# Modelling diapause termination of *Rhipicephalus appendiculatus* using statistical tools to detect sudden behavioural changes and time dependencies.

**Speybroeck N.[1*], Lindsey P.J.[2], Billiouw M.[1], Madder M.[1],**
**Lindsey J.K.[3], and Berkvens D.L.[1].**

**E-mail: nspeybroeck@itg.be**

**E-mail: patrick.lindsey@gen.unimaas.nl**

**E-mail: mbilliouw@zamnet.zm**

**E-mail: mmadder@itg.be**

**E-mail: james.lindsey@luc.ac.be**

**E-mail: dberkvens@itg.be**

[1] Institute for Tropical Medicine, Nationalestraat 155, 2000 Antwerp, Belgium.

* Author for corresponding - Telephone: + 32 3 247 62 73.
Email: nspeybroeck@itg.be.

[2] University of Maastricht, Department of population genetics, UNS 50 , postvak 16, Postbus 616, 6200 MD Maastricht, The Netherlands.

[3] Centre for Statistics, Limburgs Universitair Centrum, Universitaire Campus, 3590 Diepenbeek, Belgium (Fax: 043664751).

## ABSTRACT

This paper presents statistical methodology to analyse longitudinal binary responses for which a sudden change in the response occurs in time. Cumulative probability plots, transition matrices, and change-point models and more advanced techniques such as generalized auto-regression models and hidden Markov chains are presented and applied on a study on the activity of *Rhipicephalus appendiculatus*, the major vector of *Theileria parva*, a fatal disease in cattle. This study presents individual measurements on female *R. appendiculatus,* which are terminating their diapause (resting status) and become active. Comprehending activity patterns is very important to better understand the ecology of *R. appendiculatus*. The model indicates that activity and non-activity act in an absorbing way. The change-point model estimates that the sudden change in activity happens on December 10. The reaction of ticks on acceleration and changes in rainfall and temperature indicates that ticks can sense climatic changes. The study revealed the underlying not visually observable states during diapause development of the adult tick of *R. appendiculatus.* These states could be related to phases during the dynamic event of diapause development and post-diapause activity in *R. appendiculatus.*

**Keywords**: behaviour, ecology, diapause, generalized auto-regression models, hidden Markov chains, *Rhipicephalus appendiculatus*, ticks.

# 1. INTRODUCTION

In ecology, an organism often remains in a resting phase for a certain period to synchronise its life cycle with fitting climatic conditions. The diapausing behaviour of *Rhipicephalus appendiculatus* vector of *Theileria parva,* the causative agent of the bovine disease East Coast fever is an example. This tick goes in diapause, to survive dry periods in southern Africa. *R. appendiculatus* uses the shortening daylenghts at the beginning of the dry season (in April-May) as an indication to enter diapause (Madder et al., 2002). At a certain point in time, the organism becomes active again in order to continue its life cycle. In southern Africa, this occurs often suddenly at the end of the dry season and the beginning of the rainy season in November-December (Speybroeck et al., 2002). By not taking into account this sudden change of activity in time in an analysis of the activity patterns, the influence of other explanatory variables, which are measured, might be misinterpreted.

In Zambia, numbers of *R. appendiculatus* adults on the hosts increase after the onset of the rains (MacLeod, 1970; Pegram *et al.*, 1986; Pegram and Banda, 1990; Berkvens *et al.*, 1995; Berkvens *et al.*, 1998, Speybroeck *et al.*, 2002), probably to ensure that the most vulnerable stages of the lifecycle, namely eggs and larvae, are exposed to favourable conditions. The main factor responsible for becoming active again is thought to be day length where a long photoperiod terminates the state of diapause and induces host seeking in the wet months (Short *et al.*, 1989b; Pegram and Banda, 1990). Berkvens *et al.* (1995) reported that in eastern Zambia the diapause of *R. appendiculatus* was terminated after the onset of the rains, apparently not after a long day signal, but due to weakening photoperiodic maintenance of the diapause because of an increased age of the ticks. This could also be demonstrated in laboratory conditions (Madder *et al.*, 2002).

Understanding activity patterns of adult ticks at the beginning of the rainy season is important for epidemiological reasons.

The aim of this paper is to study tick activity under quasi-natural conditions during the 1986-87 rainy seasons. The interest lies in presenting and applying methodology for detecting a sudden change in the behaviour of an organism and in the investigation of the remaining importance of explanatory variables after having accounted for the change in behaviour and for the dependency of measurements on the same subject in time.

First we present explorative techniques like cumulative probability plots, transition matrices, and change-point models. The benchmark is a change-point model, which does not use covariates but simply tries to locate the most likely point at which behaviour changes.

Lindsey (2001) fitted a change point model for Poisson distributed responses and indicated that hidden Markov chains could have been used to detect the change point.

We then present and apply more advanced techniques such as generalized auto-regression models and hidden Markov chains.

Guttorp (1995) and MacDonald and Zuchini (1997) provide mathematical background on the topic of hidden Markov chains. The theory and applications of generalised auto-regression models have been thoroughly discussed by Lindsey (1999). The term 'generalised' is used in this context to indicate the possibility of choosing distributions other than the Gaussian one to model the response.

## 2. DESIGN OF THE TICK ACTIVITY STUDY

The aim of the trial at hand was to investigate activity patterns of female *R. appendiculatus* ticks from October, just before the start of the rainy season until the end of March at the end of the rainy season.

The experiment was carried out in eastern Zambia. On the 11th of July 1986, newly moulted adults of four local *R. appendiculatus* strains were released into circular gauze columns of 5cm diameter and 1m tall. The bottom end of each column was glued to an open cylinder, which was secured in the soil to a depth of about 10 cm. This allowed ticks access to the soil. The top of the column was tied to a metal wire frame. The columns were placed under a cover. Coloured numbered tags used by beekeepers to identify queens were applied to the ticks, allowing individual identification. The following stocks of ticks originating from eastern Zambia were used in this trial (Berkvens *et al.*, 1995):

- *R. appendiculatus* (Michembo), collected at Michembo (altitude 1040 m).
- *R. appendiculatus* (Genda), collected at Genda (altitude 1150 m).
- *R. appendiculatus* (Nkolowondo), collected at Nkolowondo (altitude 1080 m).
- *R. appendiculatus* (Lundazi), collected at Lundazi (altitude 1250 m).

There are differences among these stocks. Ticks from Michembo have only recently appeared in eastern Zambia and are still settling, whereas ticks from the Genda and Nkolowondo locations have been around for a long time. On the other hand, Lundazi is located closer to the Equator and has a longer rainy season and a smaller difference in day length.

Data were collected on the following numbers of females: 7 from Michembo, 8 from Genda, 9 from Nkolowondo, and 9 from Lundazi.

Each tick was observed daily around 10 in the morning between October 1, 1986 and March 31, 1987. They were recorded as either absent from the column, present but inactive, or present and active. A tick was considered to be active when it reacted to the presence of the recorder by actually moving in the column, usually upwards. When ticks were unobserved for more than 7 consecutive days, they were considered dead.

Besides the tick's origin (four localities), several other explanatory variables are measured on a daily basis. Ambient temperature (average temperature in °C) and relative humidity (in %) were recorded daily at the study site. The rainfall (in mm), minimum and maximum temperature of the day (in °C) were recorded, and day length (expressed as potential hours of sunshine in hours) was calculated according to a formula of Duffett-Smith (1985) based on List (1951). List (1951) defines daylength as the interval between sunrise and sunset. According to List (1951), sunrise and sunset occur when the upper edge of the solar disk appears to be exactly on the horizon, i.e. on average when the centre of the sun is 50' (' = degree minutes) below the horizon. Thus, daylength is calculated as the interval between these two events, i.e. between (i) centre of sun is 50' below the horizon in the morning and (ii) centre of sun is 50' below the horizon in the evening.

Some of these variables were also used to construct additional explanatory variables such as the average daily vapour pressure deficits (in mmHg), some change-point indicators, and various cumulative and lagged variables. The daily average vapour pressure deficits (*vpd*) was calculated according to Rosenberg *et al.* (1983) as a nonlinear function of temperature and humidity. Vapour pressure deficit is the difference between the Saturation Vapour Pressure (the pressure that water vapour molecules would exert if the air were saturated with vapour at a given temperature) and the actual vapour pressure.

Two indicators were also created: a rain and a day length change-point variable respectively indicating by zeroes the period before and by ones the period after the highest rainfall peak (which occurred on December 11 or day 69) and the longest day (which occurred on December 21 or day 79).

Lagged variables are also used in order to take into account dependence on previous observed values. Lagged variables of the response up to three previous values were constructed, whereas for explanatory meteorological variables, only lagged variables up to two previous values were created. From a biological point of view, it is sensible to consider such lagged variables because the short term history of each particular individual most probably influences its current and further behaviour.

Only the inactive and active responses collected are considered in this paper. Few observations (mainly at the end of an individual's observation series) were actually collected in the absent or unobserved category. Towards the end of the study, an individual observed in this category could be hiding or dead. Thus, when a tick was unobserved, the observation was considered to be missing (at random).

It can also be remarked that we treated the columns with the ticks as identical. However, ticks from each of the four *R. appendiculatus* strains were released into four columns, one column for each strain. This design confounds "column" effects with the effects of strain. Although we do not think that this confounding distracts from this paper, a randomized complete block design where every strain is placed in every column or, because it is difficult for strains to cohabitate, a design with more than one column for each strain would have been more appropriate.

## 3. VISUALIZATION

The probability of a tick being observed inactive is represented by the height of the continuous line in Figure 1 a. The probability of a tick being inactive is very high at the start of the study. The probability of inactivity decreases at some point in time and remains lower until the end. The vertical dotted and dashed-dotted lines respectively represent the maximum rainfall peak (day 69) and the longest day (day 82). The other graphs in Figure 1 show recorded day lengths, rainfall, relative humidity, vapour pressure deficit, and temperature. These are all related: a change in their patterns occurs around the maximum rainfall peak day.

[Insert **Figure 1**].

### 3.1. Transition Matrices

For a first-order two-state Markov chain the transition matrix is calculated to be

$$\begin{pmatrix} p_{0|0} & p_{1|0} \\ p_{0|1} & p_{1|1} \end{pmatrix} = \begin{pmatrix} 0.94 & 0.06 \\ 0.35 & 0.65 \end{pmatrix}$$

with for example $p_{0|1}$ the conditional probability of no event following an event.

The conditional (transition) probability of no activity given that there was also no activity the previous day $p_{0|0}$ is 0.94, much higher than when there was activity the previous day (0.35). The marginal probability of being active regardless of the activity pattern of the previous day is 0.14. The transition matrix for a second-order two-state Markov chain is

$$\begin{pmatrix} p_{0|00} & p_{1|00} \\ p_{0|01} & p_{1|01} \\ p_{0|10} & p_{1|10} \\ p_{0|11} & p_{1|11} \end{pmatrix} = \begin{pmatrix} 0.95 & 0.05 \\ 0.75 & 0.25 \\ 0.48 & 0.52 \\ 0.28 & 0.72 \end{pmatrix}$$

The probability of activity is increasing respectively if there was no activity during the last two days, there was activity only two days ago, there was activity only on the previous day, and there was activity during both previous days.

### 3.2. Change-point Models

A change-point is an indicator variable, zero before a certain point in time and then one. The point in time can then be estimated by including it in the model. Hence, fitting such a model to binomial response data will allow a change in the average response at some unknown point in time. Figure 2a shows the corresponding negative normed likelihood for the binomial response. For our data, December 7 (day 69), was estimated as the optimal change-point. Figure 2b shows the fit of this model. Once the tick is active, a cyclic pattern of activity is visible. It is possible to estimate a change-point for every group separately. The 95% Confidence Intervals based on the assumption the –2 times the log-likelihood is distributed as chi-square with 1 degree of freedom. The estimates with the 95% Confidence Intervals are: ……………….

A change point could also be combined with a model, but this is not performed in this paper. Indeed, from this point on, we preferred to consider only change-point variables, which could biologically be interpreted such as the longest day or the start of the rainy season. It is important to distinguish between change-point variables, which are observed and change-point models, which are estimated.

[Insert **Figure 2**].

## 4. STATISTICAL BACKGROUND

Activity status on 33 ticks was collected daily over a period of six months resulting in 5249 observations. This provides enough degrees of freedom to be able to model complex environmental factors influencing the behaviour of the ticks.

One of the purposes of this article is to propose useful new techniques to analyse activity data where a sudden change of the behaviour of an organism occurs. The data are such that the behaviour of an individual can be related to the behaviour of that individual at other time points. Three different types of dependencies among observations on a subject in time were considered: state dependence, serial dependence, and spells (hidden Markov models). These all allow the subject's history to be taken into account by the model but in different ways. State dependence adjusts the model by taking the actual previous observation into consideration. In serial dependence, the difference between the model prediction at the previous time point and the actual previous observation is used to take the individuals history into account in the model. Finally, hidden Markov chains allow the undergoing biological process to switch over time among several hidden states (or behaviours), called spells.

### 4.1. Generalized Auto-regression Models

For observations collected at equally spaced times, a Markov process of order *M*, can be written as

$$m_t = a_0 + \sum_{h=1}^{M} r_h y_{t-h} \qquad (1)$$

with $m_t$ the location parameter of the distribution, $a_0$ the intercept, $r_h$ the Markov process parameter (constrained between 0 and 1) that quantifies the dependence strength on the previous response, and $y_{t-h}$ the *h*th previous response with respect to time *t* (Lindsey, 1999). Equation (1) can be generalised further by using an appropriate link function $g(m_t)$ to describe the location of the distribution (Lindsey, 1999). Models with this type of dependence can be referred to as state dependence models. First, second, and third order state dependence models are considered in this paper.

On the other hand, models where the location of the distribution does not depend directly on the previous response but on the difference between the previous response and its predictor

$$m_t = a_0 + \sum_{h=1}^{M} r_h \left(y_{t-h} - m_{t-h}\right)$$

can be referred to as serial dependence models (Lambert, 1996).

It is also important to assess whether the observations collected are following a stationary or non-stationary process. The biological process may not yet have reached equilibrium and is therefore still evolving suggesting that a non-stationary dependence process should be considered. The dependence structure will then have two parameters, measuring the dependence strength on just the previous residual but capturing a recursively fading dependence on all previously collected observations. When the first parameter $(j)$ tends to zero the dependence structures indicates that a stationary

Markov process, with parameter ( $r$ ) measuring the dependence strength on the previous residual, is more suitable.

Note that the above types of dependencies (state and serial) can be incorporated into a same model yielding a state and serial dependence model. The latter model is used in this paper.

Finally, the change in behavior of the ticks will only be introduced in the model by the longest day or the start of the rainy season change-point variables.

## 4.2. Hidden Markov Chains

Dependence among each subject's observations will now be induced differently for the case where the responses are generated in one of several different unknown states. Now, the dependence is induced conditional on the subject's previous hidden state history rather than its previous observed history. All possible changes of state over time must be taken into account. Each subject's possible "path" through the states corresponds to a product of conditional probabilities. Because the dependence is conditional on each subject's previous state history, it is calculated by multiplying the probability of a particular subject being in each state at each given time point by a transition probability. This ensures that summing all these products of conditional probabilities together produces the joint probability over time and all possible states for a particular subject. The sample probability is then obtained by multiplying these sums of products of conditional probabilities together over all individuals.

At the first time point, no previous information is available to estimate the probability of being in a particular state. This is solved by using the stationary marginal transition probabilities, which assumes that stationarity of the hidden process has been reached. Then $s(s–1)$ transition probabilities must be estimated along with the regression parameters.

Consider a simple case with two states and three time points. As above for ordinary Markov chains, the transition probabilities are $p_{1|1}$, $p_{1|2}$, $p_{2|1}$, and $p_{2|2}$ and the marginal probabilities are $d = \frac{p_{1|2}}{p_{2|1}+p_{1|2}}$ and $1-d$ respectively for state 1 and state 2 (MacDonald & Zucchini, 1997). The joint probability over time and all possible states for subject $i$ is then

$$
\begin{aligned}
&\Pr\left(Y_{i1}=k_1,\ldots,Y_{im}=k_m\right)= \\
&\quad d\ \left\{\Pr\left(Y_{i1}=k_1|S_{11}\right)\ \ p_{1|1}\left[\Pr\left(Y_{i2}=k_2|S_{21}\right)p_{1|1}\Pr\left(Y_{i3}=k_3|S_{31}\right)+p_{2|1}\Pr\left(Y_{i3}=k_3|S_{32}\right)\right]\right. \\
&\qquad\qquad \left.+p_{2|1}\left[\Pr\left(Y_{i2}=k_2|S_{22}\right)p_{1|2}\Pr\left(Y_{i3}=k_3|S_{31}\right)+p_{2|2}\Pr\left(\mathrm{Y}_{i3}=k_3|\mathrm{S}_{32}\right)\right]\right\} \\
&\quad+(1-d)\left\{\Pr\left(Y_{i1}=k_1|S_{12}\right)+p_{1|2}\left[\Pr\left(Y_{i2}=k_2|S_{21}\right)p_{1|1}\Pr\left(Y_{i3}=k_3|S_{31}\right)+p_{2|1}\Pr\left(Y_{i3}=k_3|S_{32}\right)\right]\right. \\
&\qquad\qquad \left.+p_{2|2}\left[\Pr\left(Y_{i2}=k_2|S_{22}\right)p_{1|2}\Pr\left(Y_{i3}=k_3|S_{31}\right)+p_{2|2}\Pr\left(Y_{i3}=k_3|S_{32}\right)\right]\right\}
\end{aligned}
$$

where $\Pr\left(Y_{ij}=k_j \mid S_{jh}\right)$ is the probability of the response being observed in category $k$ at time $j$ for subject $i$ given it is in state $(S)$ $h$ at time $j$. This expression is not computationally feasible but it can be rearranged in a recursive form over time (MacDonald & Zucchini, 1997; Lindsey, 1999). The joint probability over time and all possible states for subject $i$ can then be written as

$$\Pr\left(Y_{i1}=k_1,\ldots,Y_{im}=k_m\right)=\mathbf{d}^T\prod_{j=1}^{m}\left(\mathbf{p}D_{ijk_j}\right)\mathbf{J}^T \quad (2)$$

where **d** is a row vector containing the marginal probabilities, **p** is the transition matrix, $D_{ijk_j}$ is an $s\times s$ matrix containing on the diagonal the probabilities of the response being observed in category $k$ at time $j$ for subject $i$ given the various possible states, and **J** is a row vector of ones. The likelihood is obtained by multiplying Equation (2) over all subjects.

A hidden Markov chain can be illustrated for instance by considering ticks being active or not (a binary time series). The tick can now be in one of two unobservable states, a state where its metabolism is low and another state where it is high. There are no ways to measure this directly but a tick can behave differently (and shows a different activity outcome) depending on which state it is in. Each of the two possible events might be generated by one of the two Bernoulli distributions. The process switches from the one to the other according to the state of the hidden Markov chain, in this way generating dependence over time.

Two types of recursive probabilities can now be extracted from this model. These are obtained from the intermediate values

$$z_{ijr} = \sum_{o=1}^{s} z_{i,j-1,o} p_{or} \operatorname{Pr}(Y_{ij} = k_j \mid S_{jr})$$

calculated while constructing the joint probability over time and all possible states for subject *i* are required, where $z_{i1r} = d_r \operatorname{Pr}(Y_{i1} = k_1 \mid S_{1r})$ is obtained for the first time point.

A filtered probability is the probability that a specific subject is in a particular hidden state given this subject's previous state history. Hence, the probability of subject *i* being in state *r* at time *j* is $x_{ijr} = \frac{z_{ijr}}{\sum_{o=1}^{s} z_{ijo}}$, obtained by standardizing the ?$_{ijr}$. The probabilities of the response being observed in category *k* at time *j* for subject *i* can then be calculated by $j_{ij} = \sum_{o=1}^{s} x_{ijo} \operatorname{Pr}(Y_{ij} = k_j \mid S_{jo})$ which are the recursive probabilities for subject *i*.

Finally, changes in behavior of the ticks are introduced in the model by the longest day or the start of the rainy season change-point variables as well as by transitions between the different states. Although, multiple state transitions may occur.

# 5. ANALYSES OF TICK ACTIVITY

## 5.1. Modelling Strategy

Because the modelling process is exploratory, the inference criterion used for comparing the models under consideration is their ability to fit the observed data, that is how probable they make the data. In other words, models are compared directly through their minimized minus log likelihood. The models can be penalized by adding the number of estimated parameters, a form of the Akaike information criterion (AIC, see Akaike, 1973; Lindsey and Jones, 1998). Smaller values indicate more preferable models. This criterion allows direct comparisons among models that are not nested.

AICs are only comparable if they are calculated by fitting models based on the same (data and) number of observations. Hence, care must be taken when working with lagged variables. As lag(1), lag(2), and lag(3) response variables (previous, second previous, and third previous day activity status) are considered, all the exploratory methods and analyses included in this paper are based on the tick dataset with all observations recorded during the three first days removed. This allows all desired comparisons among results.

An intercept or null model provides a reference point for comparison with further fitted models. Models, each containing only one of the different explanatory variables, are then fitted and sorted in ascending order of their AIC value. The explanatory variable model with the lowest AIC is selected as the starting point for the model building process. Additional explanatory variables are then added to this model according to their AIC value and are only kept in the model if the AIC reduces. If the final model contains change-point variables, interactions between change-point variables and the explanatory variables present in the model are considered. This will provide additional information on differences in behaviour before and after the change-point. The last modelling step is then to check whether any explanatory variables are no longer significant and could therefore be deleted from the model. This is carried out by removing each variable and keeping it out of the model if the AIC decreases. The model must remain hierarchically valid and a non-significant main effect will not be removed from the model if this explanatory variable is still present in an interaction term.

For all models presented, a logit link was used.

## 5.2. Statistical Computations

The analyses presented in this paper are performed using packages in R (Ihaka and Gentleman, 1996). R is a fast S-Plus clone freely available (http://cran.r-project.org).

The generalised auto-regression models and the hidden Markov chains can be fitted respectively using the *gar* and *hidden* functions provided by the package *repeated* which can be downloaded from a web page (http://www.luc.ac.be/~jlindsey/rcode.html).

The functions to fit change point models, the code used to perform the analyses, and the tick data set can be obtained on another web site (http://euridice.tue.nl/~plindsey/).

# 6. GENERALISED AUTO-REGRESSION MODELS

The logistic regression only containing the intercept has an AIC of 2146. This is lowered to 1371.3 by adding a serial dependence. Adding state dependence on the previous three days activity status improves the model by lowering the AIC to 1327.4 (six-parameter model). After the addition of all the significantly contributing variables a seventeen-parameter state and serial dependence model with an AIC of 1290.1 was obtained (Deviance: 1209.126, degrees of freedom: 5212; McCullagh and Nelder, 1992, p.174)

$$
\begin{aligned}
\text{logit}(m) = \mathit{0.436} &+ 2.754\times\text{previous day activity status} + 0.474\times\text{second previous day activity status}\\
&+ 0.310\times\text{third previous day activity status} + 0.001\times\text{cumulative rainfall}\\
&+ \mathit{0.321}\times\text{longest day indicator} - 0.154\times\text{maximum temperature}\\
&+ 0.091\times\text{change in maximum temperature} - 0.055\times\text{acceleration in maximum temperature}\\
&+ 0.022\times\text{change in rainfall} - 0.012\times\text{acceleration in rainfall}\\
&- \mathit{0.0001}\times\text{vapour pressure deficit} - 0.669\times\text{Lundazi location}\\
&- 0.975\times\text{longest day indicator}\times\text{previous day observation}\\
&+ 0.140\times\text{longest day indicator}\times\text{vapour pressure deficit}
\end{aligned}
$$

where $\mu$ is the expected (or average) probability of being active and the three italised coefficients are not significantly different from zero. The dependence parameters for the non-stationary process are 0.916 ($j$) and 0.702 ($r$). The first parameter ($j$) clearly indicates that the biological process is still evolving over time and that the dependence structure can therefore not be simplified to a first-order Markov process.

The maximum temperature (-0.154 × maximum temperature) is the most important factor in the model obtained because the daily maximum temperature (which must be between 19 and 38°C) multiplied by its coefficient has the largest impact on the probability of the tick being active. An increase in maximum temperature corresponds to a decrease in the probability of a tick being active. Indeed, even if a tick has been active during the three previous consecutive days, this tick's current probability of being active will be below one half for (constant or increasing) maximum temperatures above 24.3°C.

The tick's activity history (2.754 × previous day activity status, 0.474 × second previous day activity status, 0.31 × third previous day activity status) is also important. Once a tick becomes active it will have a greater probability of remaining active. The change in amount of rainfall (0.022 × change in rainfall) respectively increases and decreases the probability of a tick being active each time a rainy period starts and ends. This factor will have the most impact on the probability of a tick being active when the rainy season starts and hence can be interpreted as one of the signals indicating to the tick to start becoming active.

Another one of the signals that could be indicating to the tick to start becoming active is the longest day (0.321 × longest day indicator) but in the model obtained the coefficient of the main effect is not significantly different from zero. Note that the AIC rises from 1290.1 to 1296.1 when the rainy season indicator replaces the longest day indicator. On the other hand after the longest day (December 21 or day 79), the effect of the previous day activity status decreases ([2.754-0.975] × previous day activity status) and the vapour pressure deficit ([0.14-0.0001] × vapour pressure deficit) proportionally affects the probability of a tick being active (due to the two interaction terms with the longest day indicator).

The cumulative rainfall (0.001 × cumulative rainfall) can be interpreted as a non-linear ageing effect, which slowly obliges the probability of the tick being active to increase over time regardless of any other factors.

The acceleration in rainfall (-0.012 × acceleration in rainfall) is the fourth most important factor, closely followed by the change and acceleration in maximum temperature. The acceleration is obtained by the following formula:

$$[(\text{today} - \text{previous day}) - (\text{previous day} - \text{second previous day})] = \text{today} - 2 \times \text{previous day} + \text{second previous day}$$

The model also distinguishes ticks from the Lundazi location (-0.669 × Lundazi-location), because the probability of these ticks being active remains lower throughout the entire study. From Figure 3, Individual tick 25 is a typical tick from Lundazi. It is only active for a very short period of time compared to ticks from other locations.

[Insert **Figure 3**].

The effect of these explanatory variables along with the ones of the state and serial dependence can also be seen from Figure 3. The underlying population curve indicated by the continuous line keeps slightly increasing (due to the cumulative rainfall) and fluctuating over time (due to the meteorological explanatory variables). The effect of adding a state and then a serial dependence can also be seen. These individual (or recursive) curves are respectively represented by the dashed and dotted lines. As expected by such dependencies, a tick’s probability of being active is pulled down or pushed up according to its history and is therefore closer to its observed activity status represented by the filled circles.

## 7. HIDDEN MARKOV CHAINS

The intercept model is identical to the one fitted for the generalised auto-regression model with an AIC of 2146.8. The AIC is lowered to 1326.3 by introducing three hidden states. The final model resulting from this modelling process has an AIC of 1265.1 and thirty-seven parameters (Deviance: 1598.329, degrees of freedom: 5232; McCullagh and Nelder, 1992, p.174). It fits considerably better than the final generalised auto-regression model but it is also much more complex. The regression equations for the three states are

$$\begin{aligned}
\text{logit}(m_1) = {} & -593.332 + 0.543\times\text{cumulative rainfall} + 1.246\times\text{rainfall} - 0.301\times\text{change in rainfall} \\
& + 0.312\times\text{acceleration in humidity} + 44.658\times\text{Genda location} \\
& + 31.278\times\text{Nkolowondo location} \\
\text{logit}(m_2) = {} & 5.528 + 1.944\times\text{previous day activity status} + 0.877\times\text{second previous day activity status} \\
& + 3.054\times\text{longest day indicator} + 3.004\times\text{rainy season indicator} \\
& - 0.261\times\text{maximum temperature} + 0.409\times\text{change in maximum temperature} \\
& - 0.294\times\text{acceleration in maximum temperature} - 0.029\times\text{rainfall} \\
& + 0.026\times\text{change in rainfall} - 0.074\times\text{humidity} + 0.104\times\text{change in humidity} \\
& - 0.068\times\text{acceleration in humidity} - 0.634\times\text{Lundazi location} \\
\text{logit}(m_3) = {} & 6.793 + 1.668\times\text{previous day activity status} + 0.213\times\text{second previous day activity status} \\
& + 0.522\times\text{third previous day activity status} - 0.230\times\text{maximum temperature} \\
& + 0.012\times\text{rainfall} + 0.032\times\text{change in rainfall} - 0.021\times\text{acceleration in rainfall} \\
& - 0.032\times\text{humidity} - 0.021\times\text{change in humidity} - 0.392\times\text{Genda location} \\
& + 1.040\times\text{Nkolowondo location}
\end{aligned}$$

where all explanatory variables are significant.

The hidden transition matrix is

$$\begin{pmatrix} 0.980 & 0.020 & 0.000 \\ 0.022 & 0.970 & 0.008 \\ 0.000 & 0.013 & 0.987 \end{pmatrix}$$

and the stationary distribution is (0.409, 0.368, 0.224).

The transition matrix shows that a tick cannot change directly from behaviour described by the first (or third) hidden state to the behaviour described by third (or first). For such a change in behaviour to occur the tick must always go through the transitional behaviour described by the second hidden state. The transition matrix shows that the probability of remaining in a particular state is quite high and that the probability of changing from a state to the next one and from this state back to the previous are almost identical.

The three states can be interpreted as follows. The first hidden state corresponds to a behavior where the ticks are in diapause but takes into account an ageing effect. Indeed, the probability of a tick being active (when following this type of behaviour) will only tend to reach one half towards the end of the study once the cumulative rainfall (0.543 × cumulative rainfall) becomes large enough. Ticks from the Genda (44.658 × Genda-location) and Nkolowondo (31.278 × Nkolowondo location) locations have a constant higher probability of being active than ticks from the other two locations when following this type of behaviour. This indicates that ticks from these two locations age faster and will become active earlier if they remained throughout the study in this type of behaviour.

The second hidden state corresponds to a behaviour where the tick is waiting for indications and once this has occurred, corresponds to an out-of-diapause behaviour. The probability of a tick being active (when following this type of behaviour) will only be greater than one half once the rainy season starts (3.004 × rainy season indicator) due to the presence of the maximum temperature (-0.261 × maximum temperature) in this part of the model. If this indicator is not sufficient to trigger certain ticks out-of-diapause, then this will most likely occur once the longest day of the year (3.054 × longest day indicator) has been reached. After the longest day is reached (on December 21 or day 79), this hidden state describes the influences of external factors on a possible behaviour followed by ticks out-of-diapause. Ticks from Lundazi (-0.634 × Lundazi location) have a slightly lower probability of being active than ticks from one of the other three locations when following this type of behaviour.

The third hidden state only describes an out-of-diapause behaviour, which is influenced by external factors. The probability of a tick being active, when following this type of behaviour, is slightly smaller and more dependent on meteorological factors than the out-of-diapause (or later) behaviour described for the second hidden state.

The differences in behaviour of colonizers (ticks from the Michembo location), settled down ticks (from Genda and Nkolowondo locations), and ticks collected closer to the Equator (Lundazi location) are also captured by this model and described by linear shifts of the probability of being active. Hence, ticks just have a lower or higher probability of being active from the start and the external factors triggering and influencing activity have the same effect on all of them.

This model also points out three different subgroups present at all locations, respectively characterizing "early", "optimal", and "late" reactors. An "optimal" reactor would be a tick that starts in the first hidden state and therefore would just be ageing. It then would go to the second state where it would be waiting for a signal indicating that it is time to start being active. The next step would be to go to the third hidden state where meteorological conditions would drive the amount of activity. Finally, it might go back to the second and then to third hidden state indicating a last intense period of activity, which would be less influenced by external factors. Such ticks would be the first three individuals on Figure 4. An "early" reactor would be a tick that starts directly in the third hidden state and therefore is already very active at the beginning of the study. Such a tick does not need to age before being active nor wait for a signal indicating the beginning of favourable conditions for being active. Hence, its behaviour is straight away influenced by meteorological conditions. Such ticks would be the three middle individuals on Figure 4.

[Insert **Figure 4**].

A "late" reactor would be a tick that starts in the first hidden state and remains ageing for quite a long time. It switches to the second hidden state after the signals indicating the beginning of favourable conditions for being active have occurred. Because it has waited too long for optimal conditions to be active, it must be very active during the remaining time of the study. Thus, such ticks remain mostly in the second hidden state, although they can on an occasion briefly go to the third hidden state. Towards the very end of the study, ageing might catch up with such ticks and switch them back to the behaviour described by the first hidden state. Such ticks would be the last three individuals on Figure 4.

The fit of this model can be assessed from Figure 5. Indeed, it can clearly be seen that the underlying population curve indicated by the continuous line increases slightly and fluctuates accordingly to the meteorological explanatory variables from the start of the rainy season or from the longest day until the end of the study. The dashed and dotted lines represent the individual (or recursive) curves. The dashed curve shows the effect of adding state dependence. Finally, the probability of an individual being active given the optimal path through the hidden states for its activity history is represented by the dotted line.

[Insert **Figure 5**].

## 8. Discussion

This article had two purposes: (1) propose useful new techniques to analyze activity data where a sudden change of the behaviour of an organism occurs, (2) the application on itself is important because new information on the behaviour of adult *R. appendiculatus* could be assessed.

Hidden Markov chains are especially appropriate to study post-dormancy behaviour, portrayed in this study by the activation of *R. appendiculatus* after behavioural diapause. Behavioural diapause involves a temporary interruption in a hierarchical sequence of behavioural patterns (Belozerov, 1982). In contrast, morphogenetic diapause comprises all categories of diapause whereby a development process is temporarily suspended or interrupted as for example the delay in oviposition of engorged *Amblyomma variegatum* females (Pegram *et al.*, 1988). Change point models probably suffice for analyzing morphogenetic diapause as the state is always observable but the unobservable levels that occur in behavioural diapause can probably only be analyzed by using hidden states through hidden Markov chain models. As pointed out by Hodek (1996), it is extremely difficult to distinguish between dormancy levels. Hidden Markov chain models could be useful in clarifying possible hidden states.

[Insert **Figure 6**].

*R. appendiculatus* adults can be in three (non observable) different states as visually shown in Figure 6 a. These states could be related to phases (internal to the tick) during the dynamic event of diapause development and post-diapause activity in *R. appendiculatus*. The first hidden state could be defined as a non-responsive dormant phase. Ticks in this state progressively terminate diapause as their age increases (Madder *et al.*, 2002). This results in a slowly increasing trend of activity. The second hidden state is a responsive dormant phase. In this state, the tick waits for a signal and once the signal occurs, it becomes active and remains active independent of climatic factors. For the data at hand, the signal was estimated to occur at the strong peak of rain at the start of the rainy season. The third hidden state is a non-dormant phase in which ticks react to microclimatic conditions. Ticks could be classified in three groups depending on the way they change or stay in the three states as shown in Figure 6 b. A first group of ticks starts in the first state and is thus not reacting on climatic conditions. These ticks then move to a second state just before the signal to become active occurs and when the ticks get the indication, they become active. Hereafter the ticks move quickly to a third state, where they will active in function of climatic conditions. We can call this group optimal, because it takes optimal advantage of the full rainy season. A second group of ticks remains in the first state for a long time (Figure 6 c) and is thus becoming older. By the end of the study these ticks have to become active because they are so old that if they would stay in state 1 they would die. They thus move to state two, and as the signal has already occurred much earlier they stay active independent of climatic conditions. This group is called a late group as the tick becomes active too late and does not optimally use the full rainy season. An early group of ticks stays in the third state throughout the study and reacts on climatic conditions from the beginning, too early to optimally use the rainy season.

It can be remarked that eastern Zambia is a transition zone, where uni- and bi-voltine populations are observed (Berkvens *et al.*, 1998). This is reflected in flexibility in the behaviour of ticks.

The model shows that activity and non-activity act in an absorbing way meaning that once a tick becomes active it shows a tendency to remain active. The auto-regression models further indicate that activity suddenly changes around the longest day, also indicated by the change-point model. The reaction of ticks on acceleration in rainfall and temperature indicates that ticks might sense climatic changes.

## 9. Acknowledgments

We wish to thank the two anonymous referees for many constructive criticisms and comments on the initial submitted version of the paper.

## 10. REFERENCES (TO ADAPT).

Akaike, H. (1973) Information Theory and an Extension of the Maximum Likelihood Principle. in Petrov, B.N. and Csàki, F., eds. *Second International Symposium on Information Theory,* Budapest: Akadémiai Kiadó, pp. 267-281.

Belozerov, V.N. (1982) Diapause and biological rhythms in ticks. In Obenchain, F.D. and Galun, R. (Eds). Physiology of ticks. Pergamon Press, Oxford.

Berkvens, D.L., Pegram, R.G., and Brandt, J. (1995) A study of the diapausing behaviour of *Rhipicephalus appendiculatus* and *Rhipicephalus zambeziensis* under quasi-natural conditions in Zambia. *Medical and Veterinary Entomology,* **9**, 307-15.

Berkvens, D.L, Geysen DM, Chaka G, Madder M, and Brandt JR. (1998) A survey of the ixodid ticks parasitising cattle in the Eastern province of Zambia. *Medical and Veterinary Entomology*, **12**, 234-40.

Duffett-Smith, P. (1985) *Astronomy with your personal computer*. Cambridge University Press, Cambridge.

Guttorp, P. (1995) *Stochastic Modelling of Scientific Data.* London: Chapman and Hall.

Hodek, I. (1996) Diapause development, diapause termination and the end of diapause. European Journal of Entomology, **93**, 475-487.

Ihaka, R. and Gentleman, R. (1996) R: a language for data analysis and graphics. *Journal of Computational Graphics and Statistics,* **5**, 299-314.

Lindsey, J.K. (1999, 2nd edn.) *Models for Repeated Measurements.* Oxford: Oxford University Press.

Lindsey, J.K. (2001) *Nonlinear Models for Medical Statistics.* Oxford: Oxford University Press.

Lindsey, J.K. and Jones, B. (1998) Choosing among generalized linear models applied to medical data. *Statistics in Medicine,* **16**, 59-68.

Lambert, P. (1996) Modelling irregularly sampled profiles of non-negative dog triglyceride responses under different distributional assumptions. *Statistics in Medicine,* **15**, 1695-1708.

MacDonald, I.L. and Zuchini, W. (1997) *Hidden Markov and other Models for Discrete-Valued Time Series.* London: Chapman and Hall.

MacLeod, J. (1970) Tick infestation patterns in the southern province of Zambia. *Bulletin of Entomological Research,* **60**, 253-274.

Madder, M., Speybroeck, N., Brandt, J. Tirry, L., Hodek, I., and Berkvens, D. (2002) Geographic variation in diapause response in the African brown ear tick Rhipicephalus appendiculatus (Acari: Ixodidae). *Experimental and Applied Acarology*, in press.

Pegram, R.G., Perry, B.D., and Shells, H.F. (1984) Seasonal activity of the parasitic and non-parasitic stages of cattle tics in Zambia. in Griffiths, D.A. and Bowmann, C.E., eds. *Acarology VI, Vol. 2,* Chichester: Ellios Horwood, pp. 1183-1188.

Pegram, R.G., Perry, B.D., Musisi, F.L. and Mwanaumo, B. (1986) Ecology and phenology of ticks in Zambia: seasonal dynamics on cattle. *Experimental and Applied Acarology*, **2**, 25-45.

Pegram, R.G., Mwase, E.T., Zivkovic, D., and Jongejan, F. (1988) Morphogenetic diapause in *Amblyomma variegatum* (Acari: Ixodidæ). *Medical and Veterinary Entomology*, **2**, 301-307.

Pegram, R.G. and Banda, D.S. (1990) Ecology and phenology of cattle ticks in Zambia: Development and survival of free-living stages. *Experimental and Applied Acarology*, **8**, 291-301.

Rosenberg, N.J., Blad, B.L., and Verma, S.B. (1983, 2nd edn.) *Microclimate: The biological environment.* New York: John Wiley and Sons.

Short, N.J. and Norval, R.A.I. (1981) Regulation of seasonal occurrence in the tick *Rhipicephalus appendiculatus* Neumann, 1901. *Tropical Animal health and Production,* **13**, 19-26.

Short, N.J., Floyd, R.B., Norval, R.A.I., and Sutherst, R.W. (1989) Survival and behaviour of unfed stages of the ticks *Rhipicephalus appendiculatus*, *Boophilus decoloratus* and *B. microplus* under field conditions in Zimbabwe. *Experimental and Applied Acarology,* **6**, 215-236.

Speybroeck, N., Madder, M., Van Den Bossche, P., Mtambo J., Berkvens, N., Chaka, G., Mulumba, M., Brandt, J., Tirry, L., Berkvens, D. (2002) Distribution and phenology of ixodid ticks in southern Zambia. *Medical and Veterinary Entomology,* **16**, 1-12.

# FIGURES

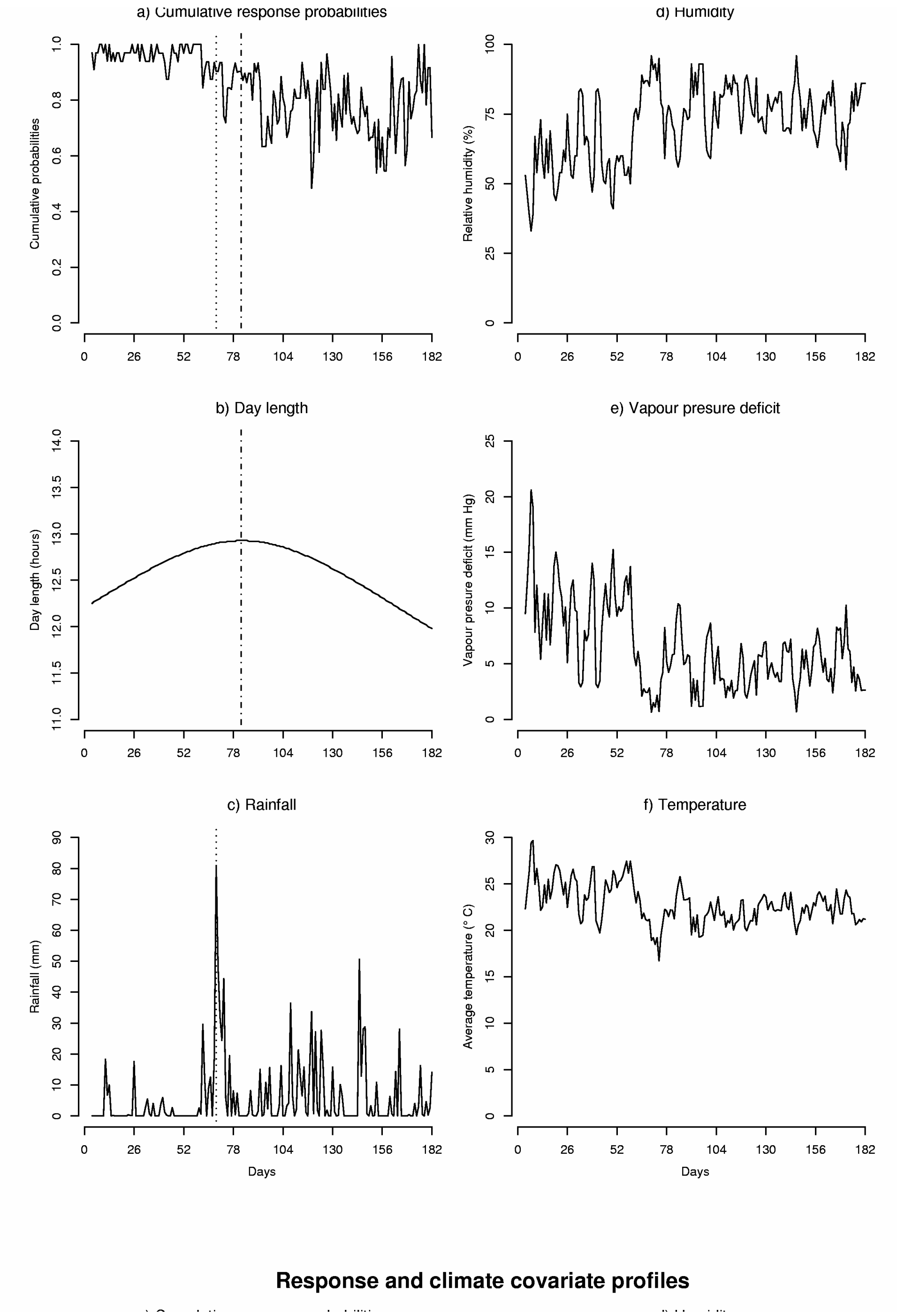


Figure 1: a) Plot of the cumulative probabilities over time for the observed responses. The height of the continuous line represents the overall probability curve of being inactive. On the other hand, the height above this continuous line represents the overall probability curve of being active. The vertical dotted and dash-dotted lines respectively represent the rainiest and the longest days. b) Plot of the day length (in hours) over time. c) Plot of the rainfall (in mm) over time. d) Plot of the relative humidity (in percentage) over time. e) Plot of the vapour pressure deficit (in mm Hg) over time. f) Plot of the average temperature (in °C) over time.

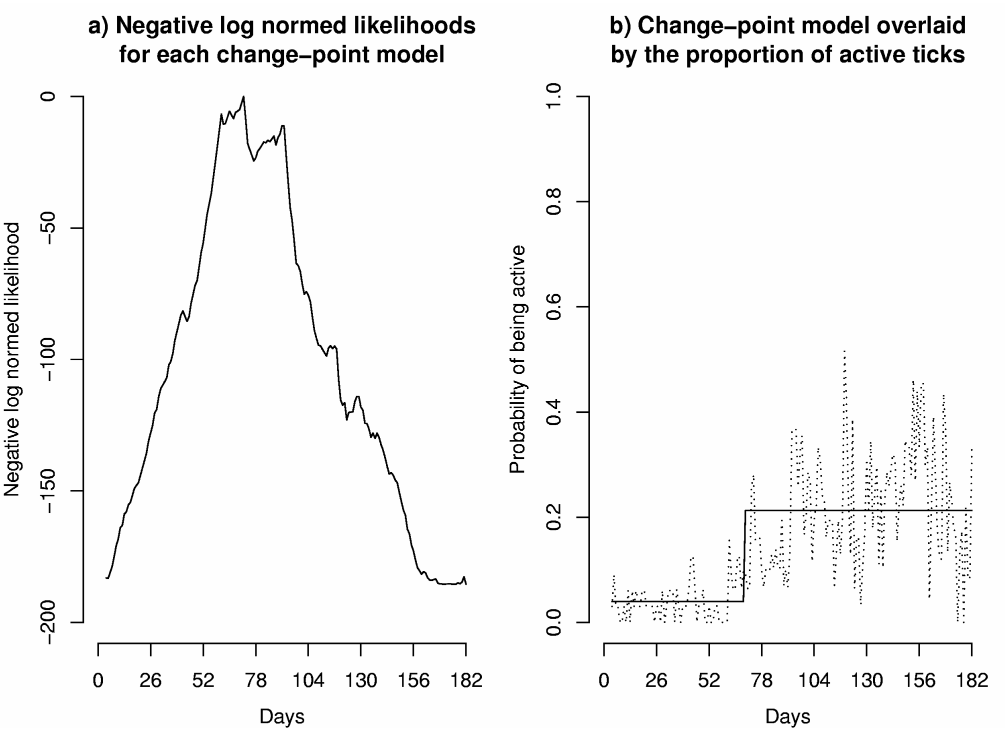


Figure 2: a) Negative log normed likelihoods for the change-point model with a change in activity occurring respectively on each corresponding day. b) Fit of the change-point model for the data (solid line) overlaid by the daily average probability of tick activity (dotted line).

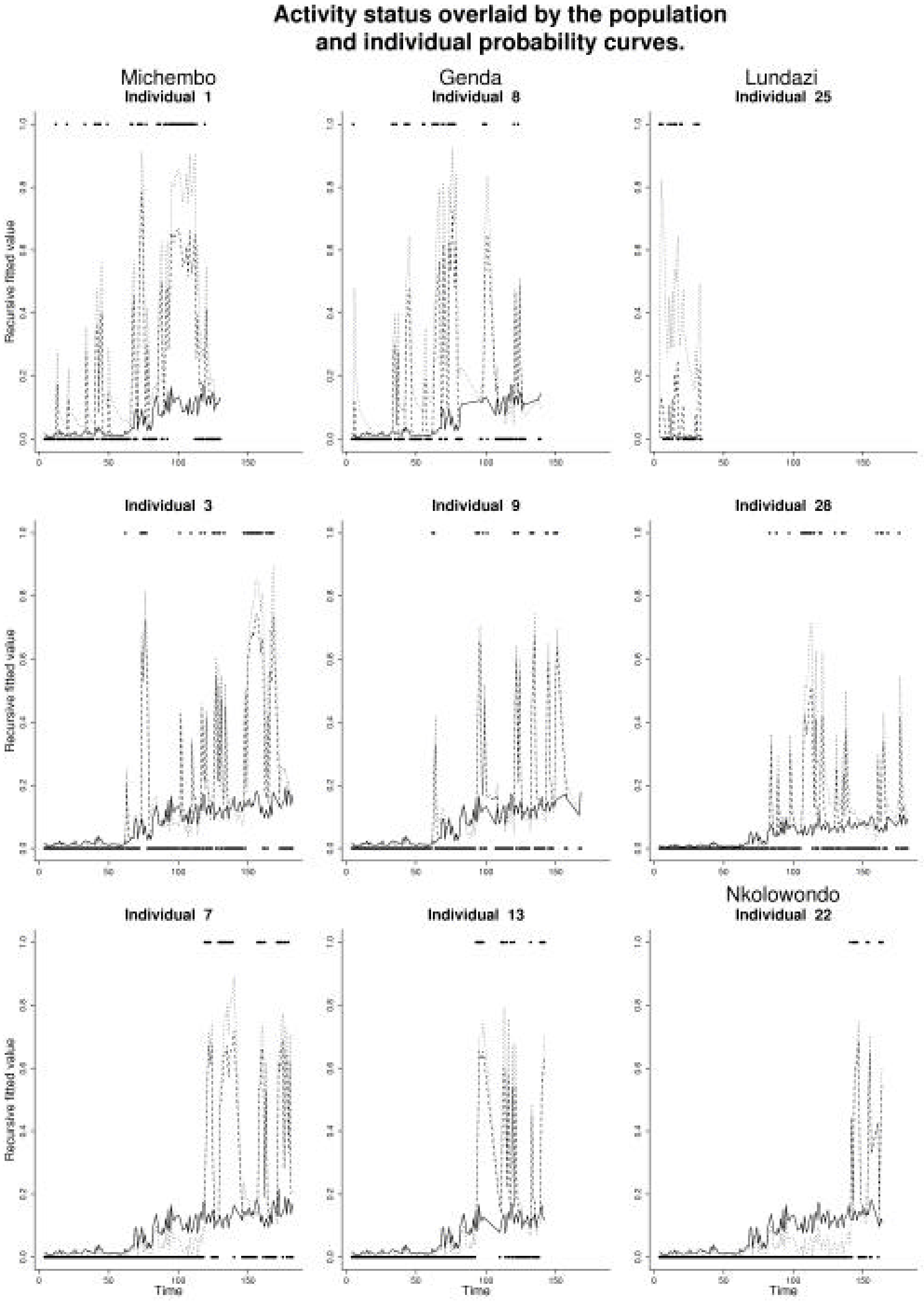


Figure 3: Plots of the activity status (filled points where 0 and 1 respectively indicates inactive and active) for some ticks selected from the different locations. The height of the solid line indicates the underlying population probability curve of being active for the generalized auto-regression model. The dashed and dotted lines represent individual (or recursive) curves obtained for this model. The height of the dashed curve represents a tick's probability of being active taking only the state dependence into account, whereas the dotted curve is obtained by taking also the serial dependence into account.

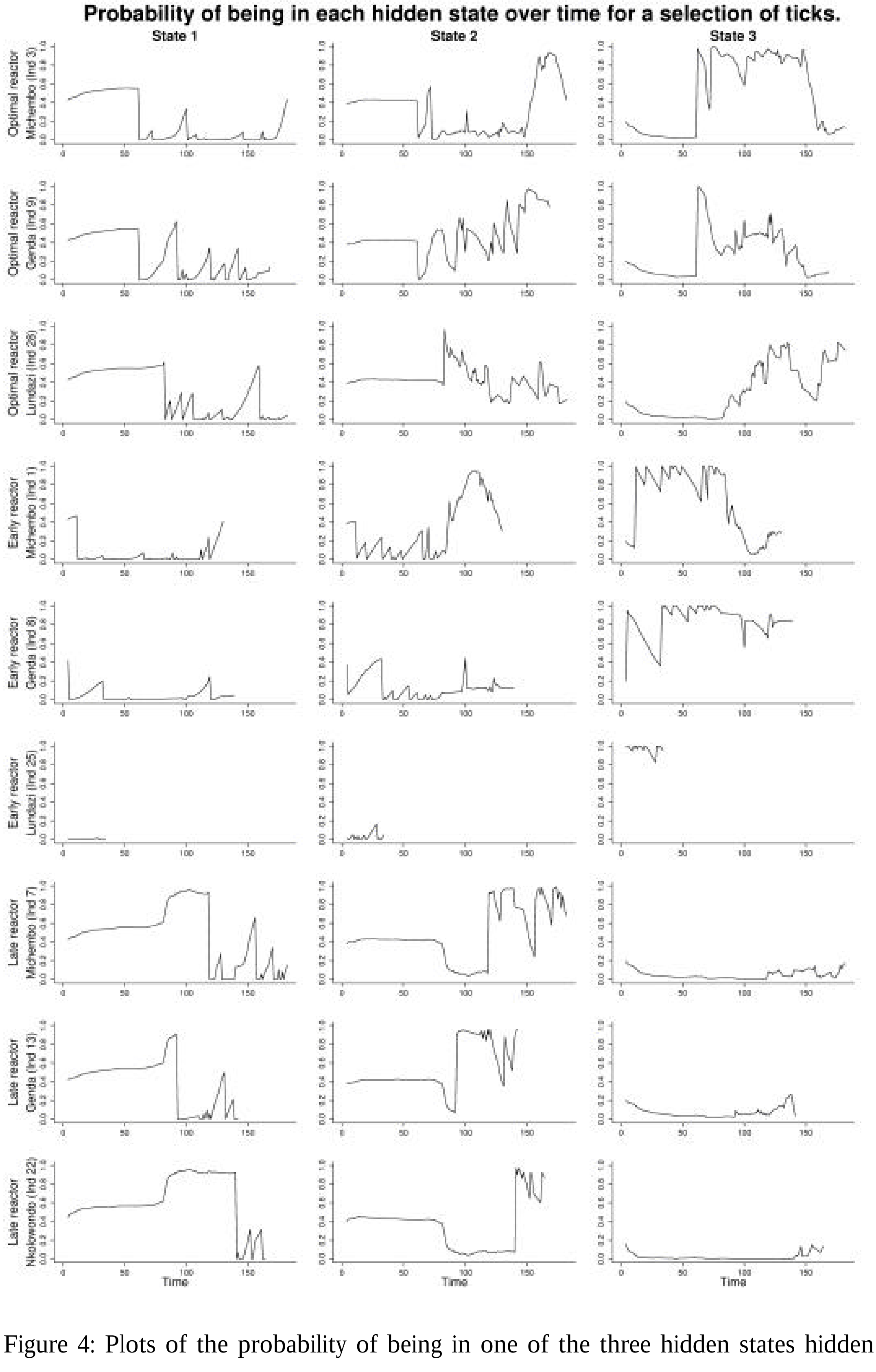


Figure 4: Plots of the probability of being in one of the three hidden states hidden states over time for some ticks selected from the different locations and classified according to reactor type.

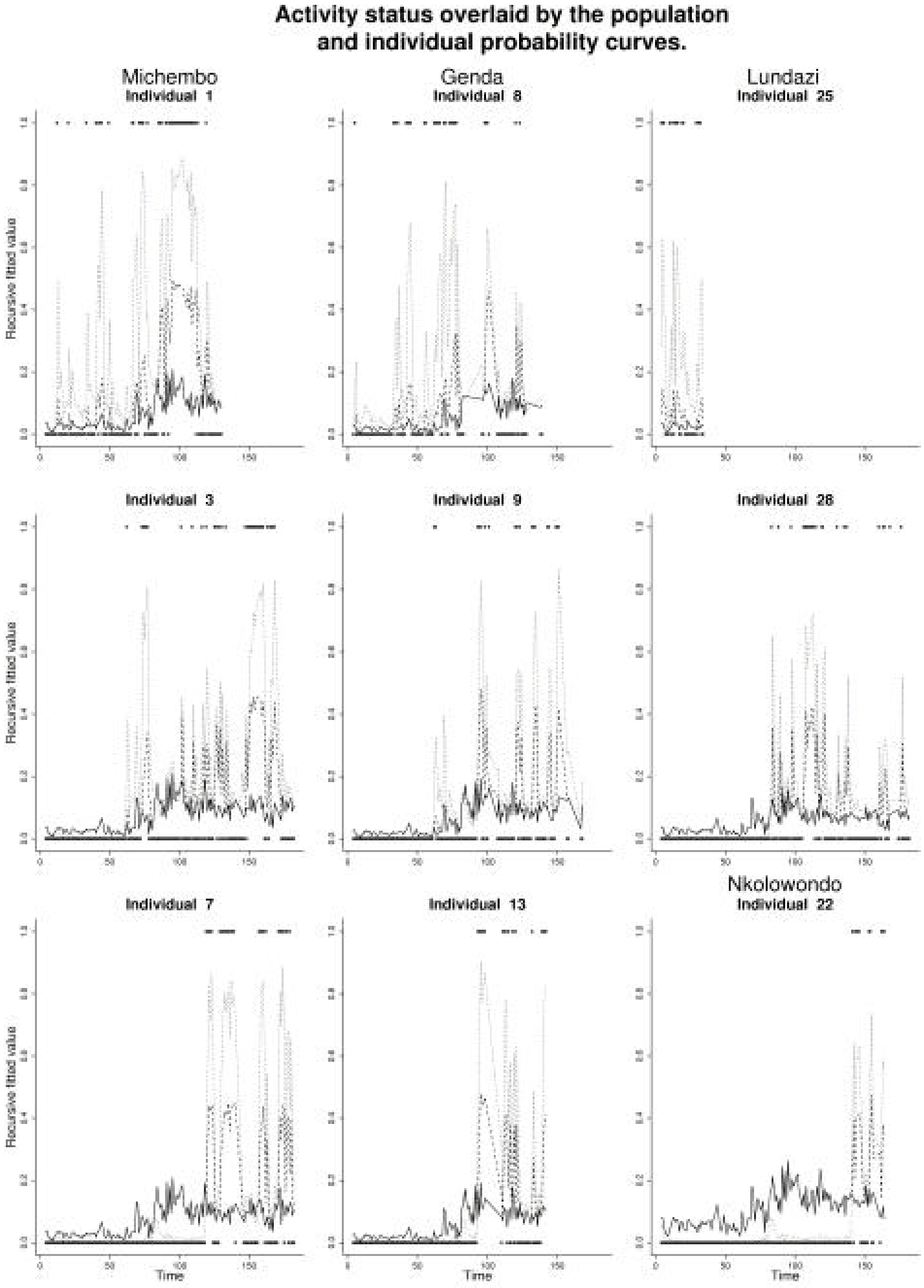
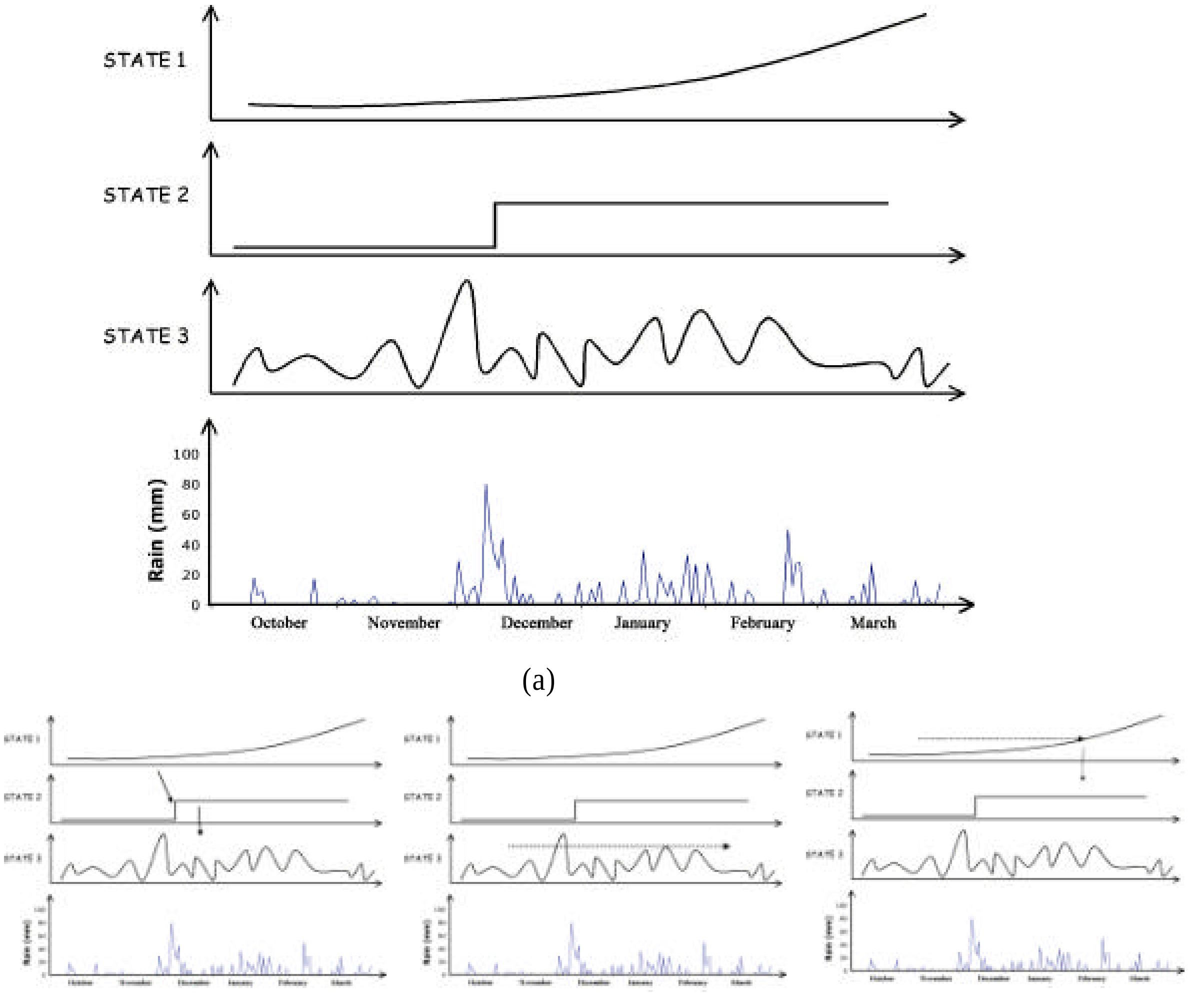


Figure 5: Plots of the activity status (points where 0 and 1 respectively indicates inactive and active) for some ticks selected from the different locations. The height of the solid line indicates the underlying population probability curve of being active for the hidden Markov chain model. The dashed and dotted lines represent individual (recursive) curves obtained for this model. The height of the dashed curve represents a tick's probability of being active taking only state dependence into account, whereas the dotted curve is obtained by taking also a tick's optimal path through the hidden states into account.



(a)

(b)

Figure 6 a: Activity patterns (probability of being active in time) for the different states in which *R. appendiculatus* ticks can be when terminating their diapause, and the rainfall (in mm) in time (below). Top: State 1, Slowly increasing activity independent of climatic conditions. Middle: State 2: No activity until ticks get an indication, thereafter ticks remain active independent of climate. State 3: Activity is a function of climate.

b: Three groups of ticks depending how the ticks change from one state to another.

Left (Optimal Group): Ticks start in State 1, change to State 2 before the rainy season starts , followed by a change to State 3, in which climatic conditions influence the activity patterns. Middle (Early Group): Ticks start and remain in State 3 in which climatic conditions influence the activity pattern. Right (Late Group) Ticks start and remain in State 1 in which climatic conditions do not influence the activity patterns but ticks slowly increase activity because of getting older and at the end of the study the ticks switch to State 2 where tick are active independent of climatic conditions.

## BIOGRAPHICAL SKETCHES

Niko Speybroeck is a biological statistician at the Institute of Tropical Medicine in Antwerp, Belgium, Patrick Lindsey is a statistician at EURANDOM in the Netherlands, Michel Billiouw is a field epidemiologist in Zambia, Maxime Madder is an acarologist at the Institute of Tropical Medicine, James Lindsey is a Professor in Biostatistics at the University of Liege, Belgium and Dirk Berkvens is Professor in epidemiology at the Institute of Tropical Medicine.

The research interest of Niko Speybroeck is ecology, statistics and epidemiology. Patrick Lindsey is mainly involved in statistical genetics, while Michel Billiouw mainly studies the epidemiology of vector-borne diseases of bovines. Maxime Madder has a strong background as an acarologist and conducts experiments with ticks under laboratory conditions. James Lindsey's primary interest lays in nonlinear modelling. He has written a large number of books in the field of statistics. Dirk Berkvens is mainly involved in the epidemiology of animal diseases.

**CONTACT AUTHOR:**

**Niko Speybroeck**

**Mailing Address:**

Institute for Tropical Medicine,

Nationalestraat 155,

2000 Antwerp,

**Belgium.**

**Telephone**: + 32 3 247 62 73.

**Fax**: + 32 3 247 62 68.

**email**: nspeybroeck@itg.be.