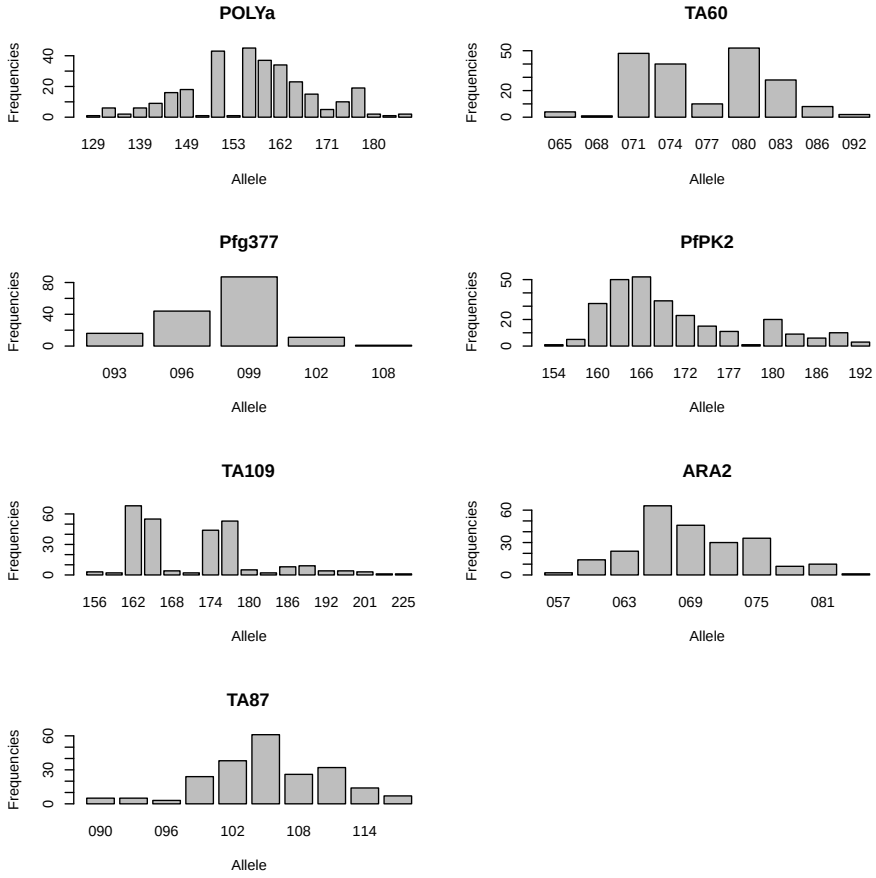

Figure 1: Alleles detected for the 7 microsatellite loci and their frequencies in *Plasmodium falciparum* gametocyte carriers.

2 Material and methods

2.1 Data collection and description

The data we considered consist of parasite densities and genotypes at 7 microsatellite loci for gametocyte isolates of *Plasmodium falciparum* on one hand, and oocyst counts 7 days post feeding for each engorged females on the other hand. *Plasmodium falciparum* gametocyte carriers were identified among asymptomatic children aged from 5 to 11 in primary schools of the locality of Mfou, a small town located 30 km apart from Yaounde, the Cameroon capital city. Volunteers were enrolled upon signature of an informed consent form by their parents or legal guardian. The protocol was approved by the National Ethics Committee of Cameroon. Gametocyte densities were expressed as the number of parasites seen against 1 000 leukocytes in a fresh thick blood smear, assuming a standard concentration of 8 000 leukocytes per μl (see Table 1 for summary of log-transformed gametocyte densities). Venous blood (2 to 3 mL)

was taken from consenting gametocyte carriers, centrifuged and the serum replaced by a non-immune AB serum. This procedure avoids the introduction of human transmission blocking factors in the experiment. 3 to 5 old females of a laboratory strain of *Anopheles gambiae* mosquito were used for the membrane feeding assays placed in cups of approximately 60-80 mosquitoes. Females were allowed to feed for 20 minutes through a Parafilm membrane on glass feeders maintained at $37°C$ and fully engorged females were kept in insectar until dissections 7 days post-infection. Midguts were removed, stained in a 0.4% Mercurochrome solution and the number of developed oocysts counted by light microscopy ($X20\ lens$). A total of 7 364 mosquitoes (see Table 1) were dissected, giving a mean of 39 females per experiment.

Gametocytes were separated from 1 mL of serum free blood using MACS® columns as previously described (Ribaut et al., 2008). DNA extractions from purified gametocytes were performed with DNAzol® and 20 ng of gametocyte DNA were subjected to whole-genome amplification (WGA) using the GenomiPhi V2 DNA Amplification Kit to generate sufficient amounts of DNA for microsatellite genotyping. Genetic polymorphism was assessed at 7 microsatellite loci as previously described (Annan et al., 2007). Their chromosome location and GenBank accession number are as follows: POLYa (chr. 4, G37809), TA60 (chr. 13, G38876), ARA2 (chr. 11, G37848), Pfg377 (chr. 12, G37851), PfPK2 (chr. 12, G37852), TA87 (chr. 6, G38838), and TA109 (chr.6, G38842). Alleles were analyzed using `GeneMapper`® software. Multiple alleles were scored when minor peaks were at least 20% of the height of the predominant allele. The number of observed alleles per locus is 21, 9, 10, 5, 15, 10 and 17 respectively (see Figure 1).

Table 1: Summary of the numbers of mosquitoes per isolate (N) and log-transformed of gametocyte densities (**log_gameto**).

	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
N	11.000	29.000	38.000	39.380	47.000	79.000
log_gameto	1.816	3.156	3.832	3.973	4.612	7.742

Feedings for which the number of dissected mosquitoes was below 20 were not considered. Then 110 experiments were included in the analysis.

2.2 Variable selection procedure

The selection procedure we considered is based on variable importances (VI) from random forests (RF). The principle of RF is to aggregate regression or classification trees built on several bootstrap samples drawn from the learning set (more details are given in Appendix A). It is shown to exhibit very good performance for lots of diverse applied situations (Breiman, 2001). Moreover, it computes a variable importance index, defined in Appendix A. Roughly, this index is a measure of the degradation of forest predictions when values of a variable are permuted.

RF variable importance is the key point of the selection procedure (see Genuer et al. (2010) for more backgrounds on RF variable importance). This procedure presents two main benefits. First the method is completely non-

parametric and can be applied on data with lots of variables of various types. Second, it achieves two main variable selection objectives: (1) to magnify all the variables related to the response variable, even with high redundancy, for interpretation purpose; (2) to find a parsimonious set of variables sufficient for prediction of the outcome variable.

Let us now describe the procedure, which comes from Genuer et al. (2010), with the following algorithm. The R package `randomForest` (Liaw and Wiener, 2002; R Development, 2009) was used in all computations.

To both illustrate and give details about this procedure, we apply it on a simulated dataset with $n = 200$ observations described by 25 continuous variables and 25 binary variables. We assume standard normal distribution $\mathcal{N}(0,\ 1)$ for all continuous variables and binomial distribution $\mathcal{B}(0.5)$ for all binary variables. We consider the following linear model

$$Y = \sum_{j=1}^{25} \beta_{c_j} X_{c_j} + \sum_{j=1}^{25} \beta_{b_j} X_{b_j}$$

in which only 8 over a total of $p = 50$ variables are related to the outcome, the others being just noise. The set of significant variables is composed by the first 4 continuous variables $\left(X_{c_j}\right)_{1\leq j\leq 4}$ and the first 4 binary ones $\left(X_{b_j}\right)_{1\leq j\leq 4}$. Their associated coefficients are given by

$$(\beta_{c_j})_{1\leq j\leq 25} = (\beta_{b_j})_{1\leq j\leq 25} = (4, 4, 2, 2, 0, \ldots, 0).$$

We also assume a 0.9 correlation between X_{c_1} and X_{c_2}, X_{c_3} and X_{c_4}, X_{b_1} and X_{b_2}, and X_{b_3} and X_{b_4}.

The selection process uses a certain number $nfor$ of random forests. In addition of this number, the user has also to provide the number $ntree$ of trees in each random forest, and the number $mtry$ of variables among which to select the best split at each node. The default parameters in the **R** package `randomForest` we used are $mtry = p/3$, $ntree = 500$. In our example, we choose the following parameters: $mtry = p/3$, and we choose $nfor = 50$ and $ntree = 1000$ to increase the VI stability. The results are summarized in Figure 2.

Let us detail the main stages of the procedure together with, in italics, the results obtained on simulated data. In the following, out of bag (OOB) error refers to an estimation of the prediction error (which is defined in Appendix A and is close to a cross-validation estimate).

- **Elimination step**

  First the variables are sorted in descending order according to VI (averaged from the $nfor$ runs).

  The result is drawn on the top left graph. The 8 *variables of interest arrive in the first* 8 *positions of the ranking.*

  Keeping this order in mind, the corresponding standard deviations of VI are plotted. A threshold for importance is computed using this graph. More precisely, the threshold is set as the minimum prediction value given by a Classification And Regression Tree (CART) model fitting this curve (for details about CART, see Breiman et al. (1984)). Then only variables with an averaged VI exceeding this level are kept. This rule is, in general,

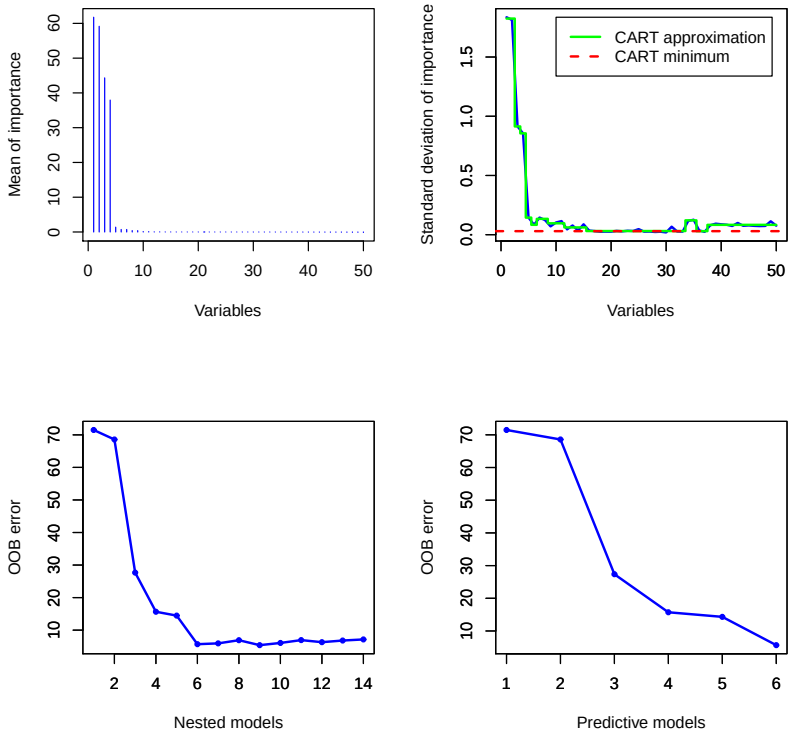

Figure 2: Variable selection procedures for interpretation and prediction for simulated data

conservative and leads to retain more variables than necessary, in order to make a careful choice later.

The standard deviations of VI can be found in the top right graph. We can see that true variables standard deviation is large compared to the noisy variables one, which is very close to zero. The threshold leads to retain $p_{elim} = 14$ *variables. Note that the threshold value is based on VI standard deviations (top right panel of Figure 2) while the effective thresholding is performed on VI mean (top left panel of Figure 2).*

- **Interpretation step**

  Then, OOB error rates (averaged on $nfor$ runs and using default parameters) of the nested random forests models are computed; starting from the one with only the most important variable, and ending with the one involving all important variables kept previously. The set of variables leading to the smallest OOB error is selected.

  Note that in the bottom left graph the error decreases and reaches its minimum when the first $p_{interp} = 9$ *variables are included in the model. This set of selected variables for interpretation contains the* 8 *true variables*

plus one noisy one. Note that the associated error is closed to the one of the model with the 6 *first variables (see bottom left panel of Figure 2) suggesting that a smaller model should be preferred for prediction purposes.*

- **Prediction step**

  Finally a sequential variable introduction with testing is performed: a variable is added only if the error gain exceeds a data-driven threshold. The rationale is that the error decrease must be significantly greater than the average variation obtained by adding noisy variables.

  The bottom right graph shows the result of this step, the final model for prediction purpose involves 6 *out of the* 8 *true variables. It is of interest that each of the two true variables non-selected is correlated to one selected variable. The threshold is set to twice the mean of the absolute values of the first order differentiated OOB errors between the model with* $p_{interp} =$ 9 *variables (the model we selected for interpretation, see the bottom left graph) and the one with all the* $p_{elim} = 14$ *variables :*

$$ave_{jump} = \frac{1}{p_{elim} - p_{interp}} \sum_{j=p_{interp}}^{p_{elim}-1} |\, errOOB(j+1) - errOOB(j) \,|$$

  where $errOOB(j)$ *is the OOB error of the RF built using the* j *most important variables.*

Since the number of variables after the variable elimination step is small (14), we tried some variants more computationally expensive, in order to validate the two last steps of the algorithm. Instead of the interpretation step, we launch a forward procedure. The principle is, at each time, to seek the best variable (in terms of OOB error rate, averaged on $nfor$ runs and using default parameters) to add in the current variable set. The set of variables leading to the smallest OOB error is then selected.

For our example, it leads, as the interpretation step, to retain the 8 *true variables plus one noisy variable (this last noisy variable being different from the one selected by interpretation step). We remark however that the initial ranking according to VI is quite changed with this procedure.*

To validate the prediction step, we tried an exhaustive procedure, i.e. we compute the OOB error rate (averaged on $nfor$ runs and using default parameters) for all models formed with the variables selected by the forward procedure. The set of variables leading to the smallest OOB error is then selected.

This procedure selects all 9 *variables selected previously.*

This validates the interpretation and the prediction step of our algorithm, since the variables sets in these variants are close to ours. In addition the errors reached by the two procedures are comparable. However this comparison was done on the easy simulated dataset we considered in this section.

2.3 Modeling oocyst count with Zero-Inflated models

The key point of our modeling is to consider that there are two possible sources of non-infected mosquitoes. First, some mosquitoes may not ingest enough parasites with sufficient sex-ratio to ensure fertilization. The reason is seemingly the

high heterogeneity in the number of gametocytes in blood-meals (Pichon et al., 2000). Second, some other mosquitoes may not be genetically susceptible to the parasites ingested (Riehle et al., 2006). We introduce a new variable U materializing this situation of non-infected mosquitoes: for mosquito j fed with blood coming from gametocytes carrier i,

$$U_{i,j} = \begin{cases} 1 & \text{if enough parasites are present in its blood-meal} \\ 0 & \text{otherwise.} \end{cases}$$

$U_{i,j}$ is an unobserved variable in our experiment. We assume that for a given i, $U_{i,1}, \ldots, U_{i,n_i}$ are independent and identically distributed. Here n_i is the number of mosquitoes associated to gametocytes carrier i. For any gametocytes carrier i, denote by

$$\pi_i := P\left(U_{i,j} = 0\right)$$

the probability that mosquito j does not ingest enough gametocytes in its blood-meal. Let $Y_{i,j}$ be the number of oocysts developed in mosquito j associated to gametocytes carrier i. The probability distribution of $Y_{i,j}$ is given by

$$P\left(Y_{i,j} = y_{i,j}\right) = \pi_i \mathbb{1}_{(y_{i,j}=0)} + \left(1 - \pi_i\right) P\left(Y_{i,j} = y_{i,j} | \; U_{i,j} = 1\right), \qquad (1)$$

where $P\left(Y_{i,j} = y_{i,j} | \; U_{i,j} = 1\right)$ is a suitable count probability distribution.

Consequently, for any gametocytes carrier i, the zero class is a mixture of two components with π_i and $1 - \pi_i$ as the mixture proportions. The resulting model of probability distribution is known as a zero-inflated count model. Such a model is a two components mixture model combining a point mass at zero with a count distribution such as Poisson, geometric or negative binomial (see Zeileis and Jackman (2008) and references therein). Thus there are two sources of zeros: zeros may come from point mass or from count component. In our framework, the zeros coming from the point mass are assumed to represent mosquitoes which did not ingest enough gametocytes to produce an infection.

Let $\lambda_i := \mathrm{E}\left(Y_{i,j} | \; U_{i,j} = 1\right)$ be the conditional mean of the count component. In the regression setting, both the mean λ_i and the excess zero proportion π_i are related to covariates vectors $\mathbf{x}_i = (x_{i,1}, \ldots, x_{i,p})$ and $\mathbf{z}_i = (z_{i,1}, \ldots, z_{i,q})$, respectively. The components of these covariates are typically the observations of the previously selected variables. They contain gametocyte density and / or their genetic profile. We consider canonical link functions **log** and **logit** for the mean of count component and the point mass component respectively. The corresponding regression equations are

$$\begin{cases} \lambda_i &= \exp\left(\beta_0 + \beta_1 x_{i,1} + \ldots + \beta_p x_{i,p}\right) \\ \pi_i &= \dfrac{\exp\left(\gamma_0 + \gamma_1 z_{i,1} + \ldots + \gamma_p z_{i,q}\right)}{1 + \exp\left(\gamma_0 + \gamma_1 z_{i,1} + \ldots + \gamma_p z_{i,q}\right)}, \end{cases}$$

where $\beta := (\beta_0, \ldots, \beta_p)$ and $\gamma := (\gamma_0, \ldots, \gamma_q)$ are the parameters to be estimated. Note that different sets of regressors can be specified for the zero inflated component and count component. In the simplest case, only an intercept is used for modeling the unobserved state (zero vs. count).

Typical candidate of zero-inflated models for count data are zero inflated Poisson (ZIP) and zero-inflated negative binomial (ZINB) (see Xiang et al. (2007) and references therein). ZINB and ZIP specifications are given in Appendix B. For the estimation of the parameters of these models, we used the package named `pscl` (Zeileis and Jackman, 2008) in **R** statistical software (R Development, 2009).

3 Application on the real data

3.1 Variable selection

Here, the results are given following the main stages of the selection procedure given in Subsection 2.2. The details are given once, in the case where the response variable is the infection prevalence of mosquitoes measured by proportion of infected mosquitoes. We will just give the selected variables at each stage in the other case where the response variable is the mean number of oocysts per infected mosquitoes. In these results, the binary variables associated to the observed alleles are coded as `locus_allele`. For example, $Pfg377\_093$ is allele 093 at locus $Pfg377$. In addition to the log-transformed of gametocytes density (log_gameto), we also consider the multiplicity of infection (MOI) defined as the maximum number of the observed alleles across the considered microsatellite loci.

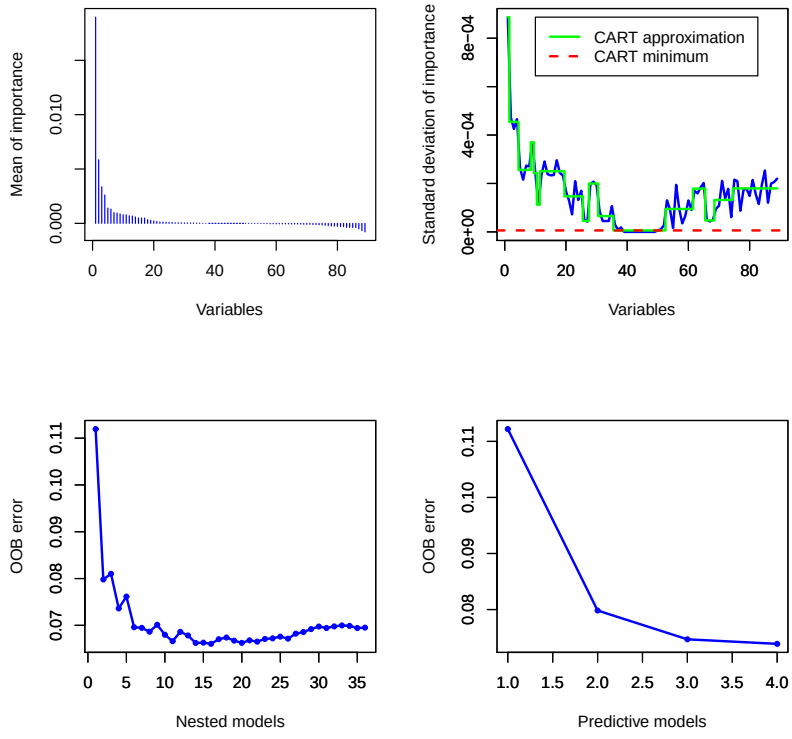

Figure 3: Variable selection for interpretation and prediction. The response variable is the infection prevalence measured by the proportion of infected mosquitoes.

Here are the main stages of the procedure.

- **Elimination Step**

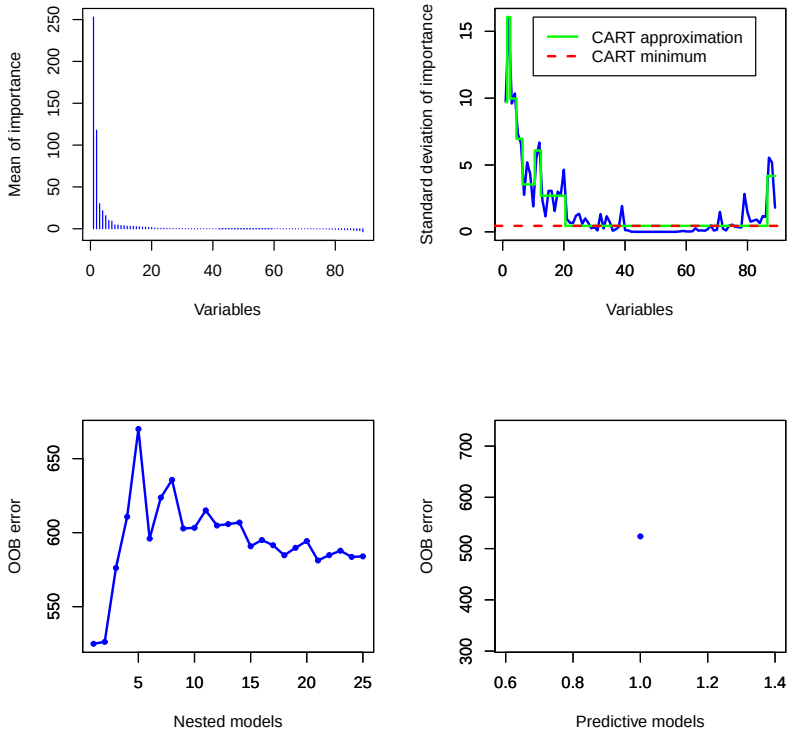

Figure 4: Variable selection for interpretation and prediction. The response variable is the mean number in infected mosquitoes.

- The top left panel in Figure 3 gives the VI mean of all the 88 variables sorted in decreasing order.
- The top right panel of Figure 3 plots the standard deviations of VI and the fitted CART model. The threshold $\min_{CART}$ represented by the horizontal dashed line leads to retain $p_{elim} = 36$ variables over 88.

- **Interpretation Step**
  This step is illustrated in the bottom left panel of Figure 3 in which the minimum OOB error rate is reached with $p_{interp} = 11$ variables for interpretation:

$$S_{interp} = \{log\_gameto,\ Pfg377\_093,\ PfPK2\_180,\ MOI,\ Pfg377\_102,\ PfPK2\_183,\ Pfg377\_099,\ TA60\_071,\ PfPK2\_169,\ PfPK2\_166,\ POLYa\_135\}.$$

- **Prediction Step**
  The bottom right panel in Figure 3 shows the behavior of the OOB error

of the nested models corresponding to the selected variables for prediction:

$$S_{pred} = \{log\_gameto,\ Pfg377\_093,\ MOI,\ PfPK2\_183\}.$$

The 4 selected variables in S_{pred} lead to the OOB error of 0.074. We also launch the variant based on forward and exhaustive search of the selection procedure. Finally it retains a set of 9 variables containing S_{pred}. The associated OOB error is 0.062 which is not far from 0.074. So we prefer a model with variables in S_{pred} which is more parsimonious.

The same procedure was applied when the outcome variable is the infection intensity as measured by the mean number of oocysts in infected mosquitoes. Figure 4 gives the behavior of VI and the OOB error at each stage of the selection procedure. 25 variables were selected by thresholding the VI in the first stage, the 2 most important being *log_gameto* and MOI. Even if only *log_gameto* is selected in the interpretation and prediction stages, we also keep MOI. Indeed, as can be seen in the bottom left graph of Figure 4, the model with these two variables is still competitive compared with the model built with *log_gameto* only.

3.2 Zero-Inflated models fitting oocyst count

Zero-Inflated negative binomial (ZINB) and Poisson (ZIP) were fitted to the data in two situations: (i) using only log-transformed of the gametocyte density as variable, (ii) using the set of variables selected for prediction of the infection prevalence or the infection intensity (see Subsection 3.1). The estimates of the parameters of ZINB and ZIP models are given in Table 2 and 3.

In situation (i), it is of interest how the zero counts are captured by the two models: they perfectly predict the observed number of non infected mosquitoes (see the left panel of Figure 5). Also, the estimates of the mean number of oocysts from both two models are similar (see the right panel of Figure 5). But according to the $\mathcal{X}^2$ goodness-of-fit test ($\mathcal{X}^2 = 48.162$, $df = 45$, $p.value \geq 0.3461$ for ZINB against $\mathcal{X}^2 = 2964.606$, $df = 46$, $p.value = 0$ for ZIP model), ZINB model is more adapted to our data. Over-dispersion is probably the main reason: there are more mosquitoes with no or few oocysts than the ones with high oocyst loads. ZIP model underestimates the number of mosquitoes with lower oocyst loads (see the left panel of Figure 5). We then consider the ZINB model in the rest of the analysis.

In situation (ii), since the data are over-dispersed, only ZINB is considered. The selected variables in the prediction step of our variable selection process using the infection prevalence as response variable are used in point mass component, and the ones using the infection intensity as response variable are used in the count component. Recall that the infection prevalence is measured by the proportion of mosquitoes that became infected, and the infection intensity by the mean number of oocysts in infected mosquitoes. It is natural to link infection prevalence and infection intensity to zero and count components respectively. We found that allele PfPK2_183 is the only variable not significant ($Z = -0.8329$, $p.value \geq 0.40$). In contrary, gametocyte density *log_gameto*, gametocyte genetic complexity MOI and allele 093 of locus $Pfg377$ significantly influence the mean oocyst load in mosquitoes in count component. The significance of the gametocyte density confirms the result obtained by ZINB

model in situation (i). The significance of the effect of MOI in both zero and count components is very interesting: it is more important in the zero component (t-test $Z = -4.5711$, $p.value < 4.9e-06$) than in the count one (t-test $Z = -2.1058$, $p.value < 3.5e-02$). Also note that the correlation is negative in both two components ($\widehat{\beta}_{MOI} = -0.0333$ and $\widehat{\gamma}_{MOI} = -0.1499$ in count and zero components respectively). So mono infected gametocyte isolates increase the probability that a mosquito do not ingest enough parasites to ensure the transmission success of *Plasmodium* through its vector mosquito. Hence, low values of MOI tend to decrease the infection prevalence. In contrary, a lower genetic diversity of gametocytes in an isolate increases the mean number of oocysts in the count component. Also note that the presence of allele 093 of the genetic marker Pfg377 increases the proportion of non-infected mosquitoes ($\widehat{\gamma}_{Pfg377\_093} = 1.2242$, $SE = 0.1204$; t-test $Z = 10.177$, $p-value < 2.7e-24$).

Table 2: Maximum likelihood estimates of the parameters of ZINB and ZIP models using only log_gameto as variable for both zero and count components. Significant codes: 0 '***'; 0.001 '**'; 0.01 '*'; 0.05 '.'; 0.1 ' '. $\mathcal{X}^2$ Goodness-of-fit test: $\mathcal{X}^2 = 47.0992$, $df = 45$, $p.value \geq 0.3866$ for ZINB against $\mathcal{X}^2 = 2834.848$, $df = 46$, $p.value = 0$ for ZIP model

		Estimate	Std. Error	z value	Pr(>|z|)	
ZINB						
Count	(Intercept)	-1.3021	0.1163	-11.1985	4.1E-29	***
	log_gameto	0.8402	0.0257	32.6835	2.7E-234	***
	Log(theta)	-0.5693	0.0557	-10.2235	1.6E-24	***
Zero	(Intercept)	0.0029	0.2405	0.0119	9.9E-01	
	log_gameto	-0.2618	0.0531	-4.9294	8.2E-07	***
ZIP						
Count	(Intercept)	-0.7941	0.0199	-40.0016	0.0E+00	***
	log_gameto	0.7717	0.0036	213.9455	0.0E+00	***
zero	(Intercept)	1.4508	0.1284	11.2996	1.3E-29	***
	log_gameto	-0.4383	0.0316	-13.8930	7.0E-44	***

4 Discussion

Plasmodium development within its vector mosquito follows complex biological processes and factors controlling parasite dynamics are not well understood. In the rodent malaria parasite *Plasmodium berghei*, it has been previously shown that the efficiency of parasite transmission from one developmental stage to another followed density-dependent models and the best fitted mathematical model differed from one developmental transition to the other one (Sinden et al., 2007). For natural populations of *Plasmodium falciparum*, the human malaria parasite, modeling becomes more challenging because of unknown genetic factors and uncontrolled environmental parameters. Nonetheless, Paul et al. (2007) found a sigmoid relationship between *Plasmodium falciparum* gametocyte density and mosquito transmission and the authors argued that parasite aggregation within mosquitoes represents an adaptive mechanism for transmission efficiency.

Table 3: Maximum likelihood estimates of the parameters of ZINB and ZIP models using $\{log\_gameto,\ Pfg377\_093,\ MOI,\ PfPK2\_183\}$ and $\{log\_gameto,\ MOI\}$ as variables in the zero and count components respectively. Significant codes: 0 '***'; 0.001 '**'; 0.01 '*'; 0.05 '.'; 0.1 ' '.

		Estimate	Std. Error	z value	Pr(>|z|)	
ZINB						
Count	(Intercept)	-0.9985	0.1436	-6.9539	3.6E-12	***
	log_gameto	0.8009	0.0261	30.6432	3.3E-206	***
	MOI	-0.0333	0.0158	-2.1058	3.5E-02	*
	Log(theta)	-0.5210	0.0500	-10.4296	1.8E-25	***
Zero	(Intercept)	0.9651	0.2679	3.6030	3.1E-04	***
	log_gameto	-0.3769	0.0534	-7.0615	1.6E-12	***
	Pfg377_093	1.2242	0.1204	10.1717	2.7E-24	***
	MOI	-0.1499	0.0328	-4.5711	4.9E-06	***
	PfPK2_183	-4.5225	5.4301	-0.8329	4.0E-01	

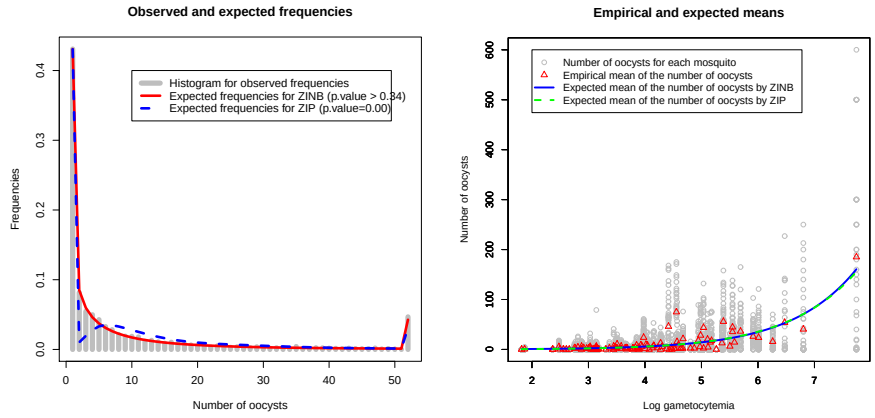

Figure 5: The right panel gives observed and predicted frequencies from ZINB and ZIP models, and the right one the empirical and predicted mean number of oocysts versus log-gametocytemia.

The great variability in *Plasmodium falciparum* oocyst numbers observed in natural *Anopheles gambiae* populations suggests that parasite transmission is the result of complex interactions between vectors and parasites, which rely on both genetic and environmental factors. Understanding factors that determine transmission intensity and then parasite population structure is of crucial importance in predicting the impact of current malaria control strategies.

In this study, we analyzed patterns of mosquito infection from experiments performed with field isolates of *Plasmodium falciparum* from Cameroon, an area of high malaria endemicity. We considered as response variables: the intensity of infection as measured by the mean of oocyst counts in infected mosquitoes, and the infection prevalence defined by the proportion of mosquitoes that became in-

fected. Gametocyte isolates were genetically characterized at seven microsatellite loci, allowing estimation of the number of co-infecting parasite clones, the MOI, and of the genetic polymorphism, given by the number of alleles at each locus. In such a situation with potentially a high number of unordered categorical variables with numerous levels and accompanying interactions, many familiar statistical techniques such as GLM over-fit the data. Then we had to face the problem of selecting the most important variables related to the outcome variables. We have addressed this issue with a selection procedure based on variable importance from random forests. The procedure has two main benefits. First, it is completely non-parametric and thus can be used on data with lots of variables of various types. Second, it answers the two distinct objectives about variable selection: (1) to find all variables related to the outcome variable and (2) to find a small number of variables sufficient for a good prediction of the outcome variable.

Recall that we are in a critical situation with the number of variables of the order of the sample size ($n = 110$). The application of the variable selection procedure on our data revealed that only 4 among the 88 variables we considered suffice to predict the infectiousness of *Plasmodium falciparum* to *Anopheles gambiae* in our experimental settings. The procedure indicates that the log-transformed of gametocyte density is the most influent variable and is positively correlated for both infection prevalence and infection intensity. This probably reflects that *Plasmodium* parasites have developed complex and diverse strategies to ensure their transmission through the mosquito vector. The fact that higher oocyst counts are found for higher gametocyte densities conforms to previous observations showing that infectiousness generally increases with gametocytemia. Interestingly, Paul et al. (2007) described upper gametocyte densities at which mosquito infection rates level off, which is consistent with our results. In their models, mosquitoes with no oocyst were treated as non infected without further consideration about the putative factors responsible of the non infected status. However, a mosquito population fed on the same gametocyte carrier results in individuals carrying high number of parasites while others do not have any. Failure to infection of a mosquito can result from various factors such the heterogeneity of gametocyte environment (Vaughan, 2007) and natural variation in mosquito susceptibility in the other hand (Riehle et al., 2006). We have described in this article an approach based on that the non-infected mosquitoes represent two distinct populations: one genetically refractory vector population and another population for which the no-oocyst status results from other biological or interacting factors. Further study to quantify the gametocyte uptake in mosquitoes fed on a single carrier would help to determine the individual variation of gametocyte density between blood-meals, and thus the real part of mosquitoes that are refractory and those that did not develop any oocyst because of other environmental factors. Nonetheless, our model perfectly predicts the number of non infected mosquitoes. Our fitting models revealed that over-dispersion of oocysts affects mosquito infection intensity. In addition, a higher over-dispersion of oocysts is observed for mosquitoes fed on blood with high gametocyte density (over 90 gametocytes/μl). The over-dispersed distribution of oocysts has often been explained as the result of the aggregation of gametocytes in the capillary blood at the time of the mosquito bite (Pichon et al., 2000). In this study, mosquitoes were membrane fed and membrane feeding is thought to suppress gametocyte over dispersion (Vaughan, 2007). Nonetheless,

the fact that the maximum aggregation is found for high gametocyte densities is indicative of aggregation of sexual stages; aggregation may occur within the mosquito midgut after parasite intake and genetic factors from the parasites may play a role in parasite recognition. This speculation is consistent with the hypothesis of adaptive aggregation, where gamete aggregation would favor fertilization and then increase infection intensity (Paul et al., 2007; Pichon et al., 2000). However, this increased oocyst burden coincided with a lower infection prevalence, possibly indicating that other factors operate in limiting mating (see below).

In malaria endemic areas, intensive use of treatments for malaria has led to the emergence of drug-resistant parasites. Despite their low efficacy, malaria therapies such as chloroquine (CQ) and sulphadoxine-pyrimethamine (SP) are still widely used in sub Saharan Africa. It has been shown that, upon treatment, drug-resistant parasites have a selective advantage, leading to higher transmission by the vector (Hallett et al., 2004, 2006). Our samples originated from an area with high drug pressure and volunteers carrying single parasite genotype may have received an early anti malarial treatment that cured them from drug-sensitive genotypes, thus allowing an optimal growth and transmission of a resistant genotype. However, children who received a malaria treatment in the one month period preceding the gametocyte carriage detection were not included in the study and genotyping of pfcrt-K76T mutation in a subset of our gametocyte samples identified single infections both as CQ resistant or sensitive parasite strains. This result indicates that other factors contribute to the better transmission capacity of the mono-infected *Plasmodium falciparum* isolates.

We found that the Multiplicity Of Infection is negatively correlated to infection intensity and positively correlated to infection prevalence (through the count and zero components respectively in the ZINB model). This indicates that the genetic complexity of gametocyte populations modulates the mosquito infection outcomes in an opposite manner: while gametocyte isolates containing a single clone of *Plasmodium falciparum* resulted in a higher mean number of oocysts in infected mosquitoes, gametocyte isolates with multiple genotypes gave rise to a higher infection prevalence. These results may suggest that malaria parasites use kin discrimination to adapt strategies allowing optimal parasite transmission.

Our results showed that the genetic complexity of gametocyte isolates affects the mosquito infection intensity. Mosquito infections with isolates of lower complexity resulted in higher oocyst counts. This may reflects a higher virulence of genotypes in these infections, where the gametocyte genotypes in the mono-infected isolates could have suppressed their competitors in a prior step of the infection, within the human host. Nonetheless, the lower infection prevalence in mono clonal infections indicates that the higher number of oocysts arises at the cost of a reduced ability to infect the mosquito vector population. This could result from blood quality/quantity such as agglutinating antibodies or anaemia. It was shown that mixed infections resulted in increased anaemia, a possible adaptive response for sex ratio adjustment (Taylor and Read, 1998; Paul et al., 2004). Sex allocation theory predicts that sex ratio becomes less female-biased as clone number increases (Read et al., 1992; Paul et al., 2002; Reece et al., 2008; Schall, 2009). Then, if parasite aggregation is an adaptive trait to promote gamete fertilization, by contrast the highly female biased sex

ratio in mono infected isolates will affect infection prevalence because male availability will constitute a limiting factor for mating.

Our results may have important implications for the genetic structuring of *Plasmodium falciparum* populations. For *Plasmodium falciparum*, fertilization of gametes can occur between genetically-identical gametes (inbreeding) or between different gametes (outbreeding). Levels of inbreeding differ from one malaria area to another but they roughly correlate with the disease endemicity (Anderson et al., 2000). In areas of high malaria endemicity, inbreeding levels are generally more reduced, mostly because parasite genetic diversity is high and multiple infections predominant. However, population genetics studies, after genotyping of oocysts from wild mosquitoes collected in intense malaria transmission areas, gave rise to conflicting results and the extent of inbreeding in natural settings remains controversial (Razakandrainibe et al., 2005; Annan et al., 2007; Mzilahowa et al., 2007). The higher fitness of inbred parasites, as suggested in this study and others (Hastings and Wedgwood-Oppenheim, 1997; Razakandrainibe et al., 2005), could explain the departs from panmixia frequently found in areas of high malaria transmission.

Finally, our results comfort the idea that malaria parasites are able to discriminate the genetic complexity of their infections and to adjust accordingly adaptive traits implicated in transmission (aggregation, sex ratio). Deciphering specific processes involved in parasite recognition and competition within the mosquito vector would help for our understanding of within host behavior of malaria parasites. This may have important implications for future malaria interventions strategies.

Appendices

A Random Forests

RF estimator

The principle of random forests is to aggregate a given number *ntree* of binary decision trees built on several bootstrap samples drawn from the learning set. The bootstrap samples are obtained by uniformly drawing n samples among the learning set with repetition. The decision trees are fully developed binary trees and the split rule is the following.

First, the whole dataset (also called the root of the tree) is split into two subsets of data (called two children nodes). To do that, one randomly chooses a given number *mtry* of variables, and computes all the splits only for the previously selected variables. A split is of the form $\{X^i \leq s\} \cup \{X^i > s\}$, which means that data with the i-th variable value less than the threshold s go to the left child node and the others to the right one. Finally the selected split is the one minimizing the variance children nodes.

Then, one restraints to one child node, randomly chooses another set of *mtry* variables and calculates the best split. And so on, until each node is a terminal node, i.e. it comprises less than 5 observations.

A new data item X, starting in the root of the tree, goes down the tree following the splits and falls in a terminal node. Then the tree predicts for X, $\bar{Y}$ the mean of response of data in this terminal node. To finally get the RF predictor, one aggregates all the tree predictors by averaging their predictions.

RF error estimate: the OOB error

Inside the variable selection procedure, we use an estimation of the prediction error directly computed by the RF algorithm. This is the Out Of Bag (OOB) error and is calculated as follows. Fix one data in the learning sample, and consider all the bootstrap samples which do not contain this data (i.e. for which the data is "out of bag"). Now perform an aggregation only among trees built on these bootstrap samples. After doing this for all data, compare to the true response and get an estimation of the prediction error (which is a kind of cross-validated error estimate).

RF variable importance

Let us now detail the computation of the RF variable importance for the first variable X^1. For each tree, one has a bootstrap sample associated with an OOB sample. Predict the OOB data with the tree predictor. Now, randomly permute the values of the first variable of the OOB observations, predict these modified OOB data with the tree predictor. The variable importance of X^1 is defined as the mean increase of prediction errors after permutation. The more the error increases, the more important the variable is (note that it can be slightly negative, typically for irrelevant variables).

B $\mathcal{ZIP}$ and $\mathcal{ZINB}$ specifications

These two models are defined by equation (1) with the count model given by:

- $\mathcal{ZIP}$:

$$\left\{\begin{array}{rcl} P\left(Y_{i,j}=y_{i,j}|\ U_{i,j}=1\right) & = & \exp\left(-\lambda_i\right)\frac{\lambda_i^{y_{i,j}}}{y_{i,j}!} \\ \lambda_i & := & \mathrm{E}\left(Y_{i,j}|\ U_{i,j}=1\right) \\ & = & \mathrm{Var}(Y_{i,j}|\ U_{i,j}=1) \end{array}\right.$$

- $\mathcal{ZINB}$:

$$\left\{\begin{array}{rcl} P\left(Y_{i,j}=y_{i,j}|\ U_{i,j}=1\right) & = & \frac{\Gamma\left(y_{i,j}+\theta\right)}{\Gamma\left(\theta\right)\cdot y_{i,j}!}\frac{\lambda_i^{y_{i,j}}\cdot\theta^{\theta}}{\left(\lambda_i+\theta\right)^{y_{i,j}+\theta}} \\ \lambda_i & := & \mathrm{E}\left(Y_{i,j}|\ U_{i,j}=1\right) \\ \mathrm{Var}(Y_{i,j}|\ U_{i,j}=1) & = & \lambda_i+\frac{1}{\theta}\lambda_i^2 \end{array}\right.$$

where $\Gamma\left(t\right)=\int_0^{\infty}x^{t-1}\mathbf{e}^{-x}\mathbf{d}x$, and θ is the over-dispersion parameter. The expectation and the variance of $Y_{i,j}$ are given by:

$$\begin{array}{rcl} \mu\left(x\right) & := & \mathrm{E}\left(Y_{i,j}|\right)=\Big(1-\pi_i\Big)\lambda_i \\ \mathrm{Var}(Y_{i,j}) & = & \left\{\begin{array}{ll} \Big(1-\pi_i\Big)\Big(\lambda_i+\pi_i\lambda_i^2\Big) & ZIP \\ \Big(1-\pi_i\Big)\Big(\lambda_i+\left(\frac{1}{\theta}+\pi_i\right)\lambda_i^2\Big) & ZINB. \end{array}\right. \end{array}$$

References

Anderson, T. J., Haubold, B., Williams, J. T., Estrada-Franco, J. G., Richardson, L., Mollinedo, R., Bockarie, M., Mokili, J., Mharakurwa, S., French, N., Whitworth, J., Velez, I. D., Brockman, A. H., Nosten, F., Ferreira, M. U., and Day, K. P. (2000). Microsatellite markers reveal a spectrum of population structures in the malaria parasite plasmodium falciparum. *Mol Biol Evol*, 17(10):1467–82.

Annan, Z., Durand, P., Ayala, F. J., Arnathau, C., Awono-Ambene, P., Simard, F., Razakandrainibe, F. G., Koella, J. C., Fontenille, D., and Renaud, F. (2007). Population genetic structure of Plasmodium falciparum in the two main african vectors, anopheles gambiae and anopheles funestus. *Proc Natl Acad Sci U S A*, 104(19):7987–92.

Babiker, H. A., Ranford-Cartwright, L. C., and Walliker, D. (1999). Genetic structure and dynamics of plasmodium falciparum infections in the kilombero region of tanzania. *Trans R Soc Trop Med Hyg*, 93 Suppl 1:11–4.

Bell, A. S., de Roode, J. C., Sim, D., and Read, A. F. (2006). Within-host competition in genetically diverse malaria infections: parasite virulence and competitive success. *Evolution*, 60:1358–71.

Boudin, C., Diop, A., Gaye, A., Gadiaga, L., Gouagna, C., Safeukui, I., and Bonnet, S. (2005). Plasmodium falciparum transmission blocking immunity in three areas with perennial or seasonal endemicity and different levels of transmission. *Am J Trop Med Hyg*, 73(6):1090–5.

Boudin, C., Van Der Kolk, M., Tchuinkam, T., Gouagna, C., Bonnet, S., Safeukui, I., Mulder, B., Meunier, J. Y., and Verhave, J. P. (2004). Plasmodium falciparum transmission blocking immunity under conditions of low and high endemicity in cameroon. *Parasite Immunol*, 26:105–10.

Breiman, L. (2001). Random forests. *Machine Learning*, 45(1):5.

Breiman, L., Friedman, J. H., Olshen, R. A., and Stone, C. J. (1984). *Classification and Regression Trees*. Statistics/Probability Series. Wadsworth Publishing Company, Belmont, California, U.S.A.

Day, K. P., Koella, J. C., Nee, S., Gupta, S., and Read, A. F. (1992). Population genetics and dynamics of plasmodium falciparum: an ecological view. *Parasitology*, 104 Suppl:S35–52.

de Roode, J. C., Helinski, M. E., Anwar, M. A., and Read, A. F. (2005). Dynamics of multiple infection and within-host competition in genetically diverse malaria infections. *Am Nat*, 166:531–42.

Drakeley, C. J., Secka, I., Correa, S., Greenwood, B. M., and Targett, G. A. (1999). Host haematological factors influencing the transmission of plasmodium falciparum gametocytes to anopheles gambiae s.s. mosquitoes. *Trop Med Int Health*, 4:131–8.

Genuer, R., Poggi, J.-M., and Tuleau-Malot, C. (2010). Variable selection using random forests. *Pattern Recognition Letters*, 31:2225–2236.

Hallett, R. L., Dunyo, S., Ord, R., Jawara, M., Pinder, M., Randall, A., Alloueche, A., Walraven, G., Targett, G. A., Alexander, N., and Sutherland, C. J. (2006). Chloroquine/sulphadoxine-pyrimethamine for gambian children with malaria: transmission to mosquitoes of multidrug-resistant plasmodium falciparum. *PLoS Clin Trials*, 1:e15.

Hallett, R. L., Sutherland, C. J., Alexander, N., Ord, R., Jawara, M., Drakeley, C. J., Pinder, M., Walraven, G., Targett, G. A., and Alloueche, A. (2004). Combination therapy counteracts the enhanced transmission of drug-resistant malaria parasites to mosquitoes. *Antimicrob Agents Chemother*, 48:3940–3.

Hastings, I. M. and Wedgwood-Oppenheim, B. (1997). Sex, strains and virulence. *Parasitol Today*, 13:375–83.

Hogh, B., Gamage-Mendis, A., Butcher, G. A., Thompson, R., Begtrup, K., Mendis, C., Enosse, S. M., Dgedge, M., Barreto, J., Eling, W., and Sinden, R. E. (1998). The differing impact of chloroquine and pyrimethamine/sulfadoxine upon the infectivity of malaria species to the mosquito vector. *Am J Trop Med Hyg*, 58:176–82.

Jaramillo-Gutierrez, G., Rodrigues, J., Ndikuyeze, G., Povelones, M., Molina-Cruz, A., and Barillas-Mury, C. (2009). Mosquito immune responses and compatibility between plasmodium parasites and anopheline mosquitoes. *BMC Microbiol*, 9:154.

Liaw, A. and Wiener, M. (2002). Classification and Regression by randomForest. *R news*, 2(3):18–22.

Mzilahowa, T., McCall, P. J., and Hastings, I. M. (2007). population structure and genetics of the malaria agent p. falciparum. *PLoS One*, 2:e613.

Nwakanma, D., Kheir, A., Sowa, M., Dunyo, S., Jawara, M., Pinder, M., Milligan, P., Walliker, D., and Babiker, H. A. (2008). High gametocyte complexity and mosquito infectivity of plasmodium falciparum in the gambia. *Int J Parasitol*, 38(2):219–27.

Paul, R. E., Lafond, T., Muller-Graf, C. D., Nithiuthai, S., Brey, P. T., and Koella, J. C. (2004). Experimental evaluation of the relationship between lethal or non-lethal virulence and transmission success in malaria parasite infections. *BMC Evol Biol*, 4:30.

Paul, R. E. L., Bonnet, S., Boudin, C., Tchuinkam, T., and Robert, V. (2007). Aggregation in malaria parasites places limits on mosquito infection rates. *Infect Genet Evol*, 7(5):577–86.

Paul, R. E. L., Brey, P. T., and Robert, V. (2002). Plasmodium sex determination and transmission to mosquitoes. *Trends in Parasitology*, 18:32–38.

Pichon, G., Awono-Ambene, H. P., and Robert, V. (2000). High heterogeneity in the number of plasmodium falciparum gametocytes in the bloodmeal of mosquitoes fed on the same host. *Parasitology*, 121 (Pt 2):115–20.

R Development Core Team. (2009). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0.

Razakandrainibe, F. G., Durand, P., Koella, J. C., Meeüs, T. D., Rousset, F., Ayala, F. J., and Renaud, F. (2005). "clonal" population structure of the malaria agent Plasmodium falciparum in high-infection regions. *Proc Natl Acad Sci U S A*, 102(48):17388–93.

Read, A. F., Narara, A., Nee, S., Keymer, A. E., and Day, K. P. (1992). Gametocyte sex ratios as indirect measures of outcrossing rates in malaria. *Parasitology*, 104 (Pt 3):387–95.

Reece, S. E., Drew, D. R., and Gardner, A. (2008). Sex ratio adjustment and kin discrimination in malaria parasites. *Nature*, 453:609–14.

Ribaut, C., Berry, A., Chevalley, S., Reybier, K., Morlais, I., Parzy, D., Nepveu, F., Benoit-Vical, F., and Valentin, A. (2008). Concentration and purification by magnetic separation of the erythrocytic stages of all human plasmodium species. *Malar J*, 7:45.

Riehle, M. M., Markianos, K., Niaré, O., Xu, J., Li, J., Touré, A. M., Podiougou, B., Oduol, F., Diawara, S., Diallo, M., Coulibaly, B., Ouatara, A., Kruglyak, L., Traoré, S. F., and Vernick, K. D. (2006). Natural malaria infection in anopheles gambiae is regulated by a single genomic control region. *Science*, 312(5773):577–9.

Schall, J. J. (2009). Do malaria parasites follow the algebra of sex ratio theory? *Trends Parasitol*, 25:120–3.

Segal, M. R., Cummings, M. P., and Hubbard, A. E. (2001). Relating amino acid sequence to phenotype: analysis of peptide-binding data. *Biometrics*, 57(2):632–42.

Sinden, R. E., Dawes, E. J., Alavi, Y., Waldock, J., Finney, O., Mendoza, J., Butcher, G. A., Andrews, L., Hill, A. V., Gilbert, S. C., and Basanez, M. G. (2007). Progression of plasmodium berghei through anopheles stephensi is density-dependent. *PLoS Pathog*, 3:e195.

Targett, G., Drakeley, C., Jawara, M., von Seidlein, L., Coleman, R., Deen, J., Pinder, M., Doherty, T., Sutherland, C., Walraven, G., and Milligan, P. (2001). Artesunate reduces but does not prevent posttreatment transmission of plasmodium falciparum to anopheles gambiae. *J Infect Dis*, 183:1254–9.

Taylor, L. H. and Read, A. F. (1998). Determinants of transmission success of individual clones from mixed-clone infections of the rodent malaria parasite, plasmodium chabaudi. *Int J Parasitol*, 28:719–25.

Vaughan, J. A. (2007). Population dynamics of Plasmodium sporogony. *Trends Parasitol*, 23(2):63–70.

Wargo, A. R., de Roode, J. C., Huijben, S., Drew, D. R., and Read, A. F. (2007). Transmission stage investment of malaria parasites in response to in-host competition. *Proc Biol Sci*, 274:2629–38.

Xiang, L., Lee, A., Yau, K., and McLachlan, G. (2007). A score test for overdispersion in zero-inflated poisson mixed regression model. *Statistics in medicine*, 26(7):1608–1622.

Zeileis, A. and Jackman, C. K. S. (2008). Regression Models for Count Data in **R**. *Journal of Statistical Software*, 27.

Centre de recherche INRIA Saclay – Île-de-France
Parc Orsay Université - ZAC des Vignes
4, rue Jacques Monod - 91893 Orsay Cedex (France)

Centre de recherche INRIA Bordeaux – Sud Ouest : Domaine Universitaire - 351, cours de la Libération - 33405 Talence Cedex
Centre de recherche INRIA Grenoble – Rhône-Alpes : 655, avenue de l'Europe - 38334 Montbonnot Saint-Ismier
Centre de recherche INRIA Lille – Nord Europe : Parc Scientifique de la Haute Borne - 40, avenue Halley - 59650 Villeneuve d'Ascq
Centre de recherche INRIA Nancy – Grand Est : LORIA, Technopôle de Nancy-Brabois - Campus scientifique
615, rue du Jardin Botanique - BP 101 - 54602 Villers-lès-Nancy Cedex
Centre de recherche INRIA Paris – Rocquencourt : Domaine de Voluceau - Rocquencourt - BP 105 - 78153 Le Chesnay Cedex
Centre de recherche INRIA Rennes – Bretagne Atlantique : IRISA, Campus universitaire de Beaulieu - 35042 Rennes Cedex
Centre de recherche INRIA Sophia Antipolis – Méditerranée : 2004, route des Lucioles - BP 93 - 06902 Sophia Antipolis Cedex

Éditeur
INRIA - Domaine de Voluceau - Rocquencourt, BP 105 - 78153 Le Chesnay Cedex (France)
http://www.inria.fr
ISSN 0249-6399