2012•2013

FACULTY OF SCIENCES

Master of Statistics: Biostatistics

Masterproef

Modelling of Malaria Incidence data: Comparison of Shared frailty Model and Count Regression model

Promotor :
Prof. dr. Paul JANSSEN
Prof. dr. Anneleen VERHASSELT

Promotor :
Prof. dr. LUC DUCHATEAU
Mr. YEHENEW GETACHEW

Alex Sila

Master Thesis nominated to obtain the degree of Master of Statistics , specialization Biostatistics

Transnational University Limburg is a unique collaboration of two universities in two countries: the University of Hasselt and Maastricht University.

universiteit hasselt
KNOWLEDGE IN ACTION

Universiteit Hasselt | Campus Hasselt | Martelarenlaan 42 | BE-3500 Hasselt
Universiteit Hasselt | Campus Diepenbeek | Agoralaan Gebouw D | BE-3590 Diepenbeek

universiteit hasselt | Maastricht University

2012•2013
FACULTY OF SCIENCES
Master of Statistics: Biostatistics

Masterproef

Modelling of Malaria Incidence data: Comparison of Shared frailty Model and Count Regression model

Promotor :
Prof. dr. Paul JANSSEN
Prof. dr. Anneleen VERHASSELT

Promotor :
Prof. dr. LUC DUCHATEAU
Mr. YEHENEW GETACHEW

Alex Sila

Master Thesis nominated to obtain the degree of Master of Statistics , specialization Biostatistics

universiteit hasselt | Maastricht University

Dedication

This thesis is dedicated to all my family members. I shall also, in a special way dedicate this piece of work to my friends Alex, David, Johnson, Michael, Ken, Alice, Francis, Sofy, Ryan, Stanley and all my other friends for their unflinching moral support.

Acknowledgements

Successful completion of this thesis has been through the support of a number of individuals. First of all, I acknowledge my supervisors Prof. dr. Paul Janssen, Prof. dr. Luc Duchateau, Prof. dr. Anneleen Verhasselt and Mr. Yehenew Getachaw who continually guided me through this work, by supplying thoughtful suggestions for improving this report. In particular, I thank Jimma University for allowing me to use their data and for providing me with all the necessary study description materials.

I also send my gratitude to VLIR-UOS scholarship for awarding me this important opportunity for developing my knowledge and expertise in the field of statistics which has really improved my views in terms of my academic perspective and statistics as a whole. Thanks to VLIR-UOS coordinator of Hasselt University Mrs Martine Machiels for all the support and timely communication during the entire study period. I thank all the professors of the master of statistics programme for their guidance and teaching expertise. I am also grateful to all my classmates and friends for making my stay in Belgium lively and enjoyable.

Most of all I thank God for giving me wisdom, strength and life to successfully complete this program.

Alex Sila

University of Hasselt, Belgium,

September, 2013

Contents

List of Figures

List of Tables

Abstract

Malaria is a major health challenge in most of the developing countries leading to high morbidity and mortality rates in the population. Seasonal variations, development of water projects and their operation have some long history of facilitating increased transmission of vector borne diseases. The study was motivated by the availability of the time-to-event dataset, which contains data on time-to-first *P.falciparum* malaria infection. The main interest of the study is to investigate the influence of seasons and distance to the Gilgel-Gibe hydroelectric dam reservoir on the time-to-first malaria infection in children aged less than 10 years. To incorporate the seasonal variation, a more flexible proportional hazards model was fitted by making some mild assumptions concerning the baseline hazard. The resulting piecewise constant hazards model is equivalent to a count regression model with aggregated event counts as response variable and the log of exposure time as an offset. However, due to the aggregation of the event counts for each period per village we expect some loss of power as compared to the piecewise constant hazards model which we investigated through simulation studies. Marginal and conditional models were used to analyse the clustered events.

Results indicate a significant seasonal effect for the time to first malaria infection. None of the models fitted had a significant distance effect, but it can be noted that three of the four models predicts an increasing incidence with increasing distance to the dam. Aggregation of the events per the season of infection lead to some loss of power though not much, hence the count regression seems a good alternative to the survival modelling.

To maximize on the economic benefits generated by Gilgel-Gibe hydroelectric dam, preventive programmes against malaria and other related vector-borne diseases need to be implemented in the households close to the dam reservoir and establishment of an early warning systems for malaria outbreak especially at the onset of rains.

Keywords: Count regression, Piecewise constant hazards model, *P.falciparum*.

1 Introduction

1.1 Background

Malaria is a major health concern in most parts of Tropical and sub - Tropical regions of Africa, Asia, South and Central America which might be attributed to the warm temperatures that provide ideal habitats for mosquito larvae development. The World Health Organization (WHO) estimates the annual malaria infections in Africa to be over 300 million and over 1 million annual deaths (WHO, 2001; WHO, 2010). Over 400 species of malaria parasites are said to exist, with five of them known to infect humans (Service, 1991). The transmission of malaria from one person to another is by a bite of an infected female *Anopheles* mosquito. In most African countries, studies have shown that the flight range of different species of *Anopheles* mosquitoes ranges from 0.8 km to a maximum flight of about 3 km (Yewhalaw et al., 2009; Thomson et al., 1995). Malaria shows a strong seasonal pattern with a lag time varying from a few weeks at the onset of the rainy season to more than a month at the end of the rainy season (Adugna, 2011; White, 1974). In Ethiopia, it is estimated that 75% of the land mass lie below 2000 metres above the sea level and is malaria prone, hence two-thirds of the country's population is at risk of malaria, with an average of 7 million reported cases and 70,000 deaths per year.

Malaria has led to high morbidity and mortality rate in the population leading to reduced production activities. Settlement patterns have been influenced by the prevalence of malaria, with concentration of population in less risk highland areas which has resulted in a massive environmental degradation and loss of productivity exposing a large proportion of the country's population to a continued poverty (Adugna, 2011). Increased school absenteeism during malaria epidemics significantly reduces learning capacity of children (Karunamoorthi and Bekele, 2012). Coping with malaria epidemics overwhelms the capacity of the health services increasing the public health expenditures substantially (Gabriel et al., 2005). This makes malaria not just a health issue but a food security and environmental issue as well.

Ecological disturbances due to human actions such as deforestation, the construction of dams and establishment of new settlements in previously unsettled areas allow for the proliferation of mosquitoes that prefer human habitation to natural settings (Tulu, 1993; Lindsay et al., 1995). In developing countries and more so in Ethiopia, dams and other related water projects continue to be planned, constructed and operated to meet human needs such as drinking water, energy generation and agricultural production (Adugna, 2011). The potential for dams to alleviate poverty leads to enhanced human health and quality of life but at the same time increasing the likelihood of human infection due to some waterborne related diseases like schistosomiasis, malaria, dysentery and river blindness (Steinmann et al., 2006). Various studies have been done to investigate malaria incidence and prevalence in dam sites compared to a distant site, however consistent results have not yet been obtained. For example, a recent study in Northern Ethiopia investigating the possible impacts of small dams on malaria transmission showed that the rate of infection among children living close to dams was more than in communities with no dams (Ghebreyesus et al., 1999). However, in other related study it was also found that dam areas displayed a lower malaria transmission compared with distant setting when integrated vector management or other control interventions had been applied. For example, in India, a study which compared the parasitological indices in dam area to forest or plain areas, recorded a prevalence and annual parasite incidence of zero in dam area (Shukla et al., 2001).

Gilgel-Gibe hydroelectric dam in South Western Ethiopia was created by impounding the waters of Gilgel-Gibe river and is currently the largest supply of power (184 MW) in Ethiopia and has been operational since 2004. During its construction many people were relocated upstream of the reservoir, although some still remained close to the buffer zone surrounding the dam. The location of the villages near the newly formed reservoir may be attributed to the increased malaria transmission assuming that the reservoir contributes directly or indirectly to the presence of breeding grounds for malaria vectors (Yewhalaw et al., 2009). The current study investigates the effects of the dam on malaria incidence among children aged below 10 years, focusing on the distribution of infection in relation to distance of villages to the reservoir

shore and season of infection using survival and count regression models.

1.2 Motivation of the study

This study was motivated by the availability of the time-to-event dataset, which contains data on time-to-first *P.falciparum* malaria infection, place of residence (where households are nested within village) and distance of the household from the reservoir. In this study we are interested in investigating the influence of different seasons and distance to the dam reservoir on time-to-first *P.falciparum* malaria event. For modelling such type of data, proportional hazards model are often used. However, South Western Ethiopia experiences three climatic seasons per year, to incorporate the seasonal variation in our model, a more flexible proportional hazards model has to be fitted by making some mild assumptions concerning the baseline hazard. The baseline hazard represents the seasonal variations which are considered constant in each period, leading to a piecewise exponential hazards model. Laird and Oliver (1981) noted that the piecewise exponential hazards model is equivalent to a log-linear model with the events indicator as the response and the log of exposure time as an offset. In the context of this study, we will model the aggregated event counts per period in each village as a response assumed to be independent Poisson observations and the effect of the distance is assessed by use of the averaged household distances per village. However, due to the aggregation of the event counts for each period per village we expect some power loss as compared to the piecewise exponential hazards model.
Since children are clustered within villages, two approaches to analyzing data with clustered events will be used, which include; the Marginal approach - assume no dependence in the data and likelihood-based random effects (frailty) model - conditional models assuming dependence in the data. The efficiency and information loss of the aggregated event times for the mixed Poisson regression model will be investigated through simulation studies.

1.3 Study Objectives

The main objectives of this study are;

- To investigate the effect of seasonal variations and distance to the dam reservoir on the time to first *P. falciparum* malaria event.
- To compare and study the efficiency of the survival model and its equivalent log-linear model in terms of coverage probabilities and power through simulation studies.

The findings of this study could assist in the proper planning and development of dam associated malaria control programmes. Information on the effect of seasonal variations is important in the implementation of effective interventions and establishment of an early warning system for malaria outbreak.

1.4 Study design

The dataset used for this study was generated by one of the VLIR-OUS Inter University Collaboration (IUC) programme with Jimma University (Ethiopia). The IUC-Jimma University aims at strengthening the institutional capacity and improving the quality of life for the surrounding communities. The project focuses on the impact of the Gilgel-Gibe hydroelectric dam in terms of human and animal health, ecology and agronomy. This particular study was undertaken with the overall aim of determining malaria incidence and patterns of its transmission among children living close to the dam. The malaria incidence study was conducted within a duration of 2 year period (July 2008 - June 2010) among children aged less than 10 years living in villages around the dam as shown in Figure 1 below. Prior to the study, villages within 10 km radius from the dam reservoir shore were first identified and 16 villages randomly selected among them. The sampled villages were classified as either 'at risk villages' - within 3 km distance from the dam shore or 'control villages' - more than 3 km distance from the dam shore based on the maximum flying ability of mosquito. Each sampled village was assumed to expe-

rience similar eco-topography, access to health facilities and to be homogeneous with respect to socio-cultural and daily economic activities (Yewhalaw et al., 2013). Children recruited to the study had to have resided in the study area for a duration of at least 6 months prior to the study and intended to remain in the study area for the entire follow up period. A total of 2082 children were recruited into the study, with each village i (i=1,...,16) having around 130 children sampled. Baseline characteristics were collected in regards to individual and household characteristics. Each of the sampled child j ($j = 1, ..., n_i$) in a village i was followed-up by a trained data collector on a weekly interval based on house to house visits until the first incidence of malaria. The primary outcome is the actual time in days from the start of the study to the day of first *P. falciparum* malaria infection. A child suspected to have malaria based on the either of this symptoms (fever, pain and sweating) had to be confirmed through screening of the blood for the plasmodium parasite. Censoring was caused by death, dropout or end of the study. Due to uncontrollable factors in the course of the study duration about 42 children were lost to follow-up (deaths, migration, among other factors), hence the final dataset used for the analysis consists of 2040 children.

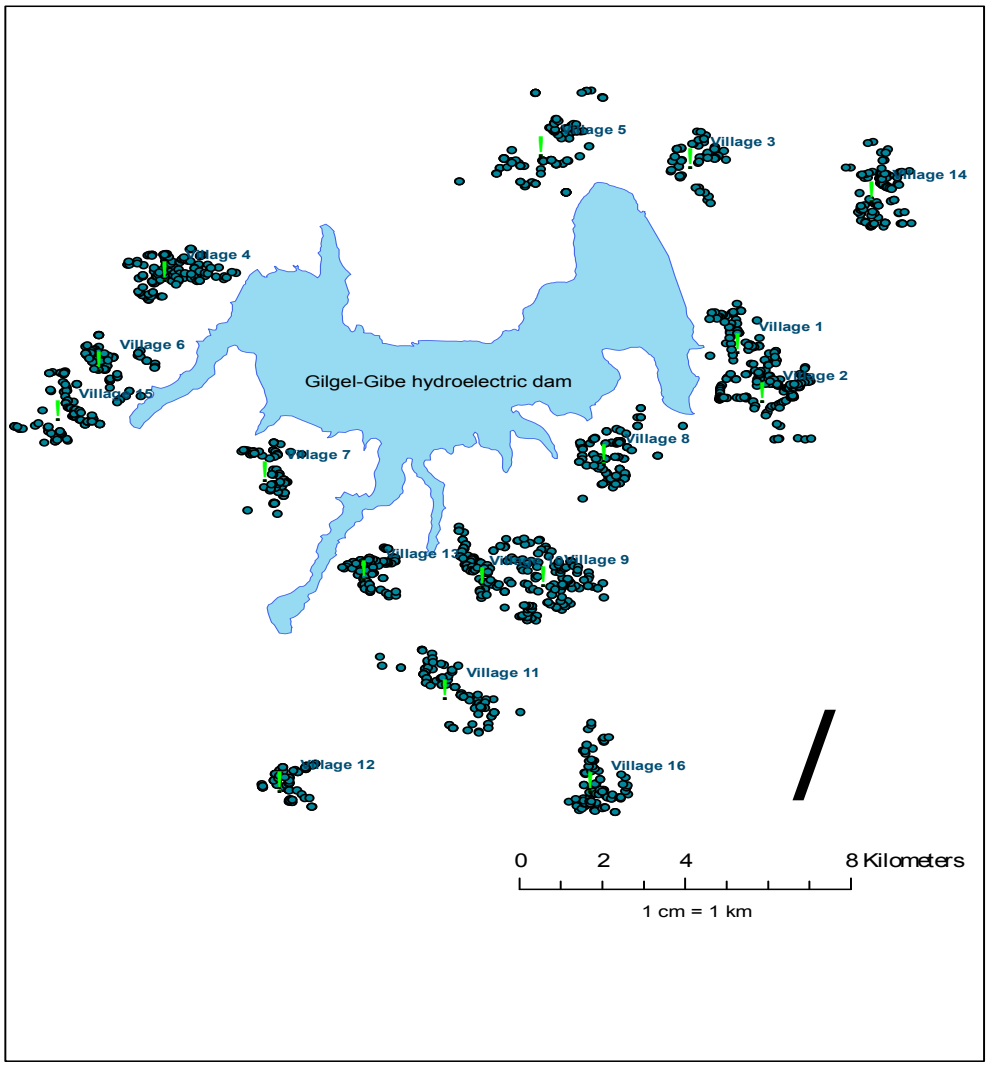

Figure 1: *Map showing the study villages and the distribution of the Gilgel-Gibe hydroelectric dam reservoir*

2 Methodology

In this study we are interested in the time to the first malaria event. The clustering of children in the villages induces dependence among the event times. Clustered multivariate time-to-event data appear predominantly in most of the clinical and epidemiological research, e.g. several types of events in an individual, time-to-event of children clustered within a village or family studies, etc. Two approaches are commonly used when modelling multivariate time-to-event data, though other computationally simpler alternatives exist with major drawbacks as compared to the two proposed approaches (Janssen and Duchateau, 2008). The marginal approach is used when the dependencies between the event times are not of interest, population parameters are estimated as if the event times were independent, but the dependence between event times is taken into account when estimating the variance of the parameter estimates (Lin, 1994). A random effect model (Hougaard, 2000) is often used when the dependence between the event times is of interest. Here the dependency of the related event times is introduced by a shared unobserved risk factor. The commonly used random effect model for clustered time-to-event data is the shared frailty model, where a single risk factor (frailty) introduces a symmetric dependence, i.e. the same dependencies between subjects in the same cluster.

To study the effect of seasonal variations on malaria infection, the study duration is partitioned into 6 periods (3 seasons per year) to represent the different seasons (long rain, short rain and dry season) as experienced in South Western Ethiopia. To incorporate the seasonal variation based on our data, a more flexible proportional hazards model will be fitted by making some mild assumptions concerning the baseline hazard. The baseline hazard will represent the seasonal variations which are considered constant in each period, leading to a piecewise constant hazards model which is equivalent to a Poisson regression model (Laird and Oliver, 1981). For the survival modelling, we model the minimum of the censoring time c_{ij} and event time t_{ij}, $y_{ij} = min(c_{ij}, t_{ij})$ and δ_{ij} the censoring indicator, taking the value one if an event is observed, otherwise zero. We assume noninformative censoring. For the count regression we will aggregate event times per period in each village. Period k starts at time τ_{k-1} and ends at time τ_k, where

k = 1,2,...,6. Aggregating event times for each period-village combination leads to small set of summary statistics for village i as shown below;

- The aggregated event time in period k,

  $d_{ik} = \sum_{j=1}^{n_i} \delta_{ij} I(\tau_{k-1} < y_{ij} \leq \tau_k)$

- Total time at risk (exposure time) for village i in period k, a_{ik}

The aggregated event counts are assumed independent between villages but dependent within the same village. Due to the aggregation of the event counts for each period per village, we expect some loss of power which will be investigated through simulation studies.

2.1 Marginal Models

In the marginal model approach, we do not take into account the cluster effect (village) and act as if the event times are independent of each other, even when the children belong to the same cluster, which result in an independent contribution to the likelihood by each child - Independence working model. The marginal parameter estimates obtained from assuming an independence working model are consistent estimators for the population based parameters. However, the standard errors are not because of correlation between survival times, therefore sandwich estimators that cope with the dependence in the data will have to be used to obtain a consistent estimate of covariance matrix by using the grouped jackknife technique (Janssen and Duchateau, 2008).

2.1.1 Survival model

A survival model adjusted for distance and seasonal variations, where the distance effect is introduced as a continuous fixed effect is fitted to the individual subject ignoring the cluster effect. The seasonal effect is introduced via the baseline hazard which is considered constant in each of the seasons. An extension of the proportional hazards model under relatively mild

assumptions about the baseline hazard $h_0(t_i)$ is fitted to the data. According to the proportional hazards model, the hazard rate can be written as a product of two functions, one merely depending on time t_i and another depending on the covariates. Unlike the constant hazards model where $h_0(t_i)$ is assumed to remain constant over the whole range of time, our proposed model assumes that the baseline hazard is constant in time intervals (seasons). The model is as described below:

$$h_{ijk}(t) = h_0(t_i)\exp(\beta_d x_{ij})$$

where we will assume that the baseline hazard is estimated from the data and is constant within each period, so that

$$h_0(t_i) = \lambda_k \text{ for } t_i \text{ in } [\tau_{k-1}, \tau_k)$$

Thus we will model the baseline hazard $h_0(t_i)$ using k parameters, λ_k, where $k = 1, ..., 6$, each representing the baseline hazard for period k and $\exp(\beta_d x_{ij})$ representing the relative risk for a child j from village i with household distance x_{ij} compared to the baseline hazard at any given time. Since we assume the risk to be periodwise constant, the corresponding survival function is often called the piecewise exponential. The choice of the cut points is based on the actual seasons in Jimma - Ethiopia.

To model the seasonal and year effects separately, we introduce the following covariates

$$x_{yk} = \begin{cases} 1 & \text{k>3} \\ 0 & \text{otherwise} \end{cases}$$

$$x_{s2,k} = \begin{cases} 1 & \text{k=2,5} \\ 0 & \text{otherwise} \end{cases}$$

$$x_{s3,k} = \begin{cases} 1 & \text{k=3,6} \\ 0 & \text{otherwise} \end{cases}$$

Hence the final piecewise exponential proportional hazards model fitted to the malaria incidence data after imposing the season and year categorization is:

$$h_{ij}(t) = \sum_{k=1}^{6} \exp(\lambda_{ijk}) I(\tau_{k-1} < t \leq \tau_k)$$

where

$$\lambda_{ijk} = \lambda_0 + \lambda_y x_{yk} + \lambda_{s2} x_{s2,k} + \lambda_{s3} x_{s3,k} + \beta_d x_{ij}$$

with λ_0, λ_y, λ_{s2}, λ_{s3} and β_d representing the year1-season1 effect, year2 effect, second season effect, third season effect and the distance effect respectively.

2.1.2 Count regression model

Holford (1980), Laird and Oliver (1981) noted that the piecewise exponential hazards model discussed above is equivalent to a log-linear model with the events indicator as the response and the log of exposure time as an offset. We will model the aggregated event counts for each period-village combination as the response variable, d_{ik}, (i = 1,...,16 k = 1,...,6) and the corresponding total time at risk, a_{ik}, as a fixed offset variable. The response variables will be assumed to be independent Poisson observations and the effect of the distance is assessed by use of the averaged village distance, $\bar{x_{i.}} = (\frac{\sum_{j=1}^{n_i} x_{ij}}{n_i})$, since the village is the basis for event times aggregation. McCullagh and Nelder (1989) provides a clear description of a log-linear model for handling multivariate count data. Due to dependence of the data, inappropriate Poisson regression model (standard) will underestimate the standard errors and overstate the significance of the regression parameters, and consequently give misleading inference about the regression parameters. The expected number of events, $E(d_{ik})$, will be represented by ξ_{ik}. Different distributions have been proposed to handle dependence in data modelling (Ismail et al., 2007). In our case we will make use of the mixed Poisson regression modelling (Lawless et al., 1989) assuming independence working correlation between the different periods within the same village. The final marginal count regression model fitted assuming the same season - year restrictions as in the survival modelling is:

$$\log(\xi_{ik}) = \log(a_{ik}) + \eta_{ik}$$

where

$$\eta_{ik} = \beta_0 + \beta_y x_{yk} + \beta_{s2} x_{s2,k} + \beta_{s3} x_{s3,k} + \beta_d \bar{x_i}.$$

and β_0, β_{s2}, β_{s3}, β_y and β_d the effect of the year1-season1, second season, third season, second year and distance respectively. Assuming $d_{ik} \sim Poisson(\xi_{ik})$ links the observed counts to the population parameter.

2.2 Conditional Models

To capture the dependence in time between counts measured at different periods in the same village and dependence of children in the same cluster, conditional models will be utilized. Children in the same village share the same environmental conditions and therefore we expect some association between them. To capture the association between the children in the same village or the aggregated events in a village for the different periods, we shall introduce a random effect, ζ_i, for each village to explain the unobserved heterogeneity not captured by the covariates. The random effects (frailty) are assumed to be observations from a probability distribution with zero mean and variance, σ^2_ζ, where the variance is to be estimated from the data. Recent studies show that ignoring heterogeneity in the data may lead to inaccurate conclusions and underestimation of the standard errors. Oakes (1989) proposed frailty models for bivariate survival times and introduced several possible frailty models. He believed that improper modelling of heterogeneity would result in biased estimates since the covariates in the model fail to explain the true effect of the covariates on the response variable. Considerable progress has been made in recent years in the area of random effects in generalised linear models (Breslow et al., 1993; McGilchrist, 1994 and Nelder et al., 1996)

2.2.1 Frailty Model

Frailty models are extensions of the proportional hazards model. In most clinical and epidemiological studies, survival analysis assumes a homogenous population. This means that all sampled individuals are subject to the same risk (e.g. risk of an infection, risk of death) apart from covariates introduced in the model. In many applications, the sampled individuals can not be assumed to be homogeneous, rather should be considered as an heterogeneous sample. Various reasons for heterogeneity exist: difference in locations, not all relevant covariates related to the event of interest are measured due to economical reasons and the importance of some covariates might still be unknown. The frailty approach is a statistical modelling concept which aims to account for this heterogeneity, caused by unmeasured covariates (Wienke, 2003). In statistical terms, a frailty model is a random effect model for time-to-event data, where the random effect has a multiplicative effect on the baseline hazard function. A natural way to model dependence of clustered event times is through the introduction of a cluster-specific random effect - the shared frailty. Children in the same cluster are assumed to share the same frailty term ζ_i, hence the name shared frailty model as introduced by Clayton (1978) and extensively studied in Hougaard (2000). The survival times are assumed to be conditionally independent with respect to the shared frailty. One important problem in the area of frailty models is the choice of the frailty distribution. The most often used frailty distributions include the gamma distribution, the positive stable distribution, the compound Poisson distribution and the log-normal distribution. For comparability with the mixed Poisson regression model we shall use a log-normal frailty as proposed by McGilchrist (1993). The shared log-normal frailty model is written as;

$$h_{ij}(t) = \sum_{k=1}^{6} \exp(\lambda_{ijk} + \zeta_i) I(\tau_{k-1} < t \leq \tau_k)$$

with ζ_i normally distributed with zero mean and variance σ_s^2. The corresponding frailty, $u_i = \exp(\zeta_i)$ has a log-normal distribution;

$$f_U(u_i) = \frac{1}{u_i \sqrt{2\pi\sigma_s^2}} \exp(-\frac{(\log u_i)^2}{2\sigma_s^2})$$

with $\sigma_s^2 > 0$, estimated from the data. Unlike other frailty distributions (Janssen and Duchateau, 2008), it is natural to assume a zero mean normal distribution for the random effect ζ_i, for the log-normal distribution. Hence we are able to compare the random effect of the frailty model and the conditional mixed Poisson regression model.

2.2.2 Conditional mixed Poisson regression model

To capture the dependence of the aggregated event counts in the different periods within the same cluster/village, a random effect ζ_i, for each village is included in the model. The conditional mixed Poisson regression model fitted is given below;

$$\log(\xi_{ik}) = \log(a_{ik}) + \eta_{ik} + \zeta_i$$

with ζ_i normally distributed with zero mean and variance σ_p^2.
In the subsequent section, it's shown that the loglikelihood of the frailty model and the conditional mixed Poisson model are equivalent.

2.3 Equivalence of the Frailty and Count regression model

One of the procedures for estimating the parameters of a specified parametric model is the method of maximum likelihood. According to this method, the partial derivatives of the log-likelihood function are set equal to zero and the obtained system of equations solved. The connection between the proportional hazards and log-linear regression models has long been recognised (Whitehead, 1980; Laird et al., 1981). The two modelling techniques lead to the same parameter estimates, though the mixed Poisson regression model uses the averaged household distances whereas the survival modelling uses the individual household distance. The equivalence of the loglikelihoods of the two statistical modelling techniques is as shown below based on a paper by (Getachew et al., 2013).
The frailty model fitted using the individual household distance as discussed previously is

$$h_{ij}(t) = \sum_{k=1}^{6} \exp(\lambda_{ijk} + \zeta_i) I(\tau_{k-1} < t \leq \tau_k)$$

where $u_i = \exp(\zeta_i)$ and $\lambda_{ijk} = \lambda_0 + \lambda_y x_{yk} + \lambda_{s2} x_{s2,k} + \lambda_{s3} x_{s3,k} + \beta_d x_{ij}$

Replacing the individual household distance to dam x_{ij} with the averaged village distance $\overline{x}_{i.}$, leads to

$$h_{ij}(t) = \sum_{k=1}^{6} \exp(\lambda_{ik} + \zeta_i) I(\tau_{k-1} < t \leq \tau_k)$$

with

$$\lambda_{ik} = \lambda_0 + \lambda_y x_{yk} + \lambda_{s2} x_{s2,k} + \lambda_{s3} x_{s3,k} + \beta \overline{x}_{i.}$$

The cumulative hazard for $\tau_{k-1} < t \leq \tau_k$ is given by

$$H_{ij}(t) = (\tau_1 - \tau_0) \exp(\lambda_{i1} + \zeta_i) + (\tau_2 - \tau_1) \exp(\lambda_{i2} + \zeta_i) + \cdots + (t - \tau_{k-1}) \exp(\lambda_{ik} + \zeta_i)$$

which is equivalent to

$$H_{ij}(t) = \sum_{k=1}^{6} I(t > \tau_{k-1})(min(\tau_k, t) - \tau_{k-1}) \exp(\lambda_{ik} + \zeta_i)$$

Therefore

$$\log S_{ij}(t) = -H_{ij}(t) = -\sum_{k=1}^{6} I(t > \tau_{k-1})(min(\tau_k, t) - \tau_{k-1}) \exp(\lambda_{ik} + \zeta_i)$$

The conditional loglikelihood contribution of the j^{th} child in the i^{th} village is given by

$$\ell s_{ij} = \delta_{ij} \log(h_{ij}(y_{ij})) + \log(S_{ij}(y_{ij}))$$

hence

$$\ell s_{ij} = \sum_{k=1}^{6} \delta_{ijk}(\lambda_{ik} + \zeta_i) - I(y_{ij} > \tau_{k-1})(min(\tau_k, y_{ij}) - \tau_{k-1}) \exp(\lambda_{ik} + \zeta_i)$$

with δ_{ijk} denoting whether an event takes place, $(\delta_{ijk} = 1)$, or not, $(\delta_{ijk} = 0)$, in period k for the particular child.

Summing over all children in the village and splitting up the sum over the six different periods, we obtain

$$\ell s_{ik} = d_{ik}(\lambda_{ik} + \zeta_i) - a_{ik} \exp(\lambda_{ik} + \zeta_i)$$

The mixed Poisson loglikelihood contribution for the i^{th} village in the k^{th} period is derived as follows. Assuming the aggregated counts d_{ik} are independent observations with $d_{ik} \sim Poisson(\xi_{ik})$ then the loglikelihood contribution is given by;

$$\ell p_{ik} = d_{ik} \log(\xi_{ik}) - \xi_{ik} - \log(d_{ik}!)$$

Dropping the last term which does not contribute to the likelihood and replacing ξ_{ik} as described for the conditional mixed Poisson regression model, we have;

$$\ell p_{ik} = d_{ik} \log(a_{ik}) + d_{ik}(\eta_{ik} + \zeta_i) - a_{ik} \exp(\eta_{ik} + \zeta_i)$$

This expression is equivalent to the loglikelihood obtained for the frailty model except for the term $d_{ik} \log(a_{ik})$, which is a constant depending on the data and not on the parameters, so it can be ignored from the point of view of estimation hence the parameter estimates will thus be exactly the same.

2.4 Simulations

Simulation studies present an important statistical tool to investigate the performance and estimation techniques in comparing statistical models in pre-specified situations (Bender et al., 2005). Since the difference between the mixed Poisson regression model and the frailty model is in the use of the aggregated and individual risk factors, we aim to study the effect of the averaged and individual risk factors (distance effect) on the coverage and power based on different parameter values of the risk factor. Based on the parameter estimates obtained from the frailty model and altering the risk parameter values, we will generate event times from a piecewise constant hazards distribution. In generating the event times the dependence between children in the same village will be captured to reflect the real situation as depicted in the study design. Frailty terms representing each village are randomly sampled from a normal distribution with mean zero and variance of the random effects in the frailty model θ as estimated from the data. Using the individual household distance as observed in the malaria

incidence study as risk factors, a total of 5000 data sets for each parameter value of the risk factor will be generated.

A detailed discussion on the generation of event times can be found in Walke (2010) and Bender et al., (2005). The generated event times will be classified into each of the six periods depending on the season of infection and those exceeding the study duration considered as censored observations. At the end of simulation exercise we will have a complete dataset with information on the survival time, censoring indicator and distance. For the count regression modelling, the events are aggregated as per each season of infection in each village. The two statistical methodologies discussed previously will be compared with respect to their efficiency by use of the coverage probability and power using the 5000 generated data sets. The coverage probabilities are given by the number of times the true parameter is contained in the 95% confidence interval divided by the number of evaluated data sets. For both models, results will be summarized by median, standard error, 5^{th} and 95^{th} quantile and the 95% coverage probability for the risk factor, β_d. The power of the statistical test is the probability that the test will reject the null hypothesis when the null hypothesis is false, hence in our simulation study it will be represented by the percentage of data sets that do not include the true risk parameter in the 95% confidence interval.

3 Results

3.1 Exploratory Data Analysis

To get insights into the dataset, summary statistics and graphs will be used. Distance to the dam as the main risk factor under investigation was used to classify the households as either, at risk or control based on the flying ability of mosquitoes. The mean distance of the sampled households from the dam was 2.53 km with a standard deviation of 2.03 km, minimum and maximum distance from the dam of the sampled households were 0.055 and 9.046 km respectively. About 73.48% children resided in households which were classified as at risk. The minimum and maximum observed time-to-first malaria event in this study were 7 and 698 days respectively.

During the follow up period, 548 children had at least one malaria event. The Kaplan-Meier survival function for the control and at risk group is displayed in Figure 2. We observe that the estimated survival for the children in control households is higher than those at risk households over the entire follow up period, giving evidence for lower risk of malaria infection in children living at a distance to the dam as compared to those children living in a close proximity to the dam.

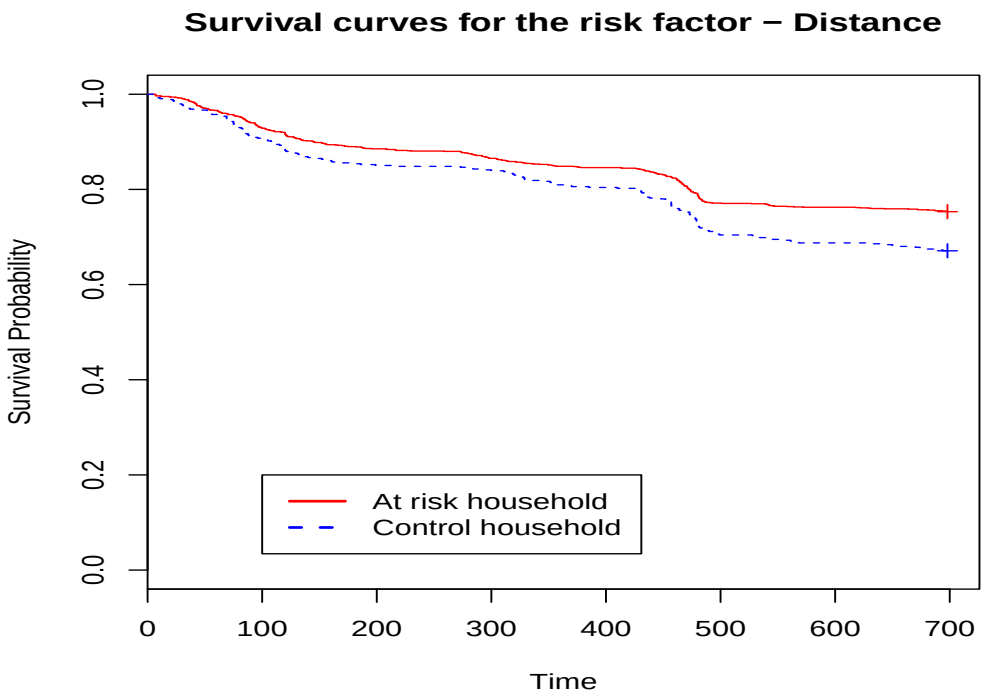

Figure 2: *Survival curves for the risk and control groups*

3.2 Statistical Analysis

3.2.1 Marginal Models

The marginal model parameter estimates are obtained by ignoring the dependence structure in the data. The event times were classified into each of the 6 different season depending on the season of infection. In survival modelling the season effect was captured by assuming a constant risk to malaria infection in each of the seasons and the distance effect captured by introducing the individual household distance to the dam as a continuous fixed effect covariate. The risk of malaria is found to be related to the season of infection. The sandwich estimators were used to obtain consistent estimates of the covariance matrix by using the grouped jackknife technique. Marginal model results for the two statistical models are as presented in Table 1.

Table 1: *Marginal models.*

	Survival model			Poisson model		
Parameter	Estimate	Std Error	P-value	Estimate	Std Error	P-value
Season1	-7.2690	0.1523	$< .0001$	-7.3025	0.1637	$< .0001$
Season2	-1.5432	0.1831	$< .0001$	-1.5427	0.1830	$< .0001$
Season3	-0.8269	0.1032	$< .0001$	-0.8264	0.1037	$< .0001$
Year2	-0.1925	0.0802	0.0164	-0.1914	0.0805	0.0175
Distance	0.0564	0.0561	0.3148	0.0687	0.0591	0.2450

On the other hand, a marginal count regression model was fitted for the aggregated event times ignoring the dependence between the periods within a village (use of independence working correlation). We used the mean distance to the dam since we aggregated the event times per village in each period. We observe similar parameter estimates as obtained for the survival model and the results show a significant seasonal effect. We observe from the marginal model results that the risk to first malaria event is high in the first season (long rain season) and decreases as we move to the dry season and the relative risk to first incidence of malaria is high for households close to dams with respect to the baseline seasonal effect.

3.2.2 Conditional Models

Due to the dependence of children within the same village, a log-normal frailty model with village as frailty and constant baseline hazard for each season was fitted. Accounting for the unobserved heterogeneity (village effect) led to insignificant distance effect and change of sign of the parameter estimate as compared to the marginal hazards model. The association parameter is significant hence reflecting dependence of subjects within the same cluster. On the other hand, the conditional mixed Poisson regression model was fitted with village as random effect and we note that the parameter estimates of the two models are the same. From Table 2 below we observe a significant period effect between aggregated event counts within a cluster. It can be observed that there is no significant distance effect in any of the two conditional models, however, it can be noted that the distance effect changes direction from one model to the other. The frailty model predict a decreasing malaria incidence with increasing distance to the dam

which might be attributed to the use of individual household distance. In the estimation of the random effects, Figure 4 in the Appendix show a positive relationship between the best linear unbiased random effects predictors obtained from the frailty model, ζ_i, and the distance to the dam.

Table 2: *Conditional models.*

	Frailty model			Conditional Mixed Poisson model		
Parameter	Estimate	Std Error	P-value	Estimate	Std Error	P-value
Season1	-7.1677	0.2147	$< .0001$	-7.3893	0.2443	$< .0001$
Season2	-1.5242	0.1356	$< .0001$	-1.5244	0.1356	$< .0001$
Season3	-0.8024	0.1192	$< .0001$	-0.8027	0.1192	$< .0001$
Year2	-0.1538	0.0876	0.0995	-0.1541	0.0876	0.0988
Distance	-0.0344	0.0570	0.5549	0.0539	0.0743	0.4797
Random effect	0.3486	0.1478	0.0323	0.3100	0.1290	0.0282

3.3 Simulation Results

Assuming different values for the risk factor and using the parameter estimates of the frailty model, the event times were generated from a piecewise constant hazard function. The generated data sets were fitted using the frailty model and the conditional mixed Poisson regression models as described in the methodology section. To investigate the coverage probabilities, we varied the risk factor parameter value and calculated the number of times the true parameter value was contained in the 95% confidence interval. For example, assuming a risk factor of $\beta_d = 0.05$, the estimated average of the overall median estimate of the risk factor (distance effect) for 5000 generated data sets was 0.0462 and 0.0491 for the fitted conditional mixed Poisson regression and frailty models respectively. The number of generated data sets that contained the true parameter value of 0.05 in the 95% confidence interval were 4482 and 4589 for the models fitted using the mixed Poisson regression and frailty models, translating to 0.8964 and 0.9178 coverage probabilities respectively. Other coverage probabilities are as presented in Table 3. We observe that there is a slight difference in the coverage between the two models which might be attributed to the information loss due to the event aggregation and use of the mean risk factor (distance) as compared to the use of individual household distances.

Table 3: *Simulation results comparing the mixed Poisson regression and the Frailty models assuming different risk factor values for the distance*

Parameter value	Model	Median (P5,P95)	Std error (P5,P95)	Coverage(%)
0	Poisson	-0.0044(-0.1166,0.0777)	0.0452(0.0318,0.0606)	90.06
	Survival	-0.0005(-0.0695,0.0702)	0.0397(0.0297,0.0485)	92.16
0.05	Poisson	0.0462(-0.0322,0.1225)	0.0443(0.0304,0.0600)	89.64
	Survival	0.0491(-0.0200,0.1155)	0.0382(0.0287,0.0470)	91.78
0.1	Poisson	0.0945(0.0112,0.1729)	0.0434(0.0296,0.0591)	88.16
	Survival	0.0986(0.0322,0.1658)	0.0374(0.0273,0.0461)	91.88
0.15	Poisson	0.1434(-1.4192,0.2231)	0.0428(0.0294,0.0584)	86.94
	Survival	0.1497(0.0845,0.2136)	0.0366(0.0270,0.0448)	92.38
0.2	Poisson	0.1813(0.118,0.2762)	0.0426(0.0283,0.0573)	88.94
	Survival	0.1955(0.1376,0.2634)	0.0361(0.0270,0.0439)	92.36

Through simulation we were able to compare the efficiency of the two statistical methodologies in statistical testing. The power represents the percentage of data sets which do not include the true parameter in the 95% confidence interval. From the simulation results on power we observe that the use of aggregated events as compared to the individual risk factor leads to a reduced power. However, the power reduction is small as depicted in the Figure 3 below.

Figure 3: *Power of the statistical models testing the effect of the distance to the dam under different risk factors value*

4 Discussion and Conclusion

Malaria has posed as a major health challenge in most of the developing countries leading to reduced production activities hence continued poverty in many countries. Public health expenditure in most developing countries has been crippled due to the high costs in the provision of health services geared towards prevention and treatment of malaria hence overdependence on donor funding. Seasonal variations, and development of water projects and their operation have some long history of facilitating increased transmission of vector borne diseases. Earlier studies have shown that malaria have a strong seasonal pattern with a lag time varying from a few weeks at the onset of the rainy season to more than a month at the end of the rainy season. Reliable analysis of the climatic and environmental risks to health is therefore fundamental for the prevention and control of malaria.

In this study we investigated the relation between seasonal variations and distance to the dam, to the time to the first malaria event of children living in close proximity to the Gilgel-Gibe hydroelectric power dam using the piecewise exponential proportional hazards model and its equivalent count regression model. 548 events (Only 26% of the children experienced the event during the follow-up period) were observed over the entire study period with a minimum and maximum event times of 7 and 698 days respectively. Most of the events were recorded during the rainy seasons and in children living in a close proximity to the dam shores. Some of the villages did not record any events in some periods especially during the dry season whereas in other periods had high number of events recorded, hence a more accurate technique had to be used to capture all the information. For the survival modelling the individual household distances were used whereas for the count regression averaged village distances were used since the events were aggregated within a village.

The malaria incidence data was analyzed by using both the marginal and conditional models. By ignoring the dependence structure in the data, consistent parameter estimates of seasonal and distance effects were obtained under the two statistical techniques. The standard error were estimated by their robust sandwich estimators by using the grouped jackknife technique

for the hazards model and assuming an independent working correlation assumption for the marginal mixed Poisson model. The risk to malaria event is high in the first season (long rain season) and decreases during the dry season. The relative risk to first incidence of malaria is high for households close to dams with respect to the baseline seasonal effect. To cater for the dependence in the data, conditional models were fitted by assuming a normally distributed random effect with zero mean and variance, ζ, which was estimated from the data to capture the unexplained variability in the marginal models. For comparability with the conditional mixed Poisson regression model a shared log-normal frailty was considered for the frailty model. Since no explicit form of it's marginal likelihood exists we used a numerical integration of the normally distributed random effects based on the Gaussian quadrature (nlmixed procedure in SAS).

Although distance to the dam effect was found to be insignificant in all the four models, it can be observed that three of the four models predict an increasing incidence with increasing distance to the dam. The frailty model predicts a decreasing incidence with increasing distance to the dam, which might be attributed to the use of individual distance for the frailty model as compared to the use of averaged distance for the conditional mixed Poisson regression model. Both models show a positive relationship between the best linear unbiased predictors (BLUPS) of the random effect and the averaged distances from the dam shore.

Simulation studies carried out to check on the adequacy of the two statistical techniques showed that the coverage probabilities of the true parameter in the simulated dataset were almost similar though, a slightly lower coverage for the mixed Poisson regression which might be attributed to the aggregation of event counts in the different periods. On the other hand, the power of the statistical tests was estimated by the number of the data sets that did not include the true parameter, based on the results obtained mixed Poisson regression can serve as a good alternative for the frailty model, though with a small power loss.

4.1 Recommendation and Limitation

The economic benefits generated by Gilgel-Gibe hydroelectric dam can be maximized if preventive programmes against malaria and other related vector-borne diseases can be implemented in the households close proximity to the dam reservoir. Health package programmes including the use of bed nets, health education especially at the onset of rains and environmental management (clearing of bushes, draining stagnant water, etc) should be implemented in an integrated way and strengthened to reduce disease burden in the population. If possible future dam projects should be undertaken in the highland areas since they are less prone to malaria. The major limitation in this study was in explaining the correlation between children within clusters where we used the shared frailty which forces the unobserved factors to be the same within the cluster which may not always reflect reality.

Reference

Adugna A. (2011). Malaria in Ethiopia: Ethiopian Demography and Health. http://www.ethiodemographyandhealth.org/MedVectoredDiseasesMalaria.pdf Accessed July 20, 2013.

Agresti, A. (2002). *Categorical Data Analysis*. 2nd ed. John Wiley & Sons, Inc., New Jersey.

Bender,R., Augustin,T. and Blettner, M. (2005). Generating Survival Times to Simulate Cox Proportional Hazards Models. *Stat. Med.* **24(11)**, 1713-1723.

Breslow, N.E. and Clayton, D.G. (1993). Approximating inference in generalized linear mixed models.*Journal of the American Statistical Association* **88**, 9-25.

Clayton, D.G. (1978). A model for association in bivariate life tables and its application in epidemiological studies of familial tendency in chronic disease incidence. *Biometrika* **65,** 141-51.

Collett, D. (1994). *Modeling Survival Data in Medical Research*, London: Chapman and Hall.

Duchateau, L. and Janssen, P. (2008). *The Frailty Model*. New York: Springer.

Duchateau, L., Janssen, P., Lindsey,P., Legrand, C., Nguti, R., and Sylvester, R. (2002). The shared frailty model and the power for heterogeneity tests in multicenter clinical trials. *Computational Statistics and Data Analysis* **40**,603-620.

Duchateau, L. and Janssen, P. (2005). Understanding heterogeneity in generalized mixed and frailty models. *American Statistician* **59**, 143-146.

Gabriel, S. and James, V. (2005). Developing malaria early warning system in Ethiopia. 25th Annual ESRI International User Conference. Paper No. UC2409. San Diego.

Getachew, Y., Janssen, P., Yewhalaw, D., Speybroeck, N. and Duchateau, L. (2013). Coping with time and space in modelling malaria incidence: a comparison of survival and count regression models. *Statistics in Medicine* **32**, 3224-3233.

Ghebreyesus, T. A., Haile, M., Witten, K. H., Getachew, A., Yohannes, A. M., Yohannes, M., Teklehaimanot, H. D., Lindsay, S. W. and Byass, P. (1999). Incidence of malaria among children living near dams in northern Ethiopia: community based incidence survey. *BMJ,* **319**, 663-666.

Henderson, R., and Oman, P.(1999). Effect of Frailty on Marginal Regression Estimates in Survival analysis. *Journal of the Royal Statistical Society, Series B* **61**, 367-379

Hougaard, P. (2000). *Analysis of Multivariate Survival Data.* Springer Verlag, New York.

Ismail, N. and Jemain, A. (2007). Handling overdispersion with negative binomial and generalized Poisson regression models. *Casualty Actuarial Society Forum* **12**, 103-158.

Karunamoorthi, K., and Bekele, M. (2012). Changes in Malaria Indices in an Ethiopian Health Centre: A Five Year Retrospective Analysis. *Journal of Health Scope* **1**(3), 118-126

Laird, N. and Oliver, D. (1981). Covariance Analysis of censored survival data using log-linear analysis techniques.*Journal of the American Statistical Association* **76**, 231-240.

Lawless, J.F. and Zhan, M. (1998). Analysis of interval-grouped recurrent event data using piecewise constant rate functions. *Canadian Journal of Statistics* **26**, 549-565

Lee, Y. and Nelder, J.A. (1996). Hierarchical generalized models (with discussion). *Journal of the Royal Statistical Society. Series B (Methodological)* **58**, 619-678.

Lin, D.Y. (1994). Cox Regression analysis of multivariate failure time data: the marginal approach. *Statistics in medicine* **13**, 2233-2247

Ma, R. J., Krewsi, D., and Burnett, R.T. (2003). Random effects Cox models: A Poisson modelling Approach*Biometrika* **90**, 157-169

McCullagh, P., and Nelder, J.A. (1989). Generalized linear models. Chapman & Hall, London

McGilchrist, C.A. (1993). REML estimation for survival models with frailty. *Biometrics* **49**, 221-225.

McGilchrist, C.A. and Aisbett, C.W. (1991). Regression with frailty in survival analysis. *Biometrics* **47**, 461-466.

Oakes, D. (1989). Bivariate survival models induced by frailties. *Journal of the American Statistical Association* **84**, 487-493.

Service, M. W. (1991). Agricultural development and arthropod-borne diseases: a review. *Rev Saude Publica.* **25**, 165-178.

Shukla, R. P., Sharma, S. N., Kohli, V. K., Nanda, N., Sharma, V. P. and Subbarao, S. K. (2001). Dynamics of malaria transmission under changing ecological scenario in and around Nanak Matta Dam, Uttaranchal, India. *Indian J Malariol.* **38**, 91-98.

Steinmann, P., Keiser, J., Bos, R., Tanner, M., and Utzinger, J. (2006). Schistosomiasis and water resources development: systematic review, meta-analysis, and estimates of people at risk. *The Lancet infectious diseases* **6**(7), 411-425.

Therneau, T.M. and Grambsch, P.M. (2000). *Modelling Survival data. Extending the Cox model.* New York: Springer.

Thomson, M. C., Connor, S. J., Quinones, M. L., Jawara, M., Todd, J. and Greenwood, B.M. (1995). Movement of *Anopheles gambiae s.l.* malaria vectors between villages in The Gambia. *Med Vet Entomol.* **9**, 413-419

Tulu, A. G. (1993). *The ecology of Health and Disease in Ethiopia.* Boulder, San francisco, Oxford, Westview Press, Inc.: 341-352.

Walke, R. (2010). Example for a Piecewise Constant Hazard Data Simulation in R. Max Planck Institute for Demographic Research.

Wienke, A (2010). *Frailty Models in Survival Analysis.* Chapman & Hall/CRC biostatistics series. Taylor and Francis.

White, G. B. (1974). Anopheles gambiae complex and disease transmission in Africa. *Trans-*

actions of the Royal Society of Tropical Medicine and hygiene **68**, 278-301.

WHO (2010). World Health Organization. *World Malaria Report.*

WHO (2001). Malaria early warning systems: concepts, indicators and partners. A framework for field research in Africa.

Yewhalaw, D., Getachew, Y., Tushune, K., Kassahun, W., Duchateau, L., and Speybroeck, N. (2013). The effect of dams and seasons on malaria incidence and anopheles abundance in Ethiopia. *BMC infectious diseases* **13**(1), 1-9.

Yewhalaw, D., Legesse, W., Van Bortel, W., Gebre-Selassie, S., Kloos, H., Duchateau, L., and Speybroeck, N. (2009). Malaria and water resource development: the case of Gilgel-Gibe hydroelectric dam in Ethiopia. *Malar J.* **8**, 1-10.

Appendix

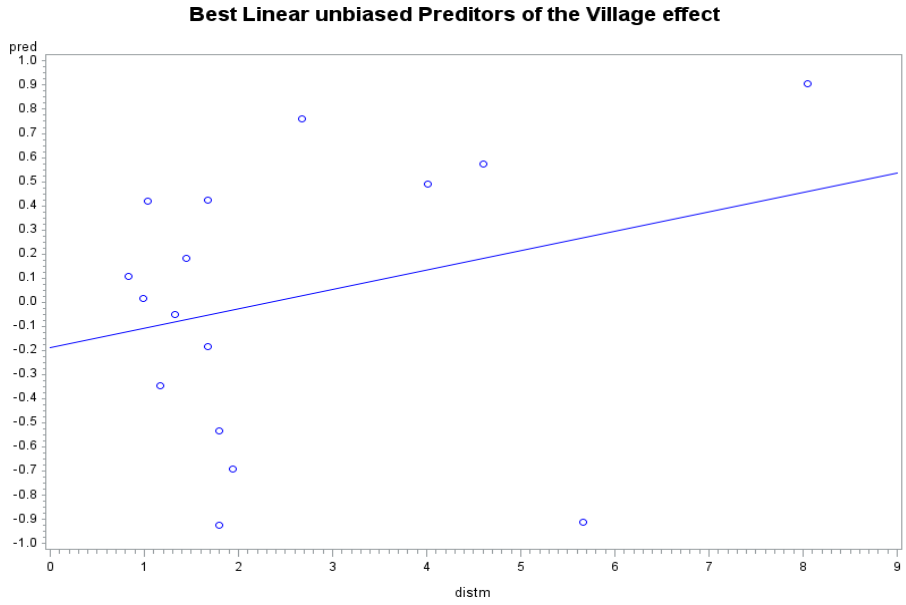

Figure 4: *Best Linear unbiased predictors (BLUPS) of the random effect as a function of the distance to the dam*

Auteursrechtelijke overeenkomst

Ik/wij verlenen het wereldwijde auteursrecht voor de ingediende eindverhandeling:
Modelling of Malaria Incidence data: Comparison of Shared frailty Model and Count Regression model

Richting: **Master of Statistics-Biostatistics**
Jaar: **2013**

in alle mogelijke mediaformaten, - bestaande en in de toekomst te ontwikkelen - , aan de Universiteit Hasselt.

Niet tegenstaand deze toekenning van het auteursrecht aan de Universiteit Hasselt behoud ik als auteur het recht om de eindverhandeling, - in zijn geheel of gedeeltelijk -, vrij te reproduceren, (her)publiceren of distribueren zonder de toelating te moeten verkrijgen van de Universiteit Hasselt.

Ik bevestig dat de eindverhandeling mijn origineel werk is, en dat ik het recht heb om de rechten te verlenen die in deze overeenkomst worden beschreven. Ik verklaar tevens dat de eindverhandeling, naar mijn weten, het auteursrecht van anderen niet overtreedt.

Ik verklaar tevens dat ik voor het materiaal in de eindverhandeling dat beschermd wordt door het auteursrecht, de nodige toelatingen heb verkregen zodat ik deze ook aan de Universiteit Hasselt kan overdragen en dat dit duidelijk in de tekst en inhoud van de eindverhandeling werd genotificeerd.

Universiteit Hasselt zal mij als auteur(s) van de eindverhandeling identificeren en zal geen wijzigingen aanbrengen aan de eindverhandeling, uitgezonderd deze toegelaten door deze overeenkomst.

Voor akkoord,

Sila, Alex

Datum: **11/09/2013**