VECTOR HOST CHOICE AND THE ENVIRONMENTAL CONTEXT OF MOSQUITO-BORNE VIRUS TRANSMISSION

Wladimir Jimenez Alonso

Wolfson College

&

Department of Zoology
University of Oxford

A THESIS
SUBMITTED FOR THE DEGREE OF DOCTOR OF PHILOSOPHY
MICHAELMAS TERM 2003

ABSTRACT

Vector host choice and
the environmental context of mosquito-borne virus transmission

Wladimir Jimenez Alonso,

Wolfson College

Thesis submitted for the Degree of Doctor of Philosophy,
Michaelmas Term, 2003

The present thesis explored ethological and geographical approaches for the investigation of vector-borne parasites. In the first part, the role of associative learning on vector preferences for hosts was investigated through a comprehensive series of behavioural experiments using the vector of dengue and yellow fever diseases, the mosquito *Aedes aegypti*. To this end, the possibility that the mosquitoes were able to associate unconditional stimuli with particular odours and visual patterns to which they were responsive was explored, but no evidence supporting the hypothesis that associative learning abilities are present in adults of this species was found. A critical review of the literature on learning in mosquitoes conducted afterward allowed the reinterpretation of findings in the field, narrowing the scope of evidence suggesting the existence of these cognitive abilities in some species.

In the second part of the thesis, the distribution and evolution of mosquito-borne viruses was investigated with the use of geo-coded environmental data and spatial statistics. Initially, the eco-climates associated with the distribution of Japanese encephalitis virus were described and modelled, allowing the production of a worldwide predictive map defining the probability of each region to develop this disease in the future. Predominating amongst those areas shown to be under high risk were the equatorial regions of South America and Africa. The methodology used to infer such patterns – non-linear discriminant analysis – was subsequently explored with a number of simulations. Overall, differences in the choice of parameters required for the analysis were shown to lead to differences in the final outputs produced, basically in those cases where the environmental range for which predictions are generated is not rigorously limited. Finally, eco-climate surrogates for the evolution of the Japanese encephalitis serocomplex were investigated, but the current environmental distances between the viruses did not seem to be associated with the events leading to their speciation.

Acknowledgments

I would like to take this occasion to thank all of the many people who made this work possible. First of all, I am very grateful to my supervisor Sarah Randolph and David Rogers, who have welcomed me into their groups and supported me in many different ways. The second part of this thesis not only drew upon their knowledge and experience, but also introduced me to an exciting field of study and very useful set of techniques. Their comments on chapter drafts have also significantly contributed to the improvement of the thesis.

I would also like to thank Tristram Wyatt for his constant support and friendship, and for giving me the opportunity to come to Oxford, and David Kelly, for the various suggestions on experiments and designs, and with whom I enjoyed interesting discussions. Tristram and David were my first supervisors in Oxford, and contributed in many ways to the development of the experiments reported in Chapter 2. The development of the first part of the thesis would also not have been possible without the assistance of many people. I am very grateful to Luke Alphey for providing me with mosquitoes, Peter Burns and José Eduardo Bracco for technical assistance, Gordon Hamilton, Tim Tully, Lester Wadhams, Esteban Fernandez-Juricic and Alex Kacelnik for helpful advices and for equipments. Armando Ulloa Garcia, Paul Johnson, Roger J. Wood, Nicola Seal, Phillip J. McCall, Steve J. Simpson, Sarah Randolph and Martin Birch also contributed with valuable comments and suggestions on the project and/or manuscript presented in Chapter 2. The development of the review presented in Chapter 3 was possible thanks to the generosity of the many researchers that provided me with reprints and supporting material.

Similarly, the support and guidance from many friends and colleagues were critical to the development of the second part of the thesis (Chapters 4, 5 and 6). The many colleagues in the TALA and the Oxford Tick Research Groups were responsible for creating a very nice and pleasant working environment, having helped me with lots of tips and assistance. A special thanks goes to Ben McCormick, for the various hints on computer programming, the enjoyable discussions and for being great company. The development of Chapter 6 would also not have been possible without the information provided by Eddie Holmes, who have kindly sent me the genetic data necessary to run all analyses presented in that Chapter. I am also most indebted to Ivan Araújo, who introduced me to Matlab programming language. Ivan was always keen to share both his time and vast knowledge on programming and statistics with me, and has played a major role on my decision to write the functions and codes used in Chapters 4, 5 and 6.

I also want to thank Steve J. Simpson, who have kindly agreed to be part of my advisory committee in the Zoology Department and has greatly supported me in the transition between research groups in the middle of my D.Phil. My college advisor Roger Hall was also a great source of support and friendship. Eddie C. Holmes and Clive Davies were also very kind in agreeing to be part of my examining board and finding the time to read this thesis.

Euphly Jalles Filho, Paolo MD Zanotto and Flavia Pinheiro Zanotto had a crucial role in giving advices and stimulating the project to come to Oxford. I have also been fortunate enough to have great friends in Oxford with whom I shared many dear moments. Being too many to name, I would like to thank you all for making my experience in Oxford so rich and enjoyable. I am also forever grateful to my whole family for their love and support and especially to my parents, Paula and Francisco, whose motivation and values made me develop an interest for science and nature from an early age.

The development of this thesis and my stay in Oxford was made possible by the financial support provided by the Brazilian government, through a full scholarship from Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES, grant BEX 1521/98-5). I have also benefited from grants and financial aid provided by Wolfson College, Federal University of Rio Grande do Norte (Brazil) and the Association for the Study of Animal Behaviour (ASAB, UK) to participate in Conferences and Scientific Meetings.

Finally, I want to especially thank the person who I love and with whom I have the privilege to share my life, Cynthia Schuck Paim, to whom words would never be enough to describe my gratitude. She was not only crucial for the completion of this thesis through her constant support and immense help in discussing and reviewing all chapters, but also for being an infinite inspiration of intellectual honesty, perseverance and brightness that I so much admire.

for Cynthia

Table of Contents

Chapter 1

FROM ETHOGRAMS TO CLIMATIC MAPS IN THE STUDY OF VECTOR-BORNE PARASITES

Part of the ideas developed in this chapter have already been presented in a review article (in Portuguese):

Alonso WJ 2002. Epidemiology and ethology in the study of the infectious agents: a desirable mutual contamination. In: Proceedings of the XX Annual Meeting of Ethology (Ed. by Albuquerque, F.D.S.), pp. 221-229. Natal, Brazil.

Foreword

Due to a number of circumstances I had the opportunity to develop this thesis in two different fields, both exploring interdisciplinary approaches in the study of parasitism. With a background in the study of insect behaviour (mainly stingless bees) I came to Oxford to develop a project on navigation systems used by wood-pest beetles. Due to an unexpected shortage of animals (*Anobium punctactum*) from various sources in the country, after approximately 6 months it became apparent that it would not be possible to develop this work. I therefore decided to develop a completely different project on host choice behaviour of *Aedes aegypti* mosquitoes with an available colony used for biomolecular studies in the Zoology Department. This gave me the opportunity to develop a study based on the use of ethological tools to investigate questions of epidemiological interest. When it came to the point of generalising my results from laboratory to wild strains of mosquitoes, however, the aggravation of dengue epidemics in several countries made it very difficult to obtain mosquitoes from the field. Together with technical limitations of the available equipment, this problem impelled me to develop my thesis in a different direction, although still within the general field of vector-borne diseases. These circumstances led to the development of the second part of the thesis, an investigation of the use of geographic information systems (GIS) to answer both applied and fundamental biological questions concerning the epidemiology of vector-borne pathogens.

The techniques employed in the thesis are therefore diverse, representing different approaches to the understanding of vector-borne disease transmission. Each chapter was written so as to stand alone, therefore having its own introductory section. Nevertheless, to assist the overall understanding of the questions that guided the development of the thesis, a general introduction and outline of the objectives of each chapter is presented below.

Introduction

In a recent review, the Russian parasitologists Alekseev and Dubinina have argued against the fragmented approach generally adopted in foreign parasitological investigations. According to the authors, while the understanding of specific, isolated factors associated with a parasitic species is essential, researches should also aim to examine parasitic systems from a broader perspective, for instance by considering the ecological scenarios under which species operate, the biology and behaviour of the interacting species as well as their population dynamics (Alekseev & Dubinina 2002). As an illustrative exception to this trend, these authors consider the important findings resulting from the studies conducted by Ronald Ross. By undertaking a systemic line of research, Ross was able to trace for the first time the complete cycle of malaria, to foresee the detrimental role of fishes upon the aquatic stages of the vector mosquitoes, and to model mathematically the epidemiological cycle of this disease.

Leaving the discussion on the effectiveness of reductionist versus holistic approaches for the understanding of living systems aside, it seems that advocating an integrative perspective in the study of parasites cannot be praised enough. By definition, parasites are often harmful organisms that live in or on another living organism (host), from which they derive part or all of its organic nutrients, shelter and other requirements. A parasite's environment is thereby a "living environment", which can be diverse, evolve to be hostile and also die (Combes 2001). To explain the maintenance of these antagonistic – but at the same time tightly fused – systems through time, disciplines as diverse as ecology, physiology, mathematics, evolution, immunology and genetics frequently need to be called upon (May & Anderson 1983, Anderson & May 1991, Frank 1996, Levin et al. 1999, Hellriegel 2001, Galvani 2003).

Epidemiological modelling and ethology

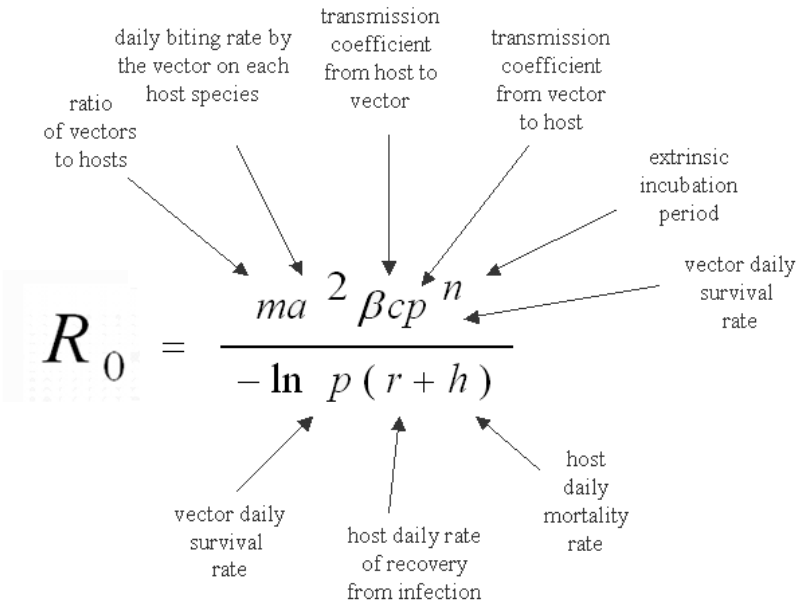
Most of the species inhabiting our planet are parasites (Windsor 1998). Such diversity (as well as abundance), associated with a high evolutionary rate, makes of these organisms one of the major selective forces acting upon living beings. Like many

other biological traits, animal behaviour is also expected to be shaped to a large extent by this evolutionary force. In effect, the last years have witnessed an increasing interest in the influence of parasitism on the behaviour of individuals, for instance in cases of host manipulation by its parasites (Hart 1990, Hart 1992, Combes 2001, Moore 2002). Yet, such an interest is still marginal in the ethological field, being only exceptionally translated into the formulation of research questions and in the design of experiments by students of behaviour. In a recent ethological conference I had the opportunity to illustrate this point by showing that only 3% of the manuscripts published in a renowned scientific journal in the field (Animal Behaviour) in the previous two years had dealt with behavioural aspects related (or potentially related) to parasitism (Alonso 2002). Confirming such observation, amongst the 229 studies presented in the same conference, only two addressed these issues (Galvão 2002, Silva 2002), but only as a result of my invitation for a symposium organized precisely with the aim of stimulating the incorporation of this approach into ethological research.

One of the reasons underlying this lack of interest (or even unawareness) on the role of parasites on behavioural problems might relate to the fact that most courses in biological sciences generally lack the relevant disciplines in parasitology (Windsor 1998). On the other hand, this apathy might be rooted on a perception that, in fact, parasitic aspects of behaviour can be better dealt with by epidemiologists and parasitologists. Such an assessment is, at least partially, founded: under the urgency of having to understand and combat the leading cause of human mortality, epidemiologists have developed a sophisticated theoretical and empirical framework to investigate the relationship between hosts and their parasites (Ross 1911, Macdonald 1957, May & Anderson 1983, Focks 1988, Mitchell 1988, Rogers 1988a, Anderson & May 1991). Still, many of the principles and concepts attached to this body of knowledge could be straightforwardly exploited by ethologists, both through the study of those behavioural patterns influenced by the presence of parasites and of those potentially affecting the transmission of diseases. Reciprocally, epidemiology would also benefit from behavioural studies oriented in this direction, as epidemiological models achieve theoretical consistency and empirical validation through the use of experimental data, but such information source is still surprisingly scarce (Anderson & May 1991).

R_0 equation as a bridging tool

A great deal of both theoretical and empirical knowledge is condensed in epidemiological models, with the most widely known and applied being those aimed at quantifying the *basic reproductive number* of parasitic infections (R_0). This parameter, which is a measure of 'parasite fitness' and expresses how many secondary infections arise in a susceptible host population from each original infection, depends to a large extent on the pattern of infectious contacts within the population of hosts, the probability that the infection is transmitted during contact and the duration of the infections in the hosts (Anderson & May 1991, Dietz 1993). Due to its relevance for epidemiological and ethological studies – and because of its bearings for the present thesis – the utility of this parameter in behavioural research is briefly discussed for the special case of parasitic cycles assisted by vectors (such as mosquitoes). Epidemiologically, vector-borne diseases are notable due to the higher than average basic reproductive numbers (Ro) they can reach: while the R_o of measles (one of the most contagious of the directly transmitted infections) can be approximately thirteen, the R_o of malaria (a mosquito-borne disease) approximates (or even exceeds) one hundred (Rogers & Packer 1993, Spielman 1999, Spielman & D'Antonio 2001). The well-known Ross-McDonald model for obtaining the basic reproductive number of vector-borne parasites is expressed in the equation (1.1) below:



[1.1]

This model was originally proposed by Ross (1911) for *Plasmodium* malaria (transmitted by *Anopheles* spp. mosquitoes) and has undergone several modifications along the years to allow for the incorporation and adjustment of particular features and parameters depending on the situation or species considered (e.g. Macdonald, 1957; Aron & May, 1982; Dietz, 1988; Rogers, 1988a; Anderson & May, 1991).

Even in one of its simplest versions (equation 1.1), the model enables the identification of those parameters potentially influenced by the behaviour of the organisms involved. In fact, all parameters can be indeed affected in this way. The less obvious ones, like transmission coefficients (β and *c*) and the extrinsic incubation period (*n*) also can be strongly influenced by behaviour. E.g., the maintenance of the R_o of the virus responsible for the Rift Valley fever in many parts of north of Africa and Middle East is only possible due to the behaviour of another parasite co-infecting the vector of this disease, the mosquito *Culex pipiens*. By perforating the mosquito's gut to reach its body cavity, the microfilariae (embryos) of the elephantiasis worms also ease the penetration of the virus, reducing its extrinsic incubation period and increasing the transmission coefficient from host to vector (Turell et al. 1984, Spielman & D'Antonio 2001).

But the model also enables the detection of parameters of greater ethological appeal, which are affected by complex behaviours such as those involved in the choice of hosts by the vector and the level of defensiveness of the host. Such behaviours can vary remarkably, even among populations of the same species, with some of them demanding sophisticated cognitive abilities. These parameters are the daily biting rate (*a*), accumulation of vectors around hosts, leading to the ratio of vector to hosts (*m*) and the daily survival rate of the vector (*p*). Their importance is illustrated in the examples described below.

Firstly, the longevity of the infected vector (related to parameter *p* in equation 1.1) after the parasite has been incubated is central in determining how many infective contacts the vector can potentially have with the host. This is of special significance in arboviral cycles (i.e., those associated with viruses able to replicate in an arthropod and

to infect a vertebrate host when the arthropod feeds on that host; Nuttall *et al*. 1988), since a host's immune system reacts to the virus invasion and frequently acquires life-long immunity. Hence, the persistence of the virus will depend on its survival in long-lived vectors (Reiter 1988). Indeed, to delay their activity to a time of better conditions, many vectors hibernate. Hibernation maintains a low physiological age structure in a vector population and still a high survival rate, therefore enhancing the vectorial capacity of parasite transmission. Not rarely, this behaviour can be highly variable within the same population, with some proportion of the individuals continuing active while the remainder becomes dormant (Reiter 1988). The behavioural mechanisms determining how vectors escape temporally (e.g. dormancy) or spatially (e.g. migration) from unfavourable conditions are therefore of crucial epidemiological importance (Mitchell 1988).

Secondly, several behaviours can influence the biting rate on hosts. This is an important epidemiological parameter and, in fact, because at least two contacts are needed for the transmission of the pathogen from one host to another (one contact between the infected host and the susceptible vector, and another between the infected vector and the susceptible host), daily biting rate (a) is squared in equation 1.1 (Aron & May 1982, Rogers 1988b). By manipulating the behaviour of vector and hosts, parasites can affect the pattern of contact between hosts and vectors (Hart 1990, Hart 1992, Combes 2001, Moore 2002). For instance, some virus species might enhance the probability that the feeding behaviour of vectors is interrupted, thus increasing the chances of multiple host contacts and consequently the likelihood that the virus colonizes a higher number of susceptible hosts (Edman & Spielman 1988). Several studies have demonstrated that the defensive behaviours exhibited by hosts, such as those involving grooming and rapid movements, constitute a major factor determining the feeding pattern (related to a in equation 1.1.) and survival of vectors (related to the parameter p) (Edman & Scott 1987, Edman & Spielman 1988, Davies 1990, Hart 1990, Clements 1999). Vector species, in turn, have also evolved strategies to reduce the risks associated with blood-feeding (Edman & Spielman 1988) leading to, e.g., heterogeneous patterns of biting activity of mosquitoes amongst their hosts (Dye & Hasibeder 1986, Hasibeder & Dye 1988, Kelly 2001, McCall & Kelly 2002).

The above-mentioned lack of randomness in the distribution of vectors among hosts, and its consequent implications for pathogen evolution and epidemiology, stimulated the development of several studies on the possible causes of this phenomenon (White 1974, Gillies 1988, Arredondo-Jimenez et al. 1992, Clements 1999, McCall et al. 2001, Ulloa et al. 2002). However, the possible role of the phenotypic plasticity provided by the existence of more sophisticated cognitive abilities is still largely unknown, and a promising area of research (Clements 1999, Kelly & Thompson 2000, Kelly 2001, McCall & Kelly 2002). Under this perspective, in Chapter 2 I report a series of laboratory experiments designed to test the existence of associative learning in the vector of yellow fever and dengue, the mosquito *Aedes aegypti*. In Chapter 3, I review those studies aimed at examining (or which have simply suggested, as based on observed results) the existence of learning abilities in mosquitoes, taking into account recent as well as pioneering investigations that mentioned the phenomenon.

Epidemiological modelling and geographic distribution of pathogens

Epidemiological models such as equation 1.1 use a biological approach, determining the threshold values of the parameters that affect the reproductive number of diseases. There is certainly no doubt about the crucial importance of collecting specific empirical data to feed these models, which not only serve as a heuristic basis for the formulation of mechanistic hypothesis about the dynamics of a disease, but also enable the detection of its reproductive number, hence allowing the understanding of how it will behave in a certain point in space and time. But such an approach does have severe limitations when the modelling process involves situations for which the environmental pathways of the pathogens are only partially understood (Skelly & Weinstein 2003) or for when the exact absolute values of the relevant parameters cannot be known. For most of vector-parasite-host systems, empirical information on appropriate values for the parameters used in the Ross-McDonald model is still lacking, and R_0 cannot be quantified rigorously even in the case of diseases intensively studied, such as malaria (Rogers & Randolph 2000) and dengue (Reiter 1992).

Given the pressure to model the dynamics of various diseases for as many regions and climatic scenarios as possible, an alternative approach has been developed. This is based on focusing on eco-environmental processes, applying computer-modeling techniques to the understanding of enteric pathogens such as *Campylobacter*. (Skelly & Weinstein 2003).

Starting with the assumption that a disease can develop and spread only in a limited subset of environmental conditions, as well as the fact that the current systems of environmental data collection are fine-tuned enough to enable the description of such boundaries, it is now possible to model the spatial and temporal distribution of the diseases even in the absence of information about the parameters used in the biological models mentioned.

Such an assumption is well founded, especially in the case of vector-borne diseases involving invertebrate vectors. In fact, there is currently little doubt that meteorological conditions have a direct influence on the parameters determining the reproductive rate of a disease, such as vector reproduction and survival rates, vector blood feeding frequency and viral development (Mitchell 1988, Reiter 1988, Martens et al. 1999, Hay et al. 2000, Gubler et al. 2001, Hales et al. 2002). These parameters act, however, in complex and interdependent ways. For instance, high humidity is needed for the appropriate functioning of most metabolic processes of vectors. In general, high levels of humidity will tend to prolong a vector's survival (mainly at higher temperatures), although the threat of fungus and bacterial pathogens also increases when humidity is above a certain level, having in such cases an opposite effect on survival. On the other hand, dehydration can, in some cases, induce higher blood-feeding rates to compensate for water loss, thereby increasing the likelihood that an infected vector transmits the disease. The spatial-temporal presence of humidity therefore constitutes one of the factors of crucial importance in the modelling process. Yet, to capture precisely such a pattern of influence the analytical tools used also need to be accurate enough to allow for the detection of the specific limits and conditions where the pathogen can survive and reproduce.

Models referred to as 'geographical' (Rogers 1988a, Rogers 1988b, Rogers & Packer 1993, Levine et al. 2004) comprise a set statistical and analytical techniques that

enable the definition of the eco-climatic variables predicting the occurrence of a disease from almost any kind of multivariate environmental data, as far as such environmental information is available for the geographic locations where the disease was registered, as well as for the geographical range where the probability of disease presence is to be determined. This modelling process therefore involves the statistical investigation and quantification of the environmental parameters determining the distribution of the pathogen, and which are available for the region under assessment. Under the assumption that similar environmental conditions should allow the presence of the pathogen in places where the disease has not yet been reported, it is possible to obtain 'predictive maps', for which every single grid-cell of land area within the map can be assigned a probability that the species is actually present, or could survive if it were introduced into that region.

Such an approach is explored in Chapter 4 to model the distribution of Japanese encephalitis, an arboviral encephalitis representing, in most Asian countries, the daily equivalent of the feared public health disaster that the potential spread of the West Nile virus in the western world could bring (Solomon & Cardosa 2000). While the investigation of the distribution of West Nile and many arthropod-borne diseases has already been carried out with these techniques, the same is not true in the case of Japanese encephalitis, despite the vast impact of this pathogen on public health in the affected areas. Here, I have conducted this investigation by adopting this approach and using non-linear discriminant analyses, an increasingly used tool to study the spatial distribution of species, especially in spatial epidemiological research (Rogers et al. 1996, Robinson et al. 1997, Rogers 1997, Randolph & Rogers 2000, Rogers 2000, Rogers & Randolph 2000, Baylis et al. 2002).

The method is preferred because it provides a straightforward test of the biological assumptions underlying the modelling process at each of its stages, as well as the opportunity of expertise input. However, this is also a source of setbacks, as researchers face several choices during the process of data analysis, which unwittingly insert a degree of subjectivity in the model that is difficult to be evaluated by final users. In Chapter 5 I evaluate how robust these spatial models are to changes in the input values of those parameters open to choice. To this end, I used the model built in the previous chapter as a template.

Once the source of environmental data and the modelling technique are shown to be appropriate to describe the niche of the species, it is possible to venture into the investigation of evolutionary questions related to the environmental context underlying the evolution of vector-borne diseases (Randolph & Rogers 2002). The possibility of reconstructing phylogenetic trees with biomolecular tools now provides us with invaluable evolutionary templates upon which to infer and test the ecological pressures involved in the evolution of a group. In Chapter 6 I explore this possibility by examining the potential eco-climatic factors associated with the phylogeny of the Japanese encephalitis serocomplex (a group of the *Flavivirus* genus composed by virus transmitted by mosquitoes and comprising several pathogens of major impact for human populations).

Chapter 2

ARE VECTORS ABLE TO LEARN ABOUT THEIR HOSTS? A CASE STUDY WITH *AEDES AEGYPTI* MOSQUITOES

This chapter is now published as Alonso, W. J., T. D. Wyatt, and D. W. Kelly. 2003. Are vectors able to learn about their hosts? A case study with Aedes aegypti mosquitoes. Mem Inst Oswaldo Cruz 98:665-672. Available at: http://dx.doi.org/10.1590/S0074-02762003000500014, and is licenced under a Creative Commons Attribution License (http://creativecommons.org/licenses/by-nc/3.0/).

The study reported in this Chapter has been published in the following source:

Alonso WJ, Wyatt TD, Kelly DW 2003. Are vectors able to learn about their hosts? A case study with Aedes aegypti mosquitoes. Memórias do Instituto Oswaldo Cruz, 98: 665 – 672.

Introduction

The spatial distribution of a population of blood-sucking insects amongst its available vertebrate hosts has important epidemiological consequences for the transmission of vector-borne diseases. In fact, after daily survival rate, their vectorial capacity is most sensitive to changes in host preference (Dye & Hasibeder 1986). Therefore, knowledge of the biological parameters that lead to host choice can be highly relevant for planning of disease control (McCall & Kelly 2002).

Regardless of their individual previous experience, mosquitoes are attracted to their hosts based on several stimuli to which each species (and even populations or individuals) responds in a complex and idiosyncratic way (Hildebrand et al. 1996, Clements 1999). Whereas a genetic basis for differences in host choice is a well documented phenomenon (e.g. (Gillies 1964), in addition, several authors (see below and Chapter 3) have suggested also that experience could change some of their behavioural repertoire, such as their response to host and oviposition odours, and even might be involved in the spatial memorization of resources.

The consensus from studies looking for evidence of learning in mosquitoes remains equivocal. Some find no evidence for learning (Rawlings & Curtis 1982, Arredondo-Jimenez et al. 1992). Others have obtained positive results. In some of them, their data do not strictly rule out alternative explanations, such as orientation towards the closest pertinent chemical source (Ribbands 1949, Giglioli 1965, Renshaw et al. 1994), or that the heterogeneity of behaviour found in the population was an effect of the existence of genetic polymorphism (Hii et al. 1991). Charlwood et al. (Charlwood et al. 1988) showed positive results that can only be attributed to learning of a memorized home range, nevertheless this work suffers from 'pseudoreplication' (Hurlbert 1984) owing to the intrinsic difficulties associated with a field study.

Recent works (Mwandawiro et al. 2000, McCall & Eaton 2001, McCall et al. 2001) have been more conclusive. Mwandawiro et al. (Mwandawiro et al. 2000) found in 3 Japanese Encephalitis vectors, that host loyalty to bloodfeed on cow or pigs was not passed onto offspring, indicating that this behaviour has a learning component, instead of a genetical one. In the work of McCall and Eaton (McCall & Eaton 2001)

odours presented during the larval/pupal stage were preferred by adults of *Culex quinquefasciatus* in choice of oviposition site. McCall et al. (McCall et al. 2001) also showed in a fieldwork in Tanzania, how 30 from 44 *Anopheles arabiensis* individuals (recovered from over 17000 found in 2 different houses and released at a intermediate point) return to their original collection places.

Specifically in the case of the search for available hosts by mosquitoes and other blood-sucking insects, Kelly and Thompson (Kelly & Thompson 2000) propose that the ability to learn would be adaptive to individuals aiming to maximise their feeding success and minimise their chances of being killed in the process, to the extent that it would allow them to feed preferentially on the least defensive hosts. For instance, individual vectors could be more 'loyal' to a sub-group of host species or populations upon which they previously fed successfully.

Aedes aegypti is the primary vector of dengue and yellow fever, both diseases caused by virus species of the *Flavivirus* genus and that have a major impact on human health. Here we conducted a comprehensive series of experiments to investigate the existence of associative learning in *Ae. aegypti* based mainly (but not exclusively) on behavioural aspects potentially related to host choice behaviour.

Materials and Methods

Insects

Ae. aegypti individuals were taken from a laboratory culture present in the Zoology Department of Oxford University, which before reaching the UK belonged to the University of Notre Dame, Indiana (USA), where it was established in the 1950's (Roger J. Wood, pers. commun.). The culture room was maintained at 26-29^{o}C and 60-80% relative humidity, under a 12:12h light:dark cycle. Larvae were fed with dog biscuits and TetraMin tropical fish food. Adults were maintained in mixed-sex colonies (cages measuring 17 x 17 x 17cm) in cohorts of the same age, and provided with water alone until used in the experiments. Experimental adults were between 1 and 3 days old, and used only once. In the case of females, only those individuals that responded positively to human breath were used. Transport between cages and testing apparatus

was carried out using a manual aspirator. In the experiments where required, warm defibrinated sheep's blood (TCS Microbiology Company®) was provided through a membrane feeding system (Hemotek®).

Experimental room

The experimental room temperature ranged between 24-27^{o}C. A single 40W lamp placed 80 cm overhead illuminated each testing apparatus. A cylindrical netting curtain screen was connected from the lamp to the bench top, completely encircling the apparatus, so that an observer could monitor the behaviour of the subjects without disturbing them. An overhead video camera installed beside the light source recorded mosquito behaviour for later analysis.

All experiments were performed between 10:00 and 18:00 h. All materials were carefully handled with disposable gloves to prevent odour contamination, and the testing apparatus was cleaned with 99.7% ethyl alcohol after each replicate.

Experiments using odours as conditioned stimuli

An apparatus (Fig. 1) was built to test whether the mosquitoes could associate either punishment or reward (the positive or negative 'unconditioned stimuli', US, respectively) with 1 of 2 odours (the 'conditioned stimuli', CS). The odours were presented on filter papers in cartridges upwind of the Y-tube (Fig. 1D). In the training stage, odour A was presented without additional stimulus, and odour B was presented with an unconditioned stimulus (for details of each experiment, see below). In different replicates the position of the odours, the order of the association with the unconditioned stimuli (first or second), and the odour that was associated with the unconditioned stimulus was varied in a balanced design, to control for these variables in the subsequent analysis.

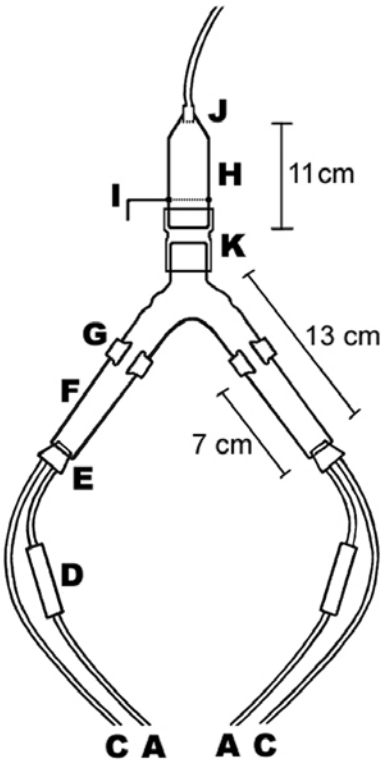


Figure 2.1. Diagram showing the apparatus used for testing whether *Aedes aegypti* mosquitoes could associate punishment or reward with particular odours. The humidified and filtered air (see text) is conducted in by a series of valves (not shown) as fresh air (C), or as odorised air in pipes A and B, after passing through the odour cartridges (D). Passing through the rubber stoppers (E) the air enters the arms of the Y tube (F) that is attached to a table by a pair of clips (G). The mosquitoes' training vial (H) has a screen door (I) that allows the air to flux constantly, until it is expelled from its other extremity (J) to outside the room. The door can be opened manually from outside to allow the mosquitoes to enter the Y tube. The manufacture of the connection (K) between H and F varies depending upon the training procedure (see text).

During the training stage the mosquitoes remained inside a plastic cartridge (the conditioning chamber: Fig. 1H), and were exposed to odours A and B consecutively. In the 'testing stage', the two odours were presented simultaneously, each flowing down one arm of the Y-tube (Fig. 1F). A metal gauze gate (Fig. 1I) was opened, allowing the mosquitoes to move into the Y-tube.

The air that entered the apparatus was pushed with a oilless pump (Charles Austen Pumps, model F.65 DE), filtered with activated charcoal and humidified to approximately 80% RH. The airflow was controlled at 4 l/min per minute so that a speed of about 4 m/min in the exit end of the Y tube was achieved. Air from the apparatus was expelled outside the room (Fig. 1J). The suitability of the apparatus was tested in preliminary trials by presenting individuals with a choice between a pure flux

of air and a flux of air mixed with human breath, which showed that most individuals consistently preferred the latter arm of the Y-tube

Unless otherwise stated, in each experiment each series of trials consisted of 20 replicates for each sex, each replicate containing 20 individuals.

Vibration as negative unconditioned stimulus

Mechanical activity, in the form of skin twitching, stamping, slapping, etc., is one of the main cues that a mosquito can use to gather information about the defensiveness of a potential host (Edman et al. 1998, Clements 1999, Kelly 2001). We tested whether *Ae. aegypti* mosquitoes would be able to associate odours with vibration of the conditioning chamber by fixing the vial (Fig. 1H) to a small aquarium pump (Tetra Whisper 400), which made it vibrate at 50 Hertz (amplitude:1mm). This stimulus seems to have provided a significant level of disturbance for the individuals inside the vial, which kept taking off and trying to land on its walls most of the time unsuccessfully when the vibration was applied. In order to provide a flexible junction between the Y tube and the mosquitoes' vial (Fig. 1K) cut fingers of powder-free latex gloves were used.

The odours presented as CS were 20 µl of carvone (Aldrich 435759) or citral (Sigma C83007). These odorants were used since (1) in preliminary trials we observed that the mosquitoes were able to perceive them (avoiding them in upwind flight when they were presented separately in the Y tube in a choice with a fresh air option), and (2) because they are chemically very distinct, and this difference seems to be noticed by the mosquitoes given their unequal distribution in the Y tube – with a small tendency towards preferring the citral – when both were presented together. (Using the same training and testing procedure, but without the presentation of the CS – see methodology elsewhere. Number of replicates for each sex: 10. Number of individuals per replicate = 20. Results – female: citral mean = 5.1 s.d. = 2.5; carvone mean = 4.2 s.d. = 1.8; male: citral mean = 6.2 s.d. = 1.5; carvone mean = 3.1 s.d. = 1.5. Dependent variable = mosquitoes in each arm after one min).

The sequence of odour presentation for each set of mosquitoes was 30 sec of one odour associated with the unconditioned stimuli and 30 sec of the other one alone. Fifteen seconds of resting (with fresh air, Fig. 1C) were allowed in between the presentations of the odours. This procedure was performed 5 times before conducting the test stage. The same experiment was used to test females and males separately.

Breath as positive unconditioned stimulus

Human breath provides reliable information about a host presence and its position and strongly attracts *Ae. aegypti* (Kellogg & Wright 1962). Odours associated with it could be used as additional cues to find a host. We therefore repeated the previous experiment but instead of providing vibration as a negative US, we provided human breath as a positive US.

As the junction between the Y tube and the mosquitoes' vial we used a hard rubber tube in which a metal pipe with small holes (for gas diffusion) was introduced. During the association period, human breath, blown continuously by the experimenter, was diffused (c.a. 0.5 l/min) into the chamber.

Blood feeding as positive unconditioned stimulus

A novel odour associated with successful blood feeding could provide important information to be used by mosquitoes when seeking a subsequent blood meal. To test this, mosquitoes up to 3 days old were randomly distributed between two cages. On the following day, one of the cages (control) was transported to another room (same conditions of humidity and temperature), left undisturbed and provided with a source of blood for 70 min. At the same time, an aquarium pump (Tetra Whisper 400) pushed a small flow of fresh air into a cartridge containing a blank filter paper and into the cage. Subsequently, another cage (treatment) was placed in the same room, in the same conditions (with new blood in the feeder), but with a filter paper containing 20 μl of citral in the cartridge connected to the air pipe. In a previous experiment we observed that this amount of citral is the minimum quantity that the subjects could detect: they were significantly repelled by it when given a choice between this amount of citral and fresh air, something that did not occur when the amount of citral was only 2 μl.

Although most of the females had engorged with the blood provided, in the following day (3rd experimental day) the procedure was repeated with both experimental groups to allow a second opportunity for feeding, as *Ae. aegypti* females regularly feed on the second as well as the first day of the gonotrophic cycle (Canyon et al. 1999). On subsequent days sugar diluted in water (1:10) was made available.

Three and 5 days later (6th and 8th exp. days), after observing that a large number of eggs had already been laid, 20 replicates (each with 10 females) of each treatment were tested in the Y tube in a balanced way. The subjects were presented with choices between filter paper containing 20 μl of citral, and blank filter paper. The number of individuals in each arm was scored as stated for the previous experiment.

The difference in the number of individuals that chose the arm with the odour, against those choosing the control arm, was then compared between the control and the treatment groups. Only females were tested.

Electric shock as negative unconditioned stimulus

Although a stimulus that is not usually found in nature by insects, electric shocks have proved to be a useful tool for investigating learning abilities in this class of organisms, notably in another dipteran species, *Drosophila melanogaster*. We adopted the method of Tully and Quinn (1985), for our apparatus. Our electrified grid was the same as that used in experiments with *Drosophila* (Quinn et al. 1974, Tully & Quinn 1985). It consisted of a copper pattern printed on a 57 x 81 mm flexible epoxy backing, which was used to line the inside of the training vial. During the test stage a separate circular grid also covered the extremity of the training vial that opened towards the Y tube. Therefore, approximately 70% of the total surface of the training vial was lined with the electric grid. The possibility that some individuals remained during the whole experimental period in the non-electrified surface was lowered since many of the subjects were prompted to fly by the electrical pulses, disrupting the resting position of the others.

The electrical stimuli consisted of pulses of 120 V (50Hz), lasting 1 sec and given at intervals of 5 sec. This voltage was chosen as the lowest one in which we were

able to observe reactions from the mosquitoes. Only one training cycle with electrical shock was applied to each group of mosquitoes, as employed in fruit fly studies.

In order to minimize the possibility of negative results as being consequence of the odours and quantities used, 2 distinct sets of experiments were conducted using the electric shock as US. In the first one, the odours and amounts were identical to those used in experiment A1 (citral 20 μl, carvone 20 μl). In the other, citronella oil (EHP-Natural Product Division, Batch No. SR1925) and eugenol (Sigma E5504) were employed, but this time the amount in the filter paper was 2 μl: sufficient to promote an avoidance response by the mosquitoes when presented against fresh air into the Y tube. Only females were used.

Analysis of the data

The response measures were obtained by analysing the video-tape records and scoring how many individuals were in each arm 1 min after the opening of the gate. Other periods were also registered (15 sec, 30 sec, 2 min, 3 min, 4 min and 5 min) in some experiments, but 1 min was chosen because it was the shortest time after the training phase when the distribution of individuals between the arms of the Y tube could be considered well established (see Fig. 2). Only those subjects located at least 7 cm beyond the entrance of either arm of the Y tube were considered.

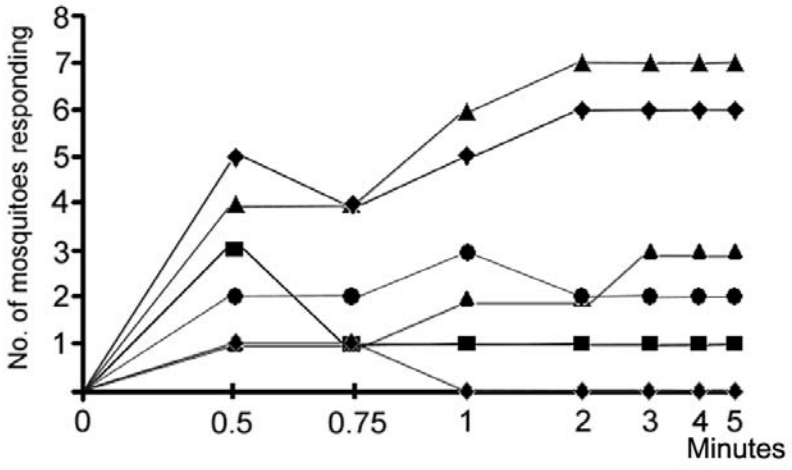


Figure 2.2. Examples (randomly chosen) of how the mosquitoes distribute between the two arms of the Y tube over a period of 5 min (intervals of 15 sec, 30 sec, 1 min, 2 min, 3 min, 4 min and 5 min). Each line represents a different observation.

The non-parametric Wilcoxon matched-pairs test was performed separately for the replicates of males and females in each experiment. One-tailed tests were performed, as we predicted a priori that subjects would avoid the arm releasing the odour associated with the US when this was a punishment (vibration, electrical shock) and would choose preferentially the arm with the other odour. In those cases where the stimulus was either human breath or blood, we expected that the subjects would use the odour released simultaneously as a cue of host presence, and during the test they would move preferentially towards the odour associated to the CS. For all statistical tests, the alpha level was set to 0.05.

Although the experiment was designed in a way so as to compensate for possible trends towards one of the sides of the apparatus and one of the odours, and the order in which the odours were presented, these variables were also analysed to investigate potential biases.

Experiments using a visual cue as Conditioned Stimuli

Here we tested the hypothesis that *Ae. aegypti* mosquitoes are able to search for their previous host (or host type) in order to resume an interrupted blood meal based on the memory of a visual pattern associated to a successful blood-meal received a few seconds before.

For this experiment, we set up a simple device where images could be presented and hidden, just by rotating a white bucket (internal bucket, 'IB') with 4 holes (8.5 cm of diameter) in its walls, inside another white bucket (external bucket, 'EB') containing black images and patterns on its inner walls. A transparent plastic lid on IB allowed the illumination, observation and video-recording of the interior of the set-up. Both buckets (HK-Plastics EC 10000) had a capacity of 10 l (c.a. 23 cm height and 26 cm diameter).

Each replicate, using a naive individual, consisted of: (a) the training stage, where a visual pattern (CS) was displayed on top of a feeding membrane offering warmed blood (US) at its centre, through a hole in the wall of the EB; (b) the testing stage, where 2 distinct patterns (positioned at a distance of 17 cm, or 90º of arc from

each other) were revealed on the opposite side from the now concealed training image and feeder.

Upon release, each individual generally landed on the visual pattern for a few seconds, and from there – and probably guided by the warmth of the feeder – moved towards the centre of the image where they could blood feed. After allowing the individual to feed for 25 sec, its feeding was interrupted by a gentle rotation of the IB. This process was repeated 3 times for each individual. *Ae. aegypti* females take about 284 ± 40 sec to take their blood meal (Jones & Pilitt 1973), and they will still try to complete their blood meal immediately if this is interrupted before a critical amount of blood is consumed. We chose three blood-feeding periods of 25 sec because, after them, the mosquitoes were clearly prone to resume their host search at once. After training, the testing stage was introduced by rotating the IB in a way such that the training shape was hidden to the mosquito, which could now see, approach and land on the testing shapes.

The images were made of a black adhesive tape stripe (8 x 2.5cm each), in one case placed vertically on the bucket's wall, and the other one horizontally. We anticipated that the mosquitoes could distinguish between these patterns in our experimental device. The eyes of adult *Ae. aegypti* have relatively poor resolution (≈ 12°, (Muir et al. 1992a)), but this should be enough to distinguish the visual cues (given their size, proximity and simplicity) we presented to them. As these mosquitoes are attracted mostly to stationary objects of low reflectance and solid colour (Muir et al. 1992a) we used only continuous black images to provide a maximum contrast over the white background, and because colour is not a stimulus to which they respond (Muir et al. 1992b). The experiment was blocked for the direction of the rotation of the bucket, the position of the images in the training stage, and the type of image used as US.

Analysis of the data

Video records of 41 replicates (out of 60) were considered for the final analysis. The other 19 were rejected, either because they were not clear enough or because the mosquito spent most of its time inactive or flying towards the light. Data records from the videotape were (blindly) analysed for 1 min from the onset of the testing stage. We

predicted that, in the existence of preference towards the image on which they had fed before (CS+), the mosquitoes would approach that image and/or land on it in preference to the control one (CS-) in the following ways: (a) choosing it first after the presentation of the 2 options in the test stage; (b) presenting a shorter latency to choose it when they chose it first; (c) choosing it with a higher frequency; (d) once landed, remaining a longer time on it.

An approach towards each image was scored if the individual entered into an area drawn around each target in front of the monitor screen. Landing was recorded when not only the individual touched the visual pattern, but also stopped flying and perched in the visual pattern surface. Several tests were employed depending upon the nature of data (see Results for details).

Results

Experiments using odours as conditioned stimuli

Vibration as US

After 20 replicates the trends obtained for both sexes were an avoidance of the odour associated with punishment (CS+), as predicted. In males (Fig. 3) this trend was not significant ($P = 0.23$). However, a significant trend ($P = 0.02$) was detected in females. As we had a high probability of obtaining positive results (learning in this species) by chance owing to the multiplicity of tests for the same hypothesis, we decided to double our sample size to 40. At this point, the significance was not maintained ($P = 0.17$) and we could not reject the null hypothesis that the higher proportion of mosquitoes in the 'escape' arm was obtained by chance. These results are summarized in Fig. 3.

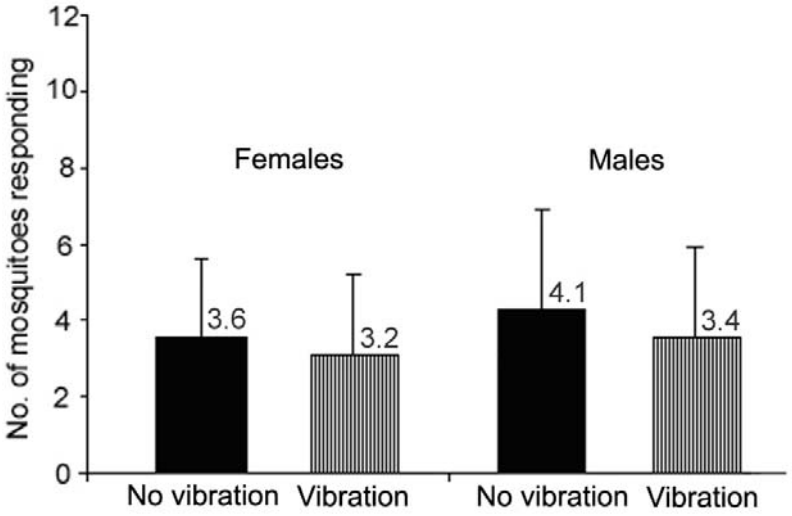


Figure 2.3. Mean number of mosquitoes in the arm that released the odour previously associated with vibration 1 min after the beginning of the test stage, versus the number of mosquitoes in the arm with the odour not associated with punishment. Females n = 40 replicates, males n = 20 replicates. Each replicate comprised 20 individuals.

Human breath as US

There was no preference for breath as the CS (Fig. 4). In males, there was a slight tendency to approach, as predicted, the 'breath' arm, but this trend was not significant ($P = 0.43$). Contrary to our expectations, females were more likely to choose the arm releasing the odour not associated with human breath.

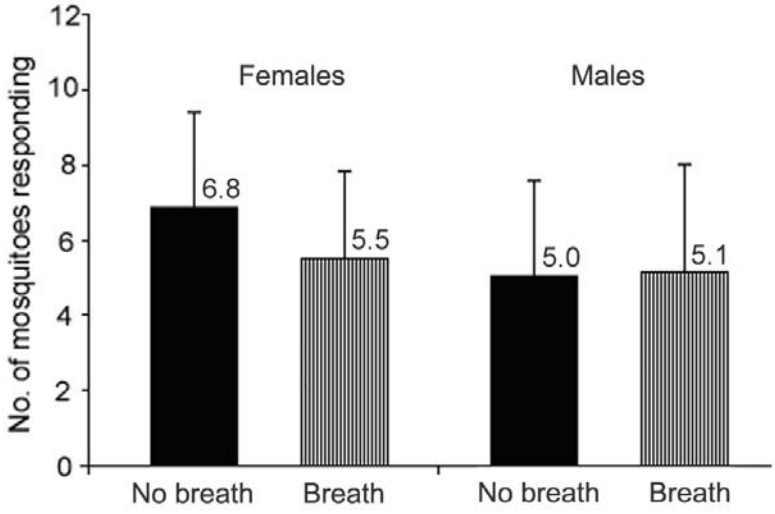


Figure 2.4. Mean number of mosquitoes in the arm that released the odour previously associated with breath 1 min after the beginning of the test stage as compared to the number of mosquitoes in the arm with the odour not associated with this reward. Females n = 19 replicates, males n = 20 replicates. Each replicate comprised 20 individuals.

Blood feeding as US

The proportion of mosquitoes that went to the citral stimulus did not differ between the groups that had been exposed to citral when blood feeding and those exposed to fresh air when blood feeding (Fig. 5). Statistical testing has not been applied in this experiment because all the control and all the treatment mosquitoes were trained separately, but all the mosquitoes in each treatment were trained together; therefore they were not independent.

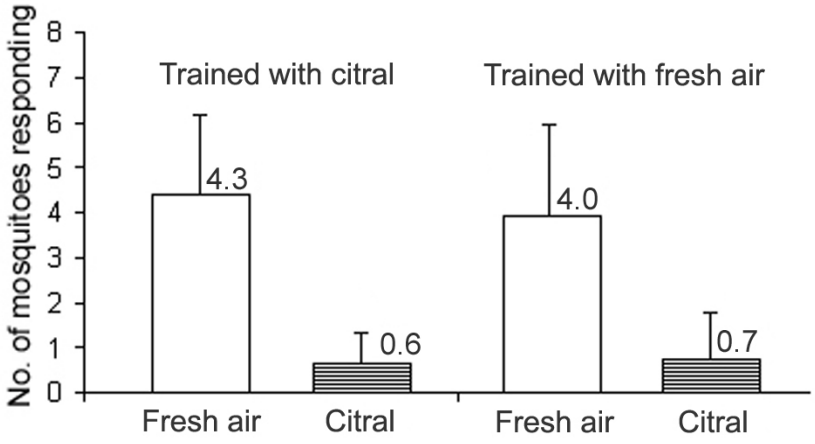


Figure 2.5. Responses to citral versus fresh air between mosquitoes submitted to exposition to citral and those submitted to only fresh air when blood feeding. Results are expressed as the mean number of individuals that 1 min after the beginning of the test stage were in each arm of the Y tube. N = 20 replicates for each treatment group (only females employed). Each replicate comprised 20 individuals.

Electric shock as US

When citral (20 µl) and carvone (20 µl) were employed as CS, a non-significant (P = 0.09) trend towards avoiding the odour associated with electrical shock was obtained. This tendency was not maintained when citronella oil (2 µl) and eugenol (2 µl) were used instead (Fig. 6).

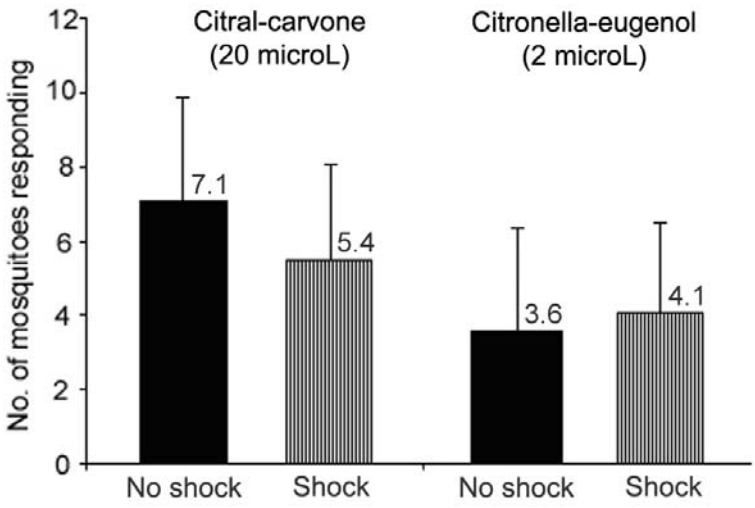


Figure 2.6. Mean number of mosquitoes in the arm releasing the odour previously associated with electric shock 1 min after the beginning of the test stage, versus the number of mosquitoes in the arm with the odour not associated with punishment. The odorants employed are indicated above the results of each comparison. N = 20 replicates for each treatment group (only females employed). Each replicate comprised 20 individuals.

Applying Bonferroni correction (Sokal & Rohlf 1995:703) on the Y-tube data also does not allow the rejection of the null hypothesis, as the critical alpha level is divided by the number of tests (in this case n = 7), which give us a $\alpha' = 0.007$, a value smaller than the p-value ($P = 0.02$) we had initially obtained.

Experiments using a visual cue as Conditioned Stimuli

All mosquitoes included in the analysis (n = 41) performed several approaches to both patterns (Table - c), and only two of them approached only one of the patterns. Most of the mosquitoes did not land on any of the visual patterns during the first minute of the test stage (Table - c). Among those that did land, 12 individuals did it only on 1 of the patterns (7 on the CS+, and 5 on the CS-), 1 did it on both patterns, and finally 1 mosquito landed twice on the CS+ and once on the CS-.

Table 2.1. Parameters and statistics relative to the data obtained in the first minute of the test stage. 'CS+' refers to the image that had been previously associated, for 'conditioning' purposes, with blood-feeding. 'CS-' refers to the image that had not been previously associated to blood-feeding. When pertinent, descriptive statistics are shown as mean ± standard deviation. One-tailed tests are summarised below each pair data. When the direction of the results is the expected, one-tailed statistics are showed below the descriptive data, otherwise only n.e.d. (non-expected direction) is shown.

	[a] First choice (number of mosquitoes)		[b] Time until 'first choice' (sec)		[c] Occurrence		[d] Duration (sec)	
	CS+	CS-	CS+	CS-	CS+	CS-	CS+	CS-
Landing	9	5	23.1 ± 16.1	38.5 ± 15.3	0.24 ± 0.49	0.17 ± 0.38	28.7 ± 25.1	18.6 ±13.8
	Binomial test *P*=0.21		Mann-Whitney 1-tailed *P*=0.06		Wilcoxon Ranks 1-tailed *P*=0.20		Mann-Whitney 1-tailed *P*=0.34	
Approach	22	19	11.8 ± 6.5	9.1 ± 7.2	8.1 ± 4.5	8.5 ± 5.7		
	Binomial test *P*=0.38		n.e.d.		n.e.d.			

Most mosquitoes approached and landed first on the image where they had been previously blood-fed. Nevertheless, these trends were not significant (Table - a). However, they did not approach the CS+ more quickly (in fact, the trend was in the opposite direction, Table - b).

The time taken to the first landing was less for the CS+ as expected, but again it was not significantly different from the time taken to land on the CS-. Landing frequency was higher for the CS+ (Table - d), but not significantly different from that for the CS-.

Therefore, overall, these results do not provide us with unequivocal evidence for the existence of a preference for the visual pattern on which subjects were partially blood-fed previously.

Discussion

There can be little doubt about the basic assumption that mosquitoes have a neurological in-built equivalent to that found in *Drosophila*, where learning and

memory skills had been extensively described (Waddell & Quinn 2001). Their phylogenetic proximity and the neuro-anatomical descriptions of mosquitoes support this hypothesis. Therefore, finding learning ability in mosquitoes would not be surprising (McCall & Eaton 2001).

However the aim of our research was to find learning abilities that could make adaptive sense to the problems that a mosquito faces in nature, such as blood-feeding. Certainly the odours we employed were not the ones that this species would naturally use to locate hosts, but we purposely chose odours that were detectable by our mosquitoes, but were not innately attractive, as would be the case for e.g. CO_2, acetone, lactic acid or 1-octen-3-ol (Hildebrand et al. 1996, Clements 1999). We did so because we predicted that, in order to detect an olfactory cue that could be used to 'qualify' a sub-set of hosts during the life-span of a mosquito, individuals could not rely on features that are shared by all hosts or do not change between mosquito generations. Instead, they should rely on very particular odours (like those derived from the food ingested by the host) that would influence in an idiosyncratic way the general odour profile of the host.

Visual patterns would also be expected to be useful for allowing an individual mosquito to find a previous host and resume an interrupted blood feeding. Experiments in a flight simulator (Ernst & Heisenberg 1999) suggested that *Drosophila* could be conditioned to discriminate vertical and horizontal bars when a heat source (a beam of infrared light) was used as punishment during conditioning. Making the assumption (quite basic, given the simplicity of the choices and the data currently available for this species; e.g. (Muir et al. 1992a, Muir et al. 1992b, Clements 1999)) that our experimental mosquitoes too can distinguish both patterns, our results do not support the hypothesis that they can use this information to return to a previous host. In any case, if they are not able to rely on the presence of these very different patterns, we tentatively conclude that *Ae. aegypti* individuals would also not be able to rely on other, more subtle, visual differences in nature.

In the series of experiments described, we used a range of unconditioned stimuli as 'punishment', from vibration, which would be also found in natural conditions, to a more severe stimulus – electric shock, routinely and successfully employed in learning

bioassays in *Drosophila*. As rewards, we allowed our subjects to blood-feed or detect human breath, both representatives of natural cues to host presence. The recurrent failure to find unequivocal and strong evidence of associative learning under these settings leads us to cast serious doubts on the possibility that *Ae. aegypti* could take advantage of associative learning under natural conditions. We still cannot rule out the possibility that our negative results derive from a failure to present the individuals with the right context in which they could learn, or to present situations where subjects would be under the proper physiological or environmental conditions for making the relevant choices based on their previous experiences. Additionally, it is also possible that Type II statistical errors prevented the detection of learning effects of small magnitude in individual tests (Cohen 1988). Although this possibility cannot be ruled out, it is weakened by the frequent observation of results in the opposite direction as that predicted by the learning hypothesis, coupled to the recurrent failure to detect significant learning effects in many and different experiments. Finally, there is the possibility that the behavioural repertoire of the mosquitoes was not representative of that observed in nature, given the long history of laboratory life (Wood 1977). Unfortunately, by the time the present experiments were performed we were prevented from using individuals collected from nature due to the aggravation of dengue epidemics in several countries.

However, if alternatively, our results do prove to reflect a lack of associative learning capacity in *Ae. aegypti*, we can discuss why it would be so. It is interesting to note that Mwandawiro et al. (Mwandawiro et al. 2000) found that the *Culex* species studied (*Cx. tritaeniorhynchus, Cx. vishnui* and *Cx. gelidus*) had an overall tendency to return to the host species upon which they had fed previously, but not *Ae. vexans*. Are *Aedes* species therefore less able to learn? Regarding *Ae. aegypti* there are reports of movements of up to near 1 km (Reiter et al. 1995) but this species does not typically fly further than a few hundred meters, and most of its displacements are within few meters (Service 1997, Clements 1999). Resting and oviposition sites and feeding sources are usually near their hosts (Clements 1999), and when these resources are not available, mosquitoes seem simply to disperse, as was found with oviposition behaviour (Reiter et al. 1995, Edman et al. 1998). Therefore, the presence of resources in a range of space where an individual can easily detect and identify them may not provide a strong selection pressure to develop and/or maintain learning abilities.

We could also speculate that learning in *Ae. aegypti* might not be relevant to assess the value of their resources on a long term (inter-gonotrophic cycle) basis, as odours associated with previous blood-feeds may not necessarily provide a more reliable and economical source of information about the host defensiveness than, for instance, the immediate evaluation made in each approach and probing stage. But, at least, as a diurnal host seeker in which the mortality risk associated with host defence is likely to be quite high, we would expect an avoidance behaviour toward individual hosts that show defensive behaviour (Gibson & Torr 1999), and here some form of short-term learning could play an important role. Our vibration or electrical shock experiments do not suggest it, and they are in agreement with the findings of Canyon et al. (Canyon et al. 1998) in which defensive host behaviour did not necessarily lead to a change of human host by individual blood-seeking *Ae. aegypti* females.

Chapter 3

LEARNING AND MEMORY IN MOSQUITOES:

A CRITICAL REVIEW

Chapter 4

PREDICTING THE RISK OF A GLOBAL DISTRIBUTION OF JAPANESE ENCEPHALITIS

Introduction

The dependence of vector-borne pathogens on the existence of appropriate eco-climatic conditions (Gubler et al. 2001) has allowed the inference of critical determinants of transmission cycles directly from the analysis of the environmental profile of those regions where the disease is present. This was made possible by the increasing availability of high-resolution data for the entire terrestrial surface through the use of remote sensing and the interpolation of climatic information obtained in ground-based meteorological stations, in conjunction with the development of powerful statistical and computational tools enabling the interpretation of such information in a biologically meaningful way. As a result, it is now possible to predict the areas in which diseases caused by these pathogens could potentially occur, and also achieve a better understanding of the dynamics of their transmission cycles (Rogers et al. 1996, Rogers 1998, Hay et al. 2000, Rogers 2000, Thomson & Connor 2000, Brooker et al. 2002, Kuhn et al. 2002).

The investigation of several vector-borne diseases through the use of these 'geographical models' (or 'spatial models'), as frequently referred to, has indeed already reached an advanced stage (Rogers 1988a, Rogers 1988b, Rogers & Packer 1993, Rogers 1998). Promising as it seems (Sabesan 2003), the use of such techniques for understanding the environmental factors underlying the endemic and epidemic manifestation of Japanese encephalitis has been, surprisingly, limited to only one study (Suwanee et al. 1999). This was based on a very small spatial scale and range of environmental conditions and has never been used for the production of predictive maps of the disease risk.

Japanese encephalitis (JE) is member of the genus *Flavivirus,* family Flaviviridae, a group of enveloped, positive sense, single stranded RNA viruses that are largely transmitted by vectors and associated with several emerging and re-emerging diseases of major medical, veterinary and economic importance (Monath 1990, Heinz et al. 2000, Gaunt et al. 2001, Gould 2002). JE is likely to have its origins in the Old World (Gould et al. 2001). In fact, there are some indications that it might have evolved only a few centuries ago from an Indonesian-Malaysian ancestor, possibly with its early roots in Africa (Solomon et al. 2003). Its demographic history was characterized, using

molecular phylogenies, by two patterns: a low or (null) rate of growth followed by a rapid expansion – probably due to the recent and massive increase in the human population (Zanotto et al. 1996) – at approximately at 85 years ago (Twiddy et al. 2003). The disease was described in Japan for the first time in 1870s, and the virus isolated in 1935 (Burke & Leake 1988, Solomon et al. 2000). The expansion of its geographic distribution to reach its current stage was probably assisted by infected migratory birds and vectors carried by wind currents (Tsai 1997, Hanna et al. 1999, Tsai et al. 1999, Gould et al. 2001, Solomon et al. 2003)

JE has become the most important cause of viral encephalitis in Asia, and of epidemic encephalitis worldwide, with an estimated annual incidence of ~175,000 cases diagnosed in humans (predominantly in rural and agricultural areas in China, Southeast Asia, India and the Pacific rim), of which approximately 43,750 result in death and 78,750 cause neurological sequelae annually (Tsai 1997, Tsai 2000). The manifestations of the disease range from a simple febrile illness to a severe meningomyeloencephalitis (Burke & Leake 1988) as well as a recently recognised polio-like acute flaccid paralysis (Solomon et al. 2000). Its main vectors are *Culex spp*. mosquitoes (from the *Culex tritaeniorhynchus* and *Culex vishnui* groups). Humans are not amongst the preferred source of blood of these vectors, but do get infected by occasional bites mainly when mosquito densities increase significantly after monsoon rains and subsequent pre-epidemic viral amplification by pigs. The enzootic cycle of the virus is developed mainly in birds, whereas its epizootic cycle is mostly based on domestic swine (Burke & Leake 1988, Tsai 1997, Solomon et al. 2000). Together with cattle, humans are considered dead-end hosts of the disease, as they do not develop sufficiently high viraemic levels to provide vectors with infective blood meals (and non-viraemic transmission has not been shown).

The influence of climatic factors on the geographic distribution of the disease and its epidemiologic patterns is thought to be critical (Burke & Leake 1988, Richman et al. 1997, Mellor & Leake 2000, Gubler et al. 2001, Endy & Nisalak 2002, Bi et al. 2003). At present the influence of such factors on the distribution of the Japanese encephalitis could be described by a temperate-tropical division (or cline). In northern temperate regions of its current distribution (e.g. northern China) JE epidemics take place mainly during summer, with the onset of cases in June, reaching its peak in

September, and disappearing in October. Further south, the pattern seems to be the same, with the only difference that the transmission season tends to start earlier and finish later in the year. However, in the tropical regions where JE is found, its epidemiological seasonal patterns are much more complex, with the density of vectors being directly associated with the monsoons and the transmission being year-round endemic (Richman et al. 1997, Endy & Nisalak 2002).

These climatic idiosyncrasies will be explored in detail below, but it is important to highlight at this early stage that socio-economic and historical factors are crucial for the understanding of the current distribution of this disease, and should also be taken into account when the geographic model is built and the results are interpreted. For instance, the highest incidence rates are observed in the humid and sub-humid ecozones of China. It is also in these zones – which comprise 50% of China's total area and where approximately 95% of its population live – where livestock is concentrated, with a biomass of approximately 25 metric tonnes/km^2, of which 36% is composed of pigs and poultry (Slingenbergh & Wint 1997). As already mentioned, pigs play an important role in virus amplification, not only because they develop high and sustained viraemia but also because their large, warm and hairless body surface attracts large numbers of mosquitoes every night (Richman et al. 1997). Additionally, they produce many offspring, thus providing the virus with a continuous supply of uninfected new hosts (Burke & Leake 1988, Solomon et al. 2000). In fact, where pig rearing is absent, such as in Bangladesh, the transmission risk of JE is much lower (WHO 2001). In turn, when swine-waterfowl is integrated into the same farming system, the risk is increased by bringing together natural and amplifying hosts, vectors and humans into close contact (Twiddy et al. 2003). The circulation of the disease is also stimulated by another human activity very common in many countries of that region: irrigation (Thornton et al. 2002). Together with the presence of dams, irrigation (used mainly for rice fields) increases the availability of ideal breeding places for mosquitoes and attracts the presence of water birds, known to be important hosts of JE (Burke & Leake 1988, Richman et al. 1997, Tsai 1997, Endy & Nisalak 2002). Due to their passive fishing strategy, herons for instance are mosquito-tolerant species leading to unusual seroreactivity prevalence (Edman & Spielman 1988).

At the same time, modernization of agriculture and livestock breeding can also help to break the transmission cycle of the disease through the use of pesticides and the centralization of livestock (Richman et al. 1997, Mackenzie et al. 2002, Solomon et al. 2003). Direct preventive measures, like vaccination, early diagnosis and treatment, together with improvements in living standards and sanitation, have also reduced the burden of the disease dramatically in countries like Japan, Taiwan and Republic of Korea and China (Tsai 2000).

Biological factors may also have contributed to the current distribution pattern of the disease. Heterologous immunity, i.e. the acquired immune protection against other species of pathogens elicited by a previous infection from a related species, could, for instance, limit the spread of JE to the west, given the presence of West Nile virus (WNV) in that region (see Figure 4.1). Yet, there is some evidence that WNV immunization only reduces the severity of the symptoms caused by JE, but does not prevent the infection (Tesh et al. 2002). Heterologous immunity could also be affecting the potential of Japanese encephalitis to spread in Australia (JE has recently emerged in that continent; Hanna et al., 1999; Van Den Hurk et al., 2001) due to the presence of the Australian Flaviviruses (Figure 4.1 and Chapter 6) in that continent. The same phenomenon could possibly affect its distribution in other parts of the world.

Even the differences in the transmission efficiency of the different vector species – among the regional strains of a same species – could be important to determine the geographic distribution of the disease. For instance, *Culex tritaeniorhynchus* from Pakistan were shown to be susceptible to JE, but their transmission efficiency was low (although high for West Nile) – in contrast with the strains from other regions (Burke & Leake 1988).

Given the existence of different sources of 'background noise' as mentioned (historical, biological and socio-economic), the justification for exploring the eco-climate factors underlying the distribution of the disease rests on the assumption that such factors indeed act as strong constraints (or instead promoters) for the development and completion of the arboviral cycle. For instance, Hales *et al.* (2002) found that vapour pressure alone proved a significant predictor of dengue fever risk, correctly classifying 89% of the worldwide distribution. The search for these critical parameters

might be relevant not only from a biological perspective, by helping to understand the evolution of the pathogen (see Chapter 6), but also from an epidemiological standpoint, by assisting the economic planning and targeting of national resources within the veterinary and public health service (Williams et al. 1992, Randolph 2000). Here I have explored the climatic factors associated with the current distribution of the JE virus with the use of GIS techniques. During this process, I have tried to examine the biological factors underlying the model outputs and check their consistency with known ecological data about the cycle of the disease.

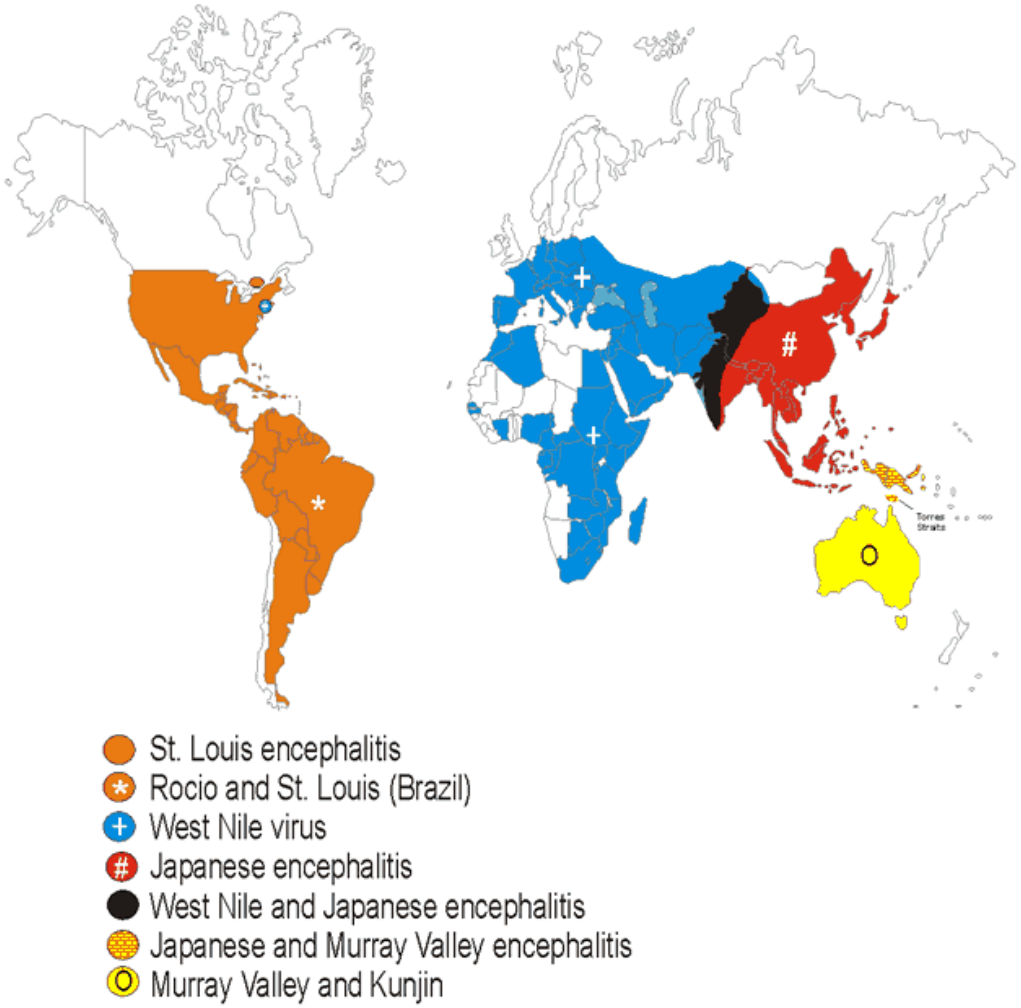


Figure 4.1. Geographic distribution of some virus species of the Japanese Encephalitis Serocomplex. Reproduced from the Centre of Diseases Control (http://www.cdc.org). The distribution of the West Nile virus is increasing widely (Campbell et al. 2002) from the introduction point in New York, as shown in the map (blue circle in North America).

Methodology

Training set

Records of virus presence and absence

The records of presence were obtained mainly through an extensive search in the bibliography and in other reliable sources of information available in the Internet – such as 'Promed' (http://www.promedmail.org) – for the period comprising the last 40 years. For each sample point, the name of the location (village, city and/or district) and country of incidence, as well as the year and source of collection were registered. I did not distinguish (despite keeping this information in the data set) whether the presence of the virus was detected from mosquito samples or serum samples from humans and other domestic and wild vertebrates. This was based on the assumption that the presence of a virus in a given location would indicate that the climatic conditions in that particular location were favourable for its existence, regardless of its association with a specific host. Each 'presence record' was geo-referenced using gazetteers available in the Internet[1]. The grain definition of the climate surfaces (see below) determined the size of the grid cells defining the global land areas. In those cases where more than one presence record fell into the same grid cell, the records were collapsed into a single 'presence point'. Information relating to the abundance of the viruses at sites was not used in the present analyses because presence records were collected by very different methods (e.g., mosquito traps or sentinel traps), mostly involving non-comparable density units. Also, abundance data were not available for many sites.

Absence points were obtained by defining a 20-degree (approximately 2000 Km) wide belt around the limits determining the distribution of Japanese Encephalitis proposed by Tsai (1997) and Tsai *et al* (1999; Figure 4.2). The Australian continent, where the presence of the virus has only recently been recorded – and where it is still not certain whether the virus will succeed in becoming established (Van Den Hurk et al. 2003) – was not considered. In this case, I preferred not to use either the presence

[1]http://fat-albert.alexandria.ucsb.edu:8827/gazetteer/
http://www.getty.edu/research/tools/vocabulary/tgn/index.html
http://164.214.2.59/gns/html/cntry_files.html

records for this continent or the absence ring (Figure 4.3). When compared to methods that make use of absence points spread around the world, this approach has the advantage of not considering those regions for which the virus is not present due to historical reasons (e.g. lack of dispersers carrying the virus to the region), but which might otherwise present ideal environmental conditions for its survival. By choosing absence points within 20^{o} of the expertise-based distribution, I have increased the likelihood of choosing those sites that have not developed the disease despite the likely and frequent exposure to the influx of organisms carrying the viruses. On the other hand, by excluding areas inside the existing distribution map, I lowered the possibility of generating 'false-negative data' (namely, false predictions of absence due to e.g., under-reporting and widespread immunization in some countries; Tsai, 1990, 2000). The collection of absence data from a belt around the limits of an existing distribution map builds on the methodology used by Rogers & Randolph (2000) to map the global distribution of malaria, with the difference that in the present study the records of presence were not obtained randomly from the distribution map, but instead from different specific records of presence.

Figure 4.2. Distribution of JE in Asia, 1970-1998 (Tsai 1997, Tsai et al 1999; image obtained from http://www.cdc.gov/ncidod/dvbid/jencephalitis/map.htm).

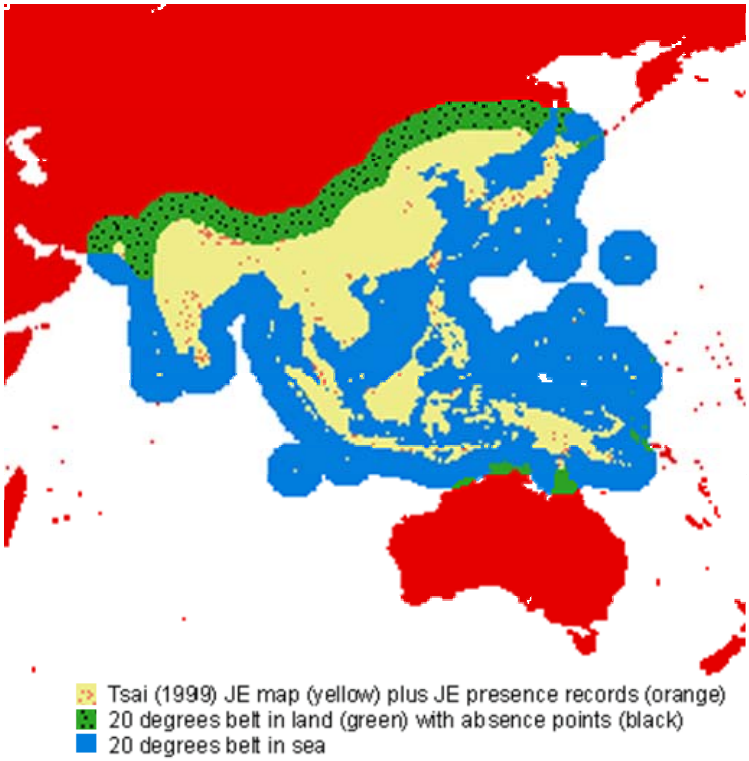


Figure 4.3. Grid map with digitalized distribution of Japanese Encephalitis (based on Tsai *et al*. 1999), showing the 20 degrees belt from which absence data was collected.

The existence of univariate and multivariate outliers (defined here as those values departing from the mean by more than 6 standard deviations) was checked and removed for each of the categories (presence/absence).

Environmental variables

Environmental variables were obtained from worldwide climate maps (excluding the Antarctic continent) generated by the interpolation of climatic information obtained from ground-based meteorological stations. The mean climate surfaces were available for the period from 1961 to 1990, with a monthly temporal resolution and 0.5º (latitude) by 0.5º (longitude) spatial resolution (New et al. 1999). The climatic variables used were: ground frost frequency (frs); precipitation (pre); mean, maximum and minimum temperature (tmp, tmx, tmn); vapour pressure (vap); wet-day frequency (wet) and diurnal temperature range (dtr). Primary variables of precipitation, mean temperature and diurnal temperature range were interpolated from extensive meteorological records using angular-distance-weighted averaging of anomaly fields for the acquisition of spatially contiguous surfaces. The secondary

variable of vapour pressure was interpolated where available, otherwise it was calculated from primary variables. The minimum and maximum monthly temperature estimates were calculated from the original climate surfaces by subtracting or adding, respectively, half the diurnal temperature range from the mean monthly temperature (New et al., 1999).

Each climatic variable was pre-processed with Fourier analysis, to allow the corresponding temporal series to be summarized in new climate surfaces containing the seasonal parameters associated with each variable (Rogers & Williams 1994, Rogers et al. 1996, Rogers 1998, Rogers 2000). This process essentially smoothes out the inter-annual variability, while leaving the seasonal characteristics intact. By this method, the variability of the monthly time-series is partitioned into uncorrelated components of cycles that are repeated up to three times per year I used the following parameters, extracted for each variable from the Fourier analysis: mean value of variable (a0), the amplitude of the annual (a1), biannual (a2) and triannual (a3) cycles and the maximum (mx) and minimum (mn) values of the combined amplitudes (sum of a1 plus a2 plus a3). Phase variables (reflecting the timing of occurrence of cyclic events as captured by the Fourier decomposition) were not considered since the timing of phenologically equivalent events would be very different throughout the continental scale used in the present study (Rogers et al. 1996, Rogers 1998).

To explore the potential interference of demographic factors on the model, I included a gridded map of the global distribution of the human population, with estimates adjusted by the United Nations for 1995 (CIESIN et al. 2000), in the variable selection process. In addition, due to the hypothesized correlation between the distribution of Japanese encephalitis and the distribution of pigs (one of its main hosts), I have also included a gridded map of pig densities around the world (FAO 1996).

Definition of Clusters

Because the distribution of spatial data can greatly depart from normality and biological systems may respond non-linearly to gradual changes in the climatic parameters defining the distribution of a species (with different covariances between critical variables), both the presence and absence data sets were partitioned into

clusters, with each cluster representing a separate multivariate description of environmentally similar parts of the JE distribution range (Rogers 1998, Rogers 2000).

The climatic variables were first scaled to correct for the discrepancies in the magnitudes and the variability of the variables, which would strongly bias the calculations. For all presence and absence data, values of each variable were therefore scaled to the interval [0,1] by the method of 'ranging' (Sneath & Sokal 1973, Milligan & Cooper 1988), by which the smallest value is subtracted from each value and the result is divided by the range between the smallest and the biggest value.

In addition, for the clustering process I did not use those variables whose distributions departed excessively from normality. I also eliminated the excess of redundant information that would be present with the inclusion of highly correlated variables (Pearson correlation > 0.75), since the uncritical use of variables has been shown to distort the clustering results (Milligan & Cooper 1987, Milligan & Cooper 1988).

Euclidean metrics were used to compute the distances between pairwise combinations of sample points within each category (presence/absence), since the basic Euclidean distance treats each variable as equally important in calculating the distance. Data were linked in the hierarchical cluster tree based on the increment of the total within-group sum-of-squares as a result of joining groups. The decision of how many clusters each category (i.e. presence and absence) should contain was based on the analyses of the resulting dendrograms obtained using the Ward linkage algorithm (Legendre & Legendre 1998). Such dendrograms were visually inspected to find the most cohesive groups, namely those with the smallest intra-group distances that were separated by the largest inter-group distances. The decision was assisted by the visual inspection of bivariate scatterplots to determine how well the clusters accounted for non-linearity in the data set, and which number of clusters best normalized the variables' distribution as shown in univariate histograms. In regard to this latter process, I have balanced the need to obtain a multi-normal distribution against the need to avoid describing any group by only a few observations, as that could lead to a distorted characterization of the resulting covariance matrices (Rogers 2000).

Variable selection

A forward stepwise algorithm was used to select the climatic variables based on their ability to increase the accuracy of the model to classify the training set data into presence or absence categories. The classification was conducted using maximum likelihood methods, by which the posterior probabilities P(i|*x*) with which each pixel 'x' belongs to each of the training-set defined clusters, 'i', were calculated as shown in the equation below (adapted from Rogers et al 1996, Rogers 1998, 2000):

$$P(i \mid x) = \frac{p_i \mid C_i \mid^{-1/2} e^{-D_i^2/2}}{\sum_{g=1}^{g} p_g \mid C_g \mid^{-1/2} e^{-D_g^2/2}} \qquad [4.1]$$

where:

- p_i is the prior probability of each pixel 'x' belonging to each cluster 'i';
- $|C_i|$ is the determinant of the covariance matrix of the cluster 'i', necessary for correcting cases where the covariance matrices is not homogeneous;
- D_i^2 is the Mahalanobis distance between each pixel '*x*' and the centroid of the cluster 'i'. The Mahalanobis distance (Mahalanobis 1936) is the covariance-adjusted distance between the centroids of two multivariate distributions, or – as it has been used in the present case – between a sample point and a centroid;
- e is the base of the natural logarithm.
- the denominator is the sum of all probabilities obtained for '*x*' across all classes in the training-set, and is used to rescale the measurement of probabilities to a same range, between zero and one.

Prior probabilities were kept constant (= 1/sum of number of presence and absence clusters) for all clusters, since I did not have any data suggesting that the probabilities should be biased towards a particular group. The posterior probabilities with which each pixel belonged to each of the presence clusters were calculated. The largest value was taken as the final posterior probability of a pixel being classified as belonging to the presence category P(Pres|*x*). This was based on the assumption that each presence cluster represents an equally 'optimal' environmental situation for the existence of the virus. Therefore, to calculate the probability of its existence and circulation in an unsurveyed environment, the cluster of presence that is environmentally closest to this area is considered.

The increasing accuracy of the prediction as each additional candidate variable was included was measured by the area under the curve (AUC) of receiver operating characteristic (ROC) plots (Cumming 1999). ROC curves plot the proportion of presences predicted correctly (true positive cases) versus the proportion of absences predicted incorrectly (false positive cases) at a number of classification thresholds. The area under these curves can range from 0 to 1, with values higher than 0.5 indicating performances that are better than random (Fielding & Bell 1997). Under this non-parametric method, the predicted probability of each observed record of presence to be classified as a presence [probability of true positives, P(Pres|*pres*)] is compared with the predicted probability of each observed absence record to be classified as presence [possibility of false positive, P(Pres|*abse*)]. A score of one, half or zero is attributed in each comparison in which the probability of occurrence at a known presence is respectively larger, equal or lower than the probability of occurrence at a known absence. The sum of all scores is then divided by the total number of comparisons (i.e., number of observed presences multiplied by the number of observed absences), so that the resulting value for AUC ranges from 0 to 1.

Variables were selected until there was no further significant improvement in the accuracy of the training set classification. Each new variable included in the model leads the environmental profile of the training-set data to be more precisely defined, therefore attaining a higher intrinsic accuracy in the classification of the data. Nevertheless, this increase in accuracy needs to be balanced against the danger of over-fitting the model to the training-set data, reducing the likelihood that a new data point can be considered 'similar' to either the presence and absence profiles, therefore diminishing the generality of the model. The decision on the number of variables to be included in the model was thus made by analysing the curve relating the 'number of variables selected' and the correspondent AUC values. Other measures of model accuracy were also employed to assist this decision (Rogers 1998, Rogers et al. 2002): sensitivity (proportion of observed presences correctly predicted), specificity (proportion of observed absences correctly predicted), and the kappa index (proportion of agreement once random agreement is removed; Cohen, 1960).

The selection process was conducted once with climatic variables only, and a second time with the log-transformed data of human population and of pig density.

This second trial was aimed at examining the impact of these latter variables in the discrimination of the presence and absence data in the model.

Predictive maps

By using the explanatory variables selected, the final posterior probability of each pixel of the worldwide land surface to be classified as belonging to the presence category P(Pres|*x*) was calculated in the same way as that explained for the variable selection model. Predictive maps were obtained separately for each cluster of presence, with the information corresponding to each cluster being subsequently merged into a final map. Although the final result is the same as that obtained by initially considering all presence clusters together, the present procedure enables the clear detection of the influence of each of the clusters in the final predictions produced.

Range of environmental resemblance used to produce the predictive maps

When generating a risk map, it is important to address the trade-off between the desired environmental breadth (and consequently geographic extension) for which the predictions of presence or absence will be made and the reliability of such predictions. In general, the larger the difference between the environmental profile of an unsurveyed area and that of the training-set data, the less reliable a prediction might be (given the low resemblance of the former with the environmental values defining both the presence and absence of a species). One way to deal with this is to set cut-off values, or inclusion criteria, which determine the maximum allowed distance between the multivariate centroids of the training set and the new data point (Rogers 1998, Rogers 2000).

I used the following procedure to set a cut-off value that maximized the likelihood of producing appropriate predictions. The Mahalanobis distance between the centroids of each cluster and its sample points were calculated, both for the presence and for the absence data set, and the largest within-cluster distance ($D_{sample\text{-}centroid}$) registered for each cluster (excluding 5% of the extreme data). In a second stage, the distances between the new pixel and the centroids of the clusters of the training-set ($D_{new\text{-}centriod}$) were calculated to determine to which cluster this new data point belonged. Finally, a prediction was produced if the latter distance ($D_{new\text{-}centroid}$) was

shorter than the within-cluster distance ($D_{sample\text{-}centroid}$) of the cluster to which the unsurveyed point was closer. This range is referred to as 'safe range' here, to distinguish it from the other ranges of inclusion that are investigated in Chapter 5.

As mentioned, the choice on whether to use one or another range implies an important trade-off between the reach and the reliability of the predictions. If employing a less stringent limit range, the area for which predictions are obtained is increased, but the gain in information is less trustworthy, since the probabilities are calculated on a range of distances beyond that associated with the empirical data. Furthermore, the results get extremely unstable for less stringent limits, as it will be shown in Chapter 5. The justification for using different ranges and the implications of each choice will be explored in detail in the following chapter.

The final 'merged' maps (all clusters together) were produced, again, on the assumption that each presence cluster represents an equally 'optimal' environmental situation for the existence of the virus. Therefore, to calculate the final probabilities of presence in this combined map I retrieved the highest probability of presence for each pixel shown in the partial predictive maps.

Model assessment

In order to verify the fit of the models, part of the training set is not included in the modelling process but kept aside to test the accuracy of the predictions. Nevertheless, when the data set is small, all data must be used in the modelling process, otherwise the covariance matrices will be ill-characterized and the likelihood that the model will perform badly in assigning the new data points to each of the defined categories will increase significantly (Rogers 2000). In this case – and in fact in many others even when additional data are available – no statistical validation of the method is possible, as I shall discuss in detail in the relevant section.

Software

All spatial analyses and statistical tests were conducted using custom-written programs that I developed on Matlab specifically for this study. Some graphics were generated also with the assistance of Matlab's inbuilt functions, the Matlab Mapping toolbox and the Richard Strauss toolbox for Matlab platform (Strauss 2003).

Results

Training data set (records and clustering)

Presence data

A total of 134 grid-cells, out of the 185 records of presence, were used. This reduction was due to the fusion of those records falling into the same grid cell. The presence records that were available at the geographic resolution required were not evenly distributed in the area of incidence of this disease, as inferred in Figures 4.3 and 4.4. Instead, some of the regions were more intensely surveyed, such as Nepal, the Indian subcontinent, Tibet, the southeast of Asia, the Indo-Malaysian region, Korea and Japan. In addition, there were very few geographically accurate descriptions from other areas where JE is known to be prevalent, such as in China. Indeed, under-reporting and widespread immunization in some countries, together with the frequent lack of precise geographic descriptions of the locations where cases were observed or data collected, usually act by limiting the size of data sets suitable to be used in spatial modelling. No outliers were detected.

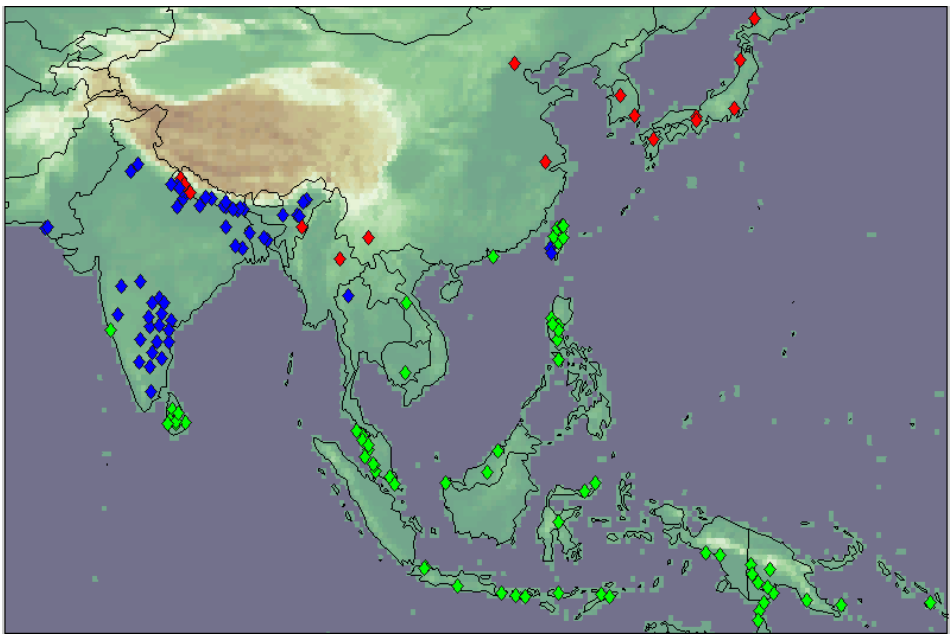


Figure 4.4. Distribution of the records of presence of the Japanese encephalitis virus. The different coloured points represent the different clusters to which the records belong. Background colour composition represent elevation.

The data were partitioned into three clusters. Such a partition was highly supported by the classification shown in the dendrogram (Figure 4.5), with the nodes

dividing these clusters located at the top of the dendrogram clearly detached from all of the remaining lower nodes. The clusters also captured well three major distinct eco-climates within the distribution of JE.

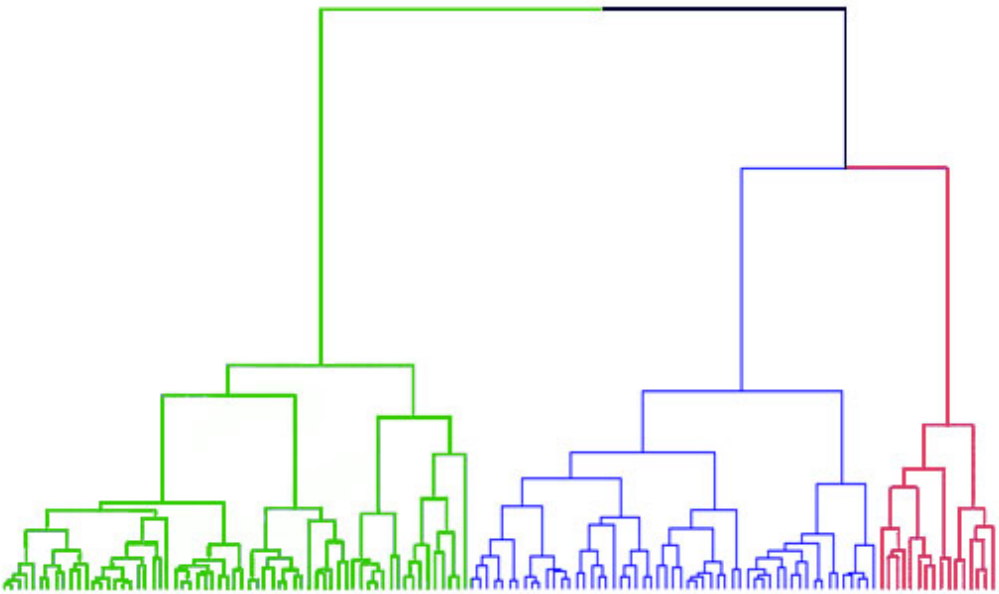


Figure 4.5. Hierarchical classification tree of the records of **presence** of the Japanese encephalitis virus as a function of their environmental similarity. The eco-climatic variables were scaled to correct for the discrepancies in the magnitudes of measurement (ranging method). Euclidean metrics were used to compute the distances between pairwise data. The Ward clustering algorithm was employed.

The records obtained for the '*Indian-subcontinent*'[2] cluster (blue points in Figure 4.4) are distributed mainly along the deciduous forests that extend throughout the centre and the eastern coast of the Indian subcontinent, as well as in the southern territories of the Himalayas, along the Ganges and Brahmaputra valleys basins. The central region of the delta of the Indus River in Pakistan is also included in this group. Those records coinciding with the humid and warm environments housing the rainforests of Southeast Asia, Sri Lanka and Indo-Pacific region were grouped into another cluster, the '*Indo-Pacific*' group (green points in Figure 4.4). This cluster also includes the records from Goa, and from the northern part of Taiwan and Hong-Kong. Finally, the '*Nipo-Chinese*' group (red points in Figure 4.4) is associated with more temperate or even colder regions within the distribution range of the virus, which embraces in the north the Japanese islands, Korean peninsula and Beijing, and in the

[2]It is important to highlight that this and the other names used to refer to the geographic locations are not accurate descriptions of the geographic limits shown for each cluster. Together with the key of colours, they are used only as easy and simple references to facilitate the identification of the same categories of records throughout the text and figures.

south locations in higher altitudes of Nepal and in the mountain range separating India from Burma.

Figure 4.6 shows the distribution of the environmental variables for each of these defined clusters. It is clear from the charts that, for most of the variables, the clusters are well defined and their individual distributions tend to approximate to a normal distribution in many of those instances where the total distributions were bimodal or trimodal (such as for the variables dtr-a0, dtr-mx, pre-a0, tmn-a0, vap-a0, wet-mx, wet-a0, wet-mn). It is interesting to notice that in the histogram (bottom right corner of Figure 4.6) showing the distribution of Mahalanobis distances for each cluster (i.e., the multivariate distance between the cluster centroid and each of its sample points), the most extremes values (i.e., highest level of environmental dissimilarity between a presence record and the centroid of its cluster) belong to the Indo-Pacific group. Indeed, one of these extreme values corresponds to a presence record (located in Papua New Guinea) for which the highest altitude in the group was registered (1500m, Figure 4.6, elevation) where the values associated with diurnal temperature range are very high (dtr-minimum 15.2 ^{o}C and dtr-maximum 16.8 ^{o}C; figure 4.6, see plots of dtr-min and dtr-mx). The other two extreme values (as measured by the Mahalanobis distance) obtained for this cluster result mainly from the high indices of precipitation associated with those records: one of the records is located in the West-Coast of India (Goa), a region affected by the south western monsoon from June to September (Figure 4.6 pre-mx; 29 mm of mean maximum values of precipitation per day); the other record is again located in Papua New-Guinea, at a site for which the value of the mean maximum values of precipitation per day was also high (17.1 mm/day; Figure 4.6, see pre-mx plot). None of the three records just described was removed from the data set, not only because they were represented by values lower than the critical value established for the removal of outliers, but also because the eco-climatic parameters (precipitation and temperature) underlying their corresponding Mahalanobis distances were represented by values suitable for the existence of the JE virus and the completion of its cycle. Therefore, there would be no biological justification for the elimination of these records.

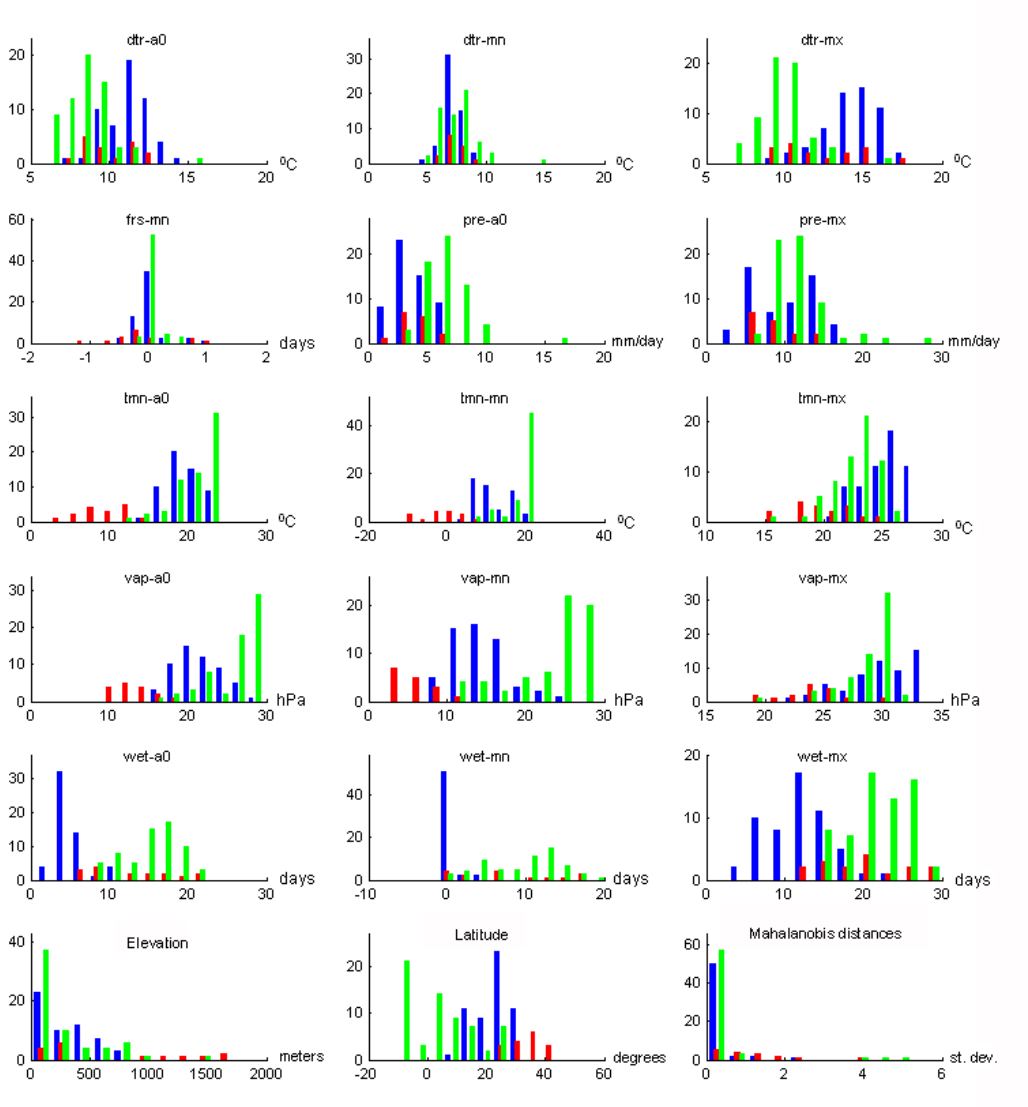


Figure 4.6. Histogram showing the distribution of the environmental variables for each of the defined clusters of presence. The eco-climatic variables shown correspond to those used in the clustering process (additionally, elevation, latitude and Mahalanobis distances – bottom row – are included). Blue, red and green bars represent the clusters referred to as 'Indian-subcontinent', 'Nipo-Chinese' and 'Indo-Pacific' in the text. Main variable codes: dtr: diurnal temperature range; vap: vapour pressure; tmn: temperature; pre: precipitation; frs: number of frost days; wet: number of wet days. Secondary codes associated to each variable: a0: mean; mx: maximum, mn: minimum. Units of measurement corresponding to each variable are shown in the plots.

Absence data

Seven outliers were identified and eliminated from the analysis (black points in Figure 4.7).

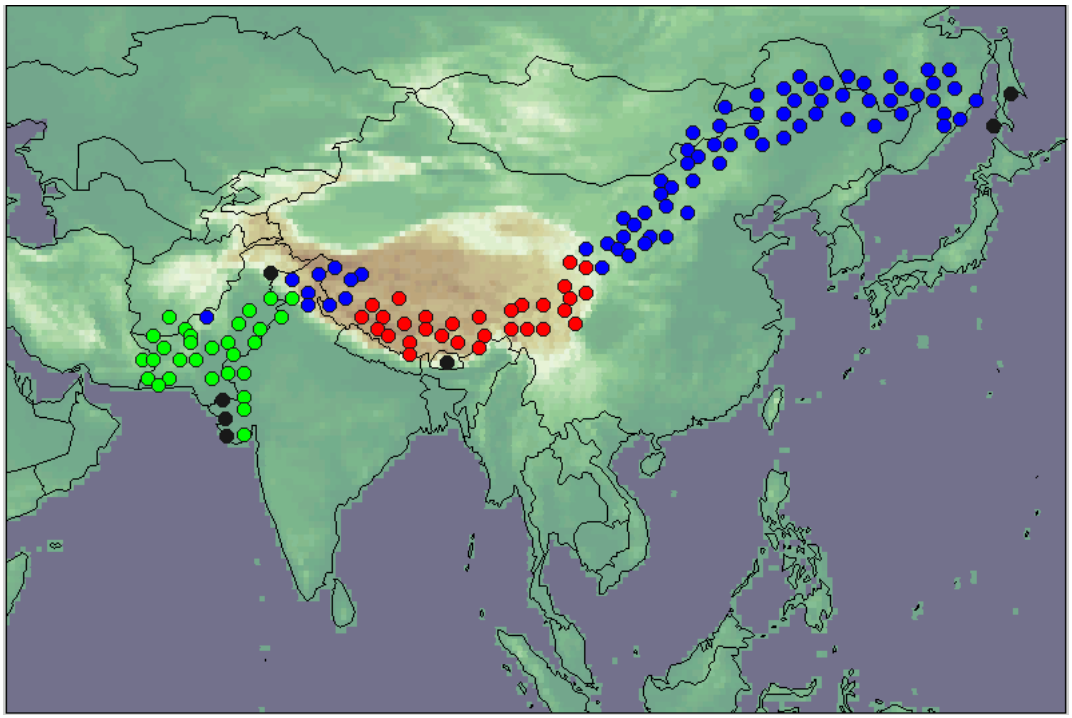


Figure 4.7. Points of estimated absence for Japanese encephalitis virus. The colours (blue, red and green) associated with the data points represent the absence cluster to which they belong. Black points represent outliers that were excluded.

By analysing the branches in the left hand end of the dendrogram of Figure 4.8, one can notice that it would have been possible to partition the remaining records into two, three or four major clusters. The partition into only two clusters, however, resulted in a poor definition (as well as lack of normality) of the variables describing wet-day frequency and vapour pressure when compared to that obtained with the partition of the data into 3 clusters (Figure 4.9). Partitioning the data set into four clusters, conversely, did not lead to any improvement in the degree to which the variables' distribution approached normality that was large enough to justify the loss of definition in the covariance matrices that follows every partition. The three clusters obtained also seemed to represent consistently the major climatic domains associated with the absence belt: the distribution of the records included in the first cluster (represented in blue in Figure 4.7) extends from the evergreen forests of maritime Siberia (near the

Artic circle) to the Mongolian steppes and the Gobi desert in China. The cold-climate profile of this 'Siberian-Mongol' cluster is thereby consequent mainly on the association of many data points to high latitudes (mostly beyond 35^{o} north, Figure 4.9, Latitude), although a separated group of points is located further south, around the region of Kashmir in the western Himalayas. The 'Tibetan' group is also associated with cold climates and tundra vegetation, but its profile derives instead from its presence in regions of high altitudes (more than 3000 m), e.g., in the central Himalayas (Figure 4.9, elevation). The last group – referred to as '*Pakistanis*' here (green colour) – is represented by a dry (Figure 4.9, pre, vap, and wet variables) and hot (Figure 4.9, tmn variables) environmental profile consisting mostly of records located in Pakistan, but also in the south of Iran and the northwest of India.

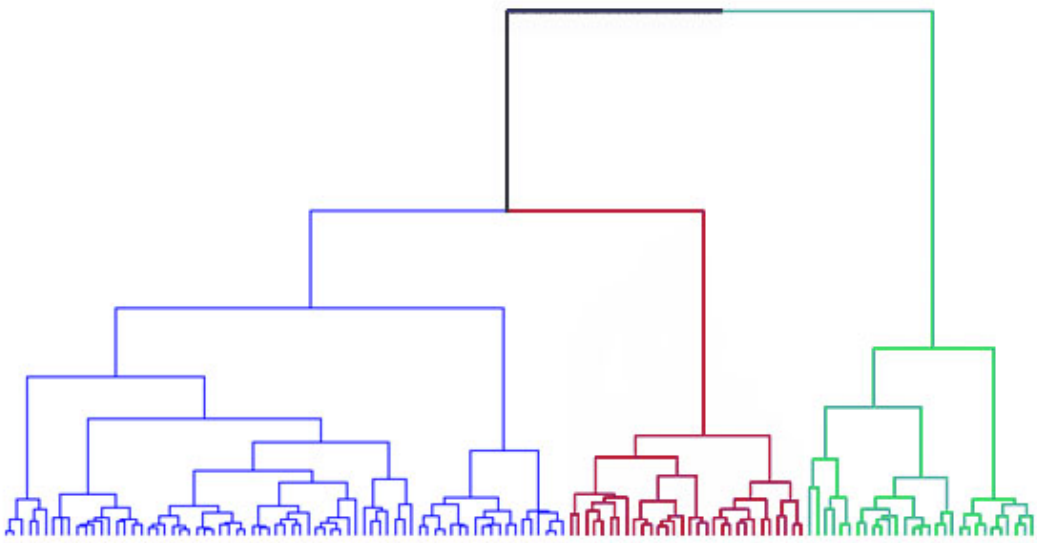


Figure 4.8 Hierarchical classification tree of the **absence** sites of the Japanese encephalitis virus as a function of their environmental similarity. The eco-climatic variables were scaled to correct for the discrepancies in the magnitudes of measurement (ranging method). Euclidean metrics were used to compute the distances between pairwise data. The Ward clustering algorithm was employed.

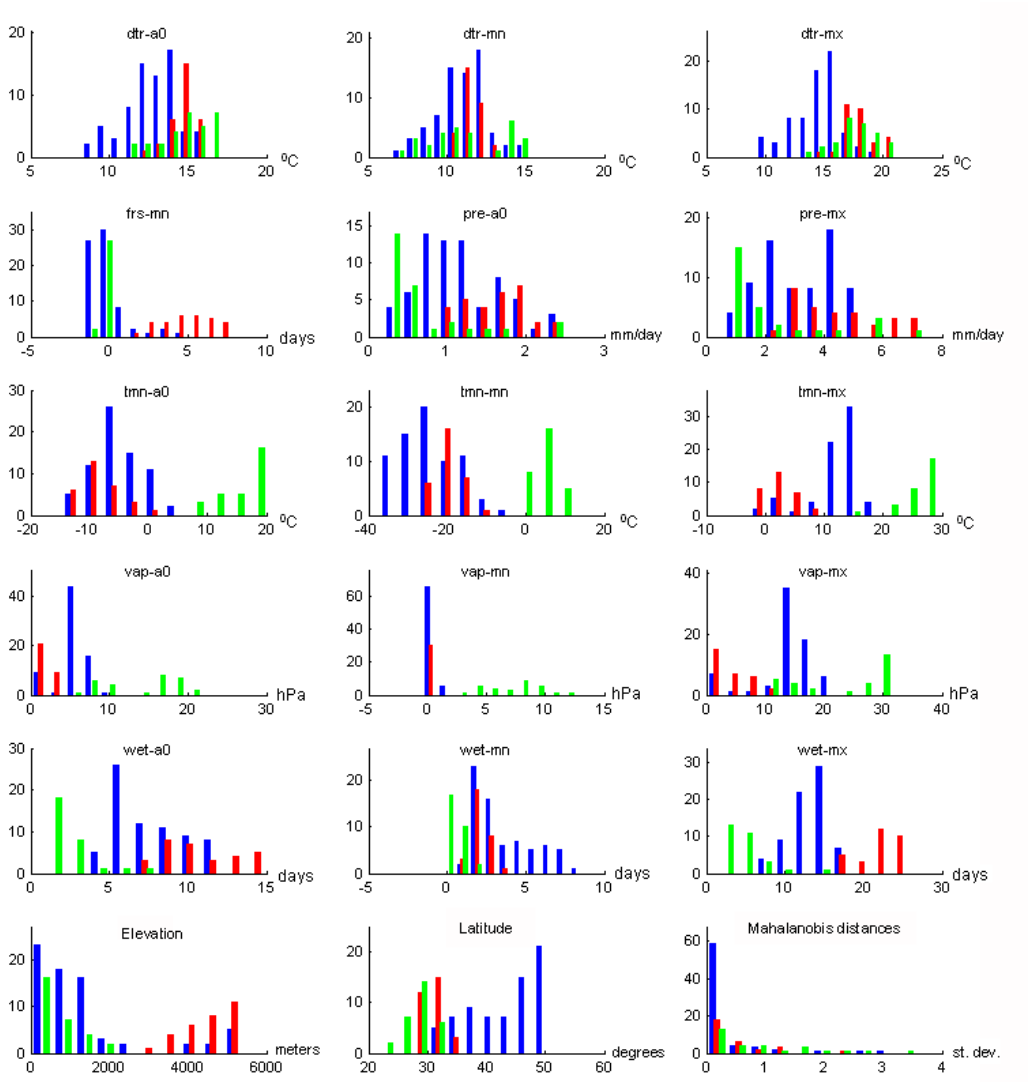


Figure 4.9. Histogram showing the distribution of the environmental variables for each of the previously defined clusters of absence. The eco-climatic variables shown correspond to those used in the clustering process (additionally, elevation, latitude and Mahalanobis distances – bottom row – are included). Blue, red and green bars represent the clusters referred to as 'Siberian-Mongol', 'Tibetan' and 'Pakistanis' in the text. Main variable codes: dtr: diurnal temperature range; vap: vapour pressure; tmn: temperature; pre: precipitation; frs: number of frost days; wet: number of wet days. Secondary codes associated to each variable: a0: mean; mx: maximum, mn: minimum. Units of measurement corresponding to each variable are shown in the plots.

Variable selection

During the process of variable selection the agreement between the observed and predicted classification of the records (in presence or absence classes) was very high even when the model contained only one explanatory variable. Figure 4.10 depicts the relationship between number of variables included in the model and the level of agreement between the observed and predicted results.

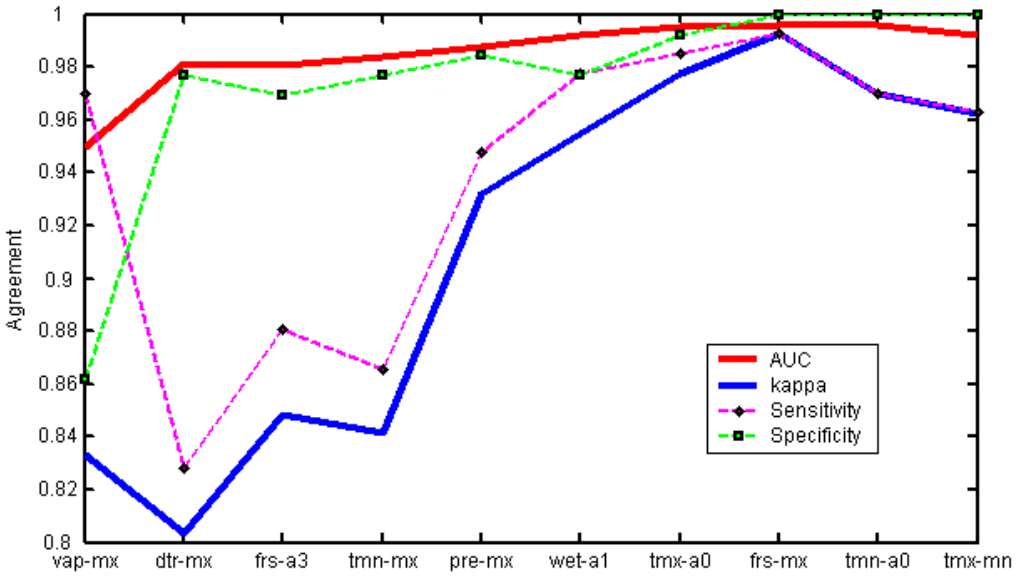


Figure 4.10 Curves used to assist the decision about the number of variables to be included in the model. Values reflect the effect of the model when the variable in the horizontal axis and all the others in its left are considered. The AUC (area under the curve) index is a threshold independent measure of model accuracy and it was the function that was maximized in the variable selection process. Sensitivity: proportion of presence points of the training-set correctly predicted. Specificity: proportion of absence points of the training-set correctly predicted. Kappa: proportion of specific (non-random) agreement. All possible measures ranges from zero to one but kappa, which ranges from –1 to +1. A value equal to one in this case is – similarly to the other measures – an indication of perfect agreement, and values of zero or less suggest a performance neither better nor worse than random agreement.

The model first selected the variable describing the maximum values of vapour pressure. This variable alone already leads to a value of AUC equal to 0.95, correctly classifying 97% of the records of absence and 86% of the records of presence of the training-set data. This result is easily understood by looking at the virtually non-overlapping ranges of vapour pressures in presence and absence clusters shown in Figures 4.6 and 4.9, and also at the coloured surface built with the values of this variable (Figure 4.11-1 (vap-mx)). In this latter figure, the green areas (representing

higher values of maximum vapor preassure) in the south and southeast of Asia and Indo-Pacific archipelago match fairly well the distribution of the presence records (diamonds). The absence points (circles), in turn, are located mostly within the limits corresponding to the blue regions (representing lower values of maximum vapor preassure) of this map. All false positive cases (Figure 4.11-1, red circles) correspond to the '*Pakistanis*' absence cluster, which is associated with a region that, athough characterized by low precipitations levels, receives most of this hydrological input during summer (e.g. in Punjab plains, Herrel et al., 2001), resulting in annual records of maximum levels of vapor preassure as high as those observed for the presence clusters (see Table 4.1, Figure 4.12). In contrast, false negative points (Figure 4.11-1, red diamonds) are located either in areas with higher elevations (above 1300 meters), or in the most northly region of the disease distribution, where colder temperatures do not allow the air to be saturated with high levels of humidity.

Table 4.1 Climatic variables used to describe the worldwide distribution of the Japanese Encephalitis virus, and their correspondent mean value (± s.d) for each group within the presence and absence categories.

	Presence Clusters			**Absence Clusters**		
	Indian Sub-Continent	Nipo-Chinese	Indo-Pacific	Siberian-Mongol	Tibetan	Pakistanis
Vapour pressure *Maximum* (hPA)	29.97 ± 2.94	23.66 ± 3.00	28.71 ± 2.30	13.98 ± 5.30	3.76 ± 3.32	23.45 ± 8.17
Diurnal temperature range *Maximum* (ºC)	14.41 ± 1.87	12.27 ± 2.57	9.65 ± 1.69	14.45 ± 1.98	17.92 ± 1.46	17.56 ± 1.85
Frost days *Amplitude of the tri-annual cycle* (days)	0.16 ± 0.23	0.89 ± 0.54	0.02 ± 0.06	2.17 ± 0.82	0.83 ± 0.45	0.84 ± 0.48
Minimum Temperature *Maximum* (ºC)	25.06 ± 1.73	19.71 ± 2.53	22.48 ± 2.00	12.27 ± 4.37	2.70 ± 2.73	25.72 ± 2.86
Precipitation *Maximum* (ml/day)	9.74 ± 3.99	8.39 ± 2.80	11.39 ± 3.68	3.18 ± 1.21	4.45 ± 1.50	2.11 ± 1.87
Wet days *Amplitude of the annual cycle* (days)	5.23 ± 1.88	5.34 ± 2.91	4.99 ± 2.85	4.40 ± 0.91	9.49 ± 1.45	1.41 ± 0.83
Maximum temperature *Amplitude of the annual cycle* (ºC)	30.75 ± 2.23	19.07 ± 4.07	29.53 ± 2.75	7.75 ± 4.05	6.59 ± 2.90	30.62 ± 3.51
Frost days *Maximum* (days)	1.65 ± 2.19	17.54 ± 8.67	0.43 ± 0.97	31.04 ± 1.41	31.07 ± 0.76	8.27 ± 6.32

Similarly, the effects of the inclusion of further variables into the model on the predictions produced are shown in the remaining maps (2-8) of Figure 4.11. In the case of the second variable selected (diurnal temperature range, dtr-mx), a high maximum daily temperature range of approximately 17.6^{o}C (Table 4.1) helped to discriminate the absence cluster 'Pakistanis' (which had been poorly classified when the model included only one variable) from the clusters of presence (Figure 4.11-2; notice that some of the absence records that were incorrectly classified in map 1 – red circles – become yellow in map 2). Such an increase in the accuracy of the predictions of absence (specificity) was, however, achieved at the cost of diminishing the sensitivity of the model (Figure 4.10) to classify correctly those data points of presence that presented higher values of dtr-mx in relation to the their vap-mx values (see increase in number of false negatives, red diamonds, in Figure 4.11-2). To exemplify how the observed improvement in the specificity of the model is achieved, consider the bivariate plot of Figure 4.12, which depicts the relationship between the first and second variables selected (bivariate scatterplots showing how all variables selected interact to discriminate between the presence and absence data are available in the appendix 1). The first variable selected (vap-mx) largely discriminates the absence points (circles) at the left side of the scatterplot from the presence points (diamonds) located mainly at the right. The inclusion of a second variable (dtr-mx) envelops even further the presence clusters, excluding from them the records of absence from the 'Pakistanis' cluster (green circles in the upper corner on the right) located above the records of presence on the right side of the plot.

Yet, with the inclusion of the third (frost days: amplitude of the triannual cycle) and further environmental variables (maps 3-8 of Figure 4.11), both the sensitivity and specificity of the model tend to converge to an optimum value that reaches its maximum with the inclusion of eight variables (Figure 4.10). Notice also that the number of red records – false positive and false negative predictions - is progressively reduced from maps 3 to 8 in Figure 4.11. Beyond that, all measures of agreement apart from specificity show a decline. Given such a consistent trend, the final model was produced with eight variables.

All the variables that were selected in the modelling process are listed in Table 4.1. Following such selection, the eco-climate of the current distribution of the

Japanese encephalitis (in general terms, without specifying the multidimensional interaction between the variables) can be compared with the eco-climate of absence sites, as follows:

1. Higher levels of maximum values of air humidity (vap-mx) (with the exception, as already mentioned, of the Pakistanis cluster) and of precipitation (pre-mx);
2. Lower extreme differences between the minimum and maximum daily temperatures (dtr-mx);
3. Lower oscillation in the triannual cycles of the number of frost days (frs-a3), especially compared to the Siberian-Mongol absence site;
4. Warmer temperatures with higher minimum values (tmn-mx). This was also captured by the fact that presence regions showed fewer frost days (frs-mx).

The later the inclusion of the variable in the selection process, the lower its contribution to the overall discrimination of presence and absence sites. When the model already includes a high number of variables, the specific effect of the inclusion of a further variable is not easily identified, for it may act through a number of multidimensional interactions that are not captured in bi or tri-dimensional plots. In the case of the amplitude of the annual cycle of the wet days (wet-a1), it is not possible to infer any discriminant role in the classification of the data set by looking at its means values (Table 4.1). Still, the effect of the inclusion of wet-a1 in the model can be noticed through the resulting improvement in the classification of five records of presence belonging to the '*Indian-subcontinent*' cluster (Figure 4.11-6).

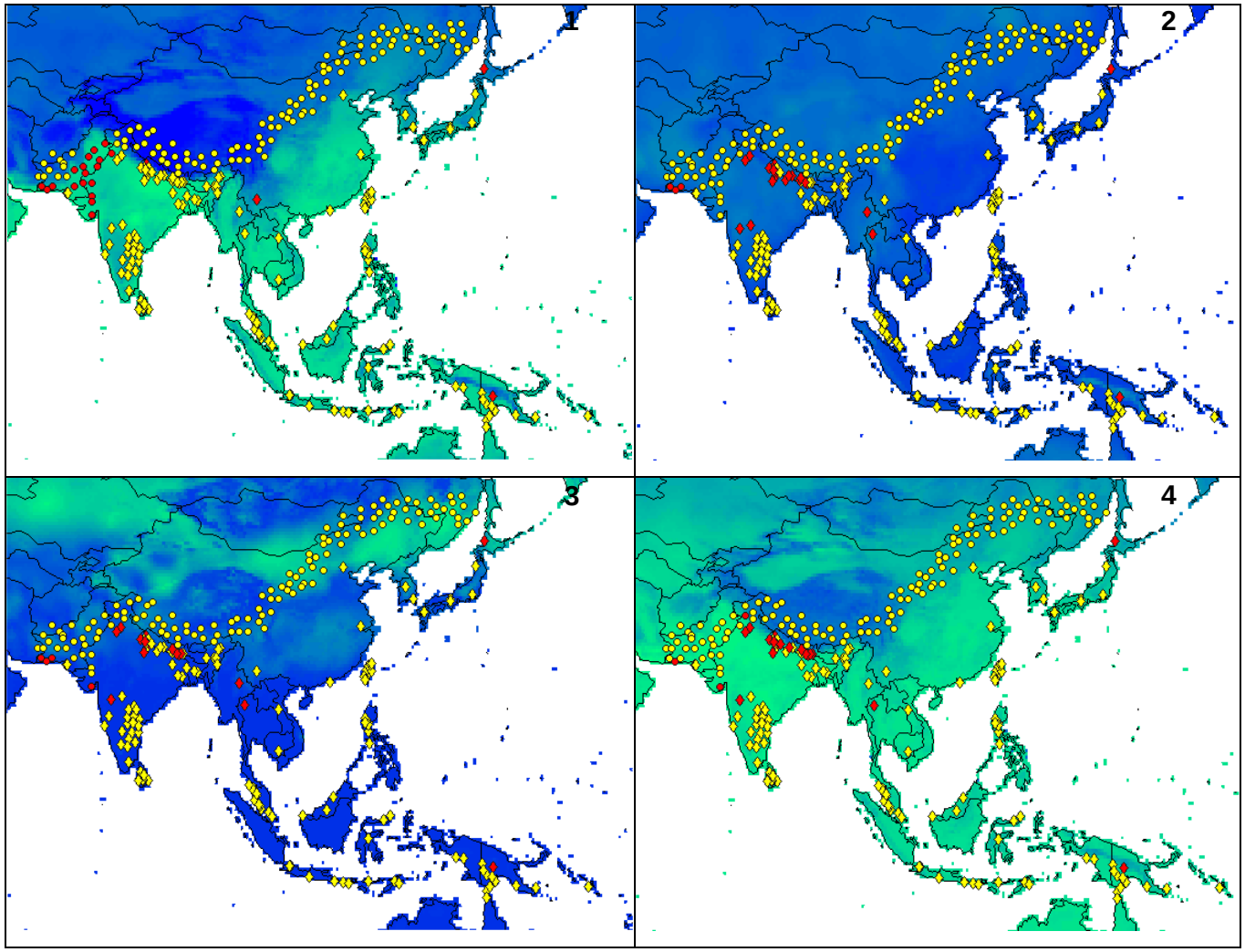


Figure 4.11 (1-4). Classification of the records of presence (diamonds) and absence (circles) with the inclusion of eco-climatic variables in the model. Symbols in yellow and red were correctly and incorrectly classified, respectively. In each of the maps the palette of colours was scaled to range from the minimum value (blue) to the maximum value (green) observed for the corresponding variable. Map1: first variable selected: vapour pressure (Maximum, as measured in hPA); map 2: diurnal temperature range (Maximum, measured in ^{o}C; second variable selected); map 3: Frost days (amplitude of the triannual cycle; third variable selected); map 4: minimum temperature (Maximum value, oC; fourth variable selected).

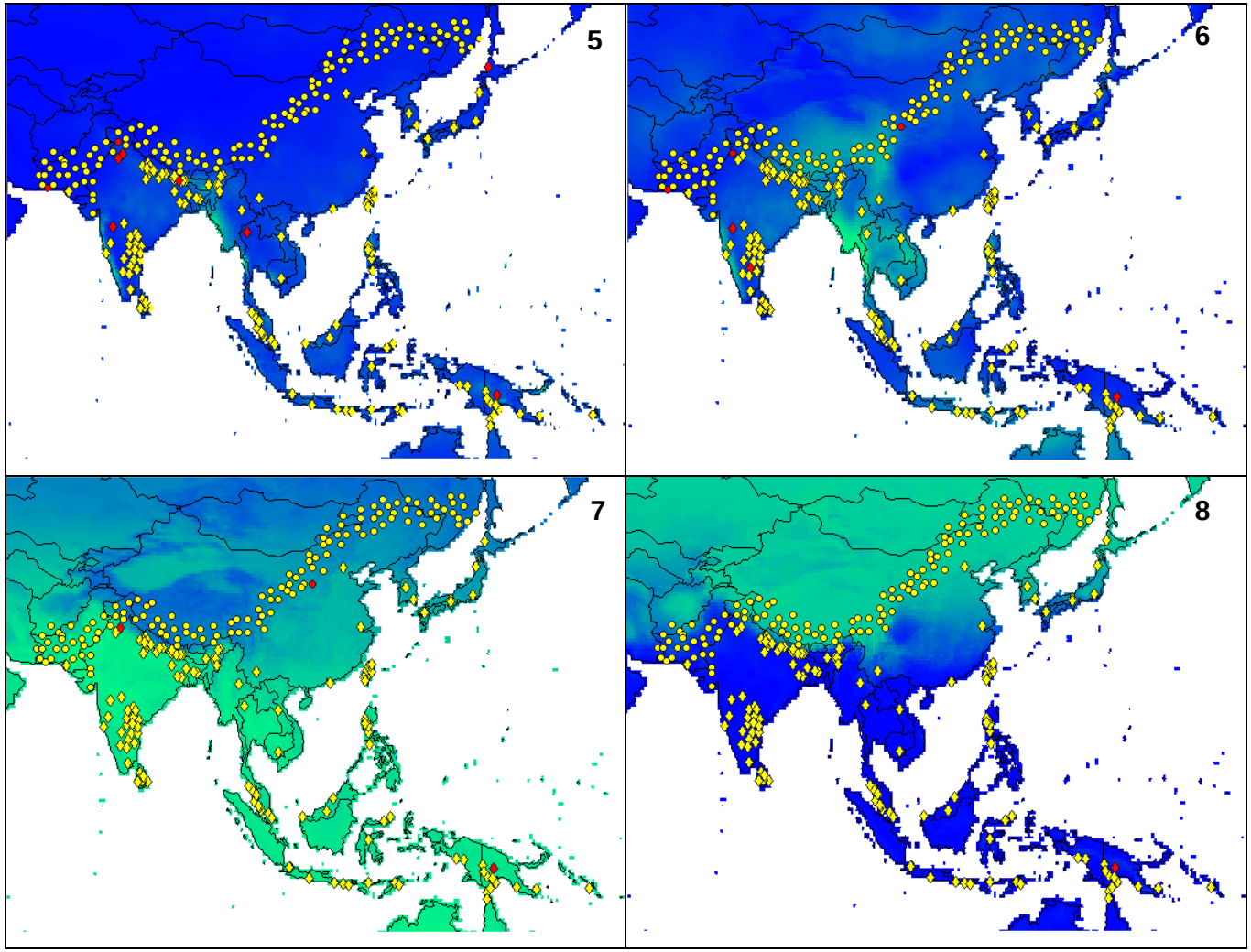


Figure 4.11 (cont.: 5-8). Classification of the records of presence (diamonds) and absence (circles) with the inclusion of eco-climatic variables in the model. Symbols in yellow and red were correctly and incorrectly classified, respectively. In each of the maps the palette of colours was scaled to range from the minimum value (blue) to the maximum value (green) observed for the corresponding variable. Map5: fifth variable selected: precipitation (maximum, as measured in ml/day); map 6: wet days (amplitude of the annual cycle, measured in days; sixth variable selected); map 7: maximum temperature (amplitude of the annual cycle; seventh variable selected); map 8: Frost days (minimum, days; eighth variable selected).

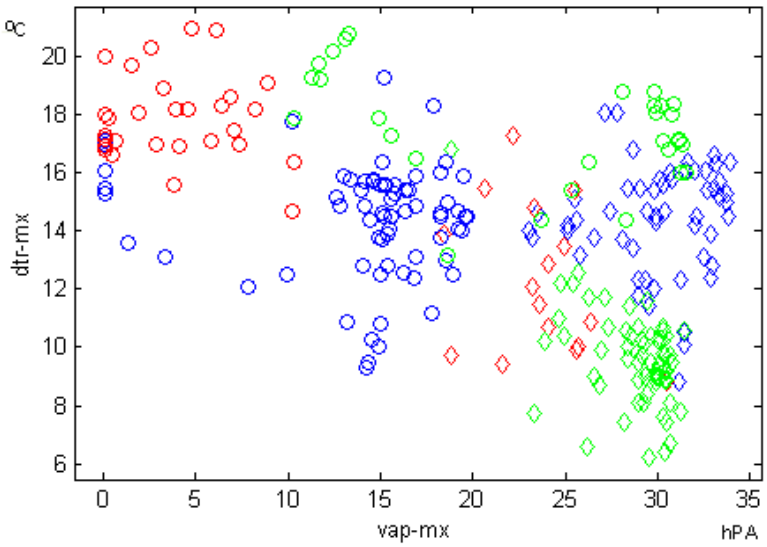


Figure 4.12. Bivariate plot of the first (vap-mx) and second (dtr-mx) variables selected. Blue, red and green diamonds represent the presence clusters 'Indian-subcontinent', 'Nipo-Chinese' and 'Indo-Pacific'. Blue, red and green circles represent the absence clusters 'Siberian-Mongol', 'Tibetan' and 'Pakistanis'.

I have also conducted the process of variable selection once more by considering distribution maps of both human population density and estimated pig density as potential explanatory variables, but the model did not select any of these variables. These additional results therefore indicate that it was not necessary to control for either of these demographic factors in the modelling process.

Predictive map

The predictive partial (cluster-based) maps generated by considering each of the clusters of presence separately are shown in Figure 4.13 (1-3). Figure 4.14 shows the final 'merged map', built from the fusion of maps 1-3 of Figure 4.13. In all maps, the green and red areas indicate those areas predicted to be risk-free and risk-prone, respectively.

If the analyses are restricted to the current region of incidence of Japanese encephalitis (Asia), it is possible to observe that the positive predictions made in each of the cluster-based maps are generally produced for those places situated nearby the geographic regions where the cluster is itself found (an expected result if the environmental similarity between two sites is inversely related to their geographic distance). For example, the positive predictions produced considering the 'Indian-

subcontinent' cluster are indeed distributed mainly across the Indian-subcontinent (Figure 4.13-1). Likewise, positive predictions for China, Korea and Japan are obtained mostly through the influence of the second cluster domain ('Nipo-Chinese', Figure 4.13-2). Finally, this is also the case for all the islands of the Indo-Pacific archipelago and all the countries of Southeast Asia, for which positive forecasts of the presence of the disease are predominantly observed in the partial map generated with the third cluster of presence ('Indo-Pacific', Figure 4.13-3). When the partial maps are merged, the resulting map (Figure 4.14) matches fairly well the distribution of Japanese Encephalitis in Asia as predicted by Tsai *et al*. (1997; Figure 4.2). The regions within this distribution for which no prediction was made (following the criterion of inclusion imposed by the adoption of the 'safe range' of environmental resemblance) correspond to the east coast of the bay of Bengal in Myanmar, the continental and some coastal areas of the southeast of China, north of North Korea and the highlands of New Guinea. In explicit disagreement with the map of Tsai (1997), marine Siberia and the northern Japanese island of Hokkaido were predicted as having a low risk of developing the disease. The lack of additional records of presence for these regions, coupled with their eco-climate similarity to the 'Siberian-Mongol' absence cluster profile, are responsible for their incorrect classification. Most of the area (91%) within the distribution limits shown in the map produced by Tsai (1997) was, nevertheless, correctly classified (and, of course, most of the training-set data, as the criterion maximized in Figure 4.10 and Figures 4.11, 1-8 was based upon this data set). Good as these high levels of agreement in the Asiatic continent are for confirming the internal consistency of the model used here, as well as for generating continuous predictive surfaces around the observed points of presence and absence, the positive predictions mentioned above relate mainly to those areas for which the potential incidence of the disease was already suspected. Therefore, the description of the results has special significance for the remaining, still unaffected, geographic regions of the world.

The global risk map (Figure 4.14) shows that, in the range of distribution beyond Southern and Eastern Asia, the eco-climatic conditions suitable for the existence of the Japanese encephalitis virus can be found in many regions of the Americas, Africa, Western Asia, Australia and the oceanic islands in the tropics. The 'Indian-subcontinent' cluster was responsible for the risk-positive predictions mainly

around the edges of the tropical regions (Figure 4.13-1). In America it included Florida, the Pacific and Caribbean coast of Mexico, the northern region of Paraguay and surroundings in Argentina and Brazil and the semiarid region in the northeast of Brazil. In Africa, positive predictions were obtained for an extensive area in the west of the sub-Saharan region, in addition to central Sudan, most of Mozambique and southwest of Madagascar. In the Middle East, positive predictions were obtained for the Red Sea coast, the east southern Arabic peninsula and the coast of Iran. Most of tropical Australia was also predicted as being climatically suitable for JE by this cluster.

The 'Nipo-Chinese' cluster seemed to be the one most largely affected by a possible over-fitting of the model, as it does not yield almost any prediction for world regions outside the surroundings of its own training data set, except for small areas in southeast of Brazil and Botswana. This is an important limitation of the model because this cluster represents the eco-climates of the presence sites situated in the north of the cold-months isotherm of 15^{o}C, therefore not generating predictions for the many regions where mosquitoes undergo a cold-weather dormancy in winter (Mitchell 1988).

The 'Indo-Pacific' cluster was responsible for most of the risk-positive predictions in tropical and equatorial areas in Asia and elsewhere (Figure 4.13-3). They include most of the oceanic islands in the tropics, the Caribbean region and Central America, areas of low altitude of the northern region of the South American continent, and the coastal, as well as some inland, regions of Brazil. In Africa, these predictions encompassed all regions surrounding the Gulf of Guinea, in addition to the Congo River basin, Madagascar and most of the tropical coast of the continent near the Indian Ocean. The coastal region of Oman was also included.

Finally, the areas for which a low risk of disease was predicted extend mainly along the northern limits of the distribution map of Tsai (1997), and also from Marine Siberia through central Asia, Tibet, north of China to European Russia. It also includes some regions in Scandinavia, in the Algerian Sahara, the North American Rocky Mountains and Alaska, and a great area in central and east Canada. The remaining surface of the map encompasses areas with environmental resemblance with the training set beyond the limits imposed by the 'safe range' for predictive purposes, and therefore it is not considered. Predictions (positive and negative) were thus obtained for

a total area of ~53.7 million Km^2, or approximately 37% of the global land surface considered (including all continents except Antarctica). Roughly half of this area (~32.2 million Km^2)– corresponds to regions predicted as JE risk-prone, where approximately 2.634 million people live (~47% of the human population; CIESIN et al., 2000). The area for which positive predictions were yielded outside the current known distribution of JE (shown in the map of Tsai 1997, Figure 4.2) – and therefore are regions where the disease is absent, but climatically is suitable for it – is still large (~21.5 million Km^2). Nevertheless still contains a much lower number of inhabitants (approximately 556 million of inhabitants) for the simple reason that the current distribution of the disease coincides with the most populated region in the world.

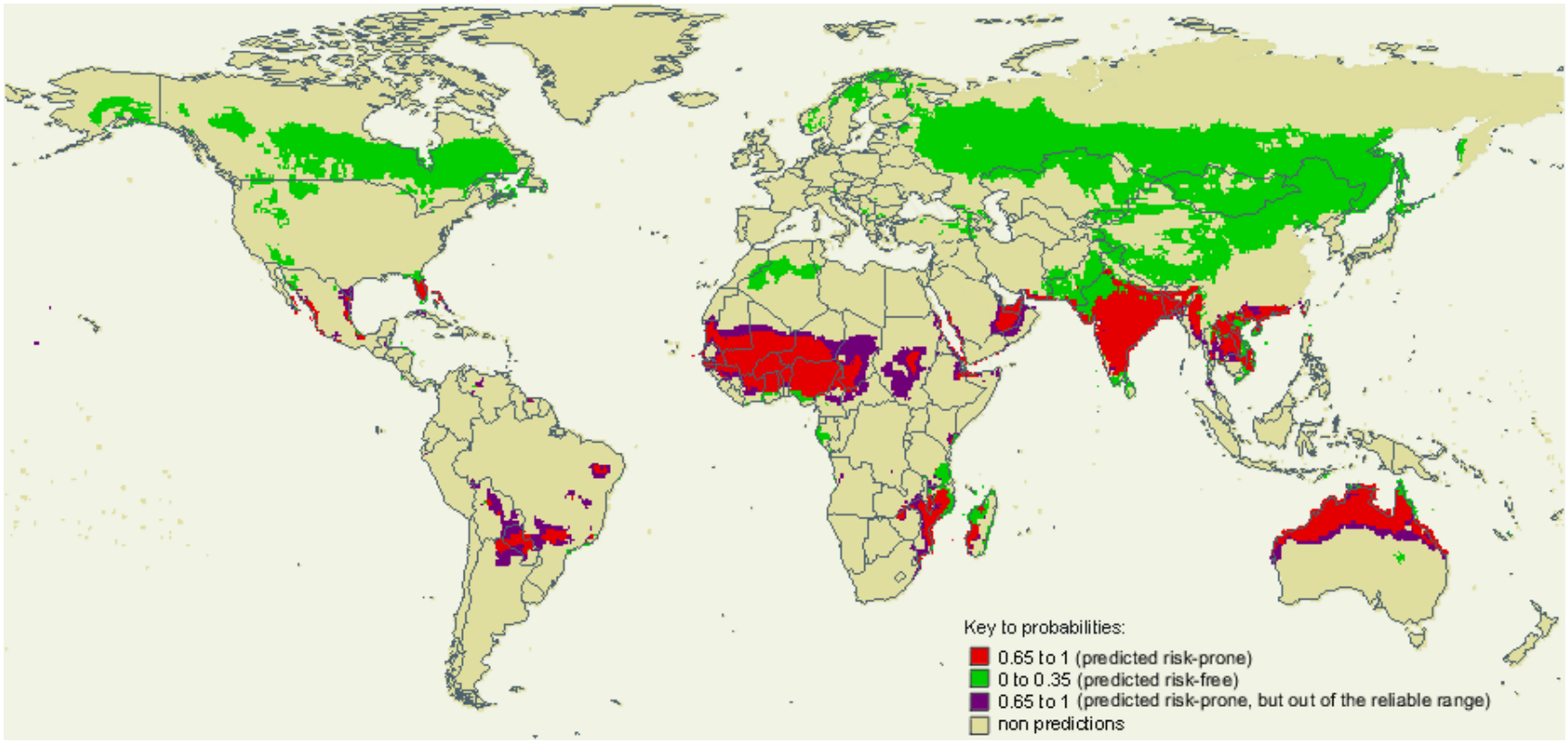


Figure 4.13 (1). Partial maps of posterior probabilities for the Japanese encephalitis virus as based on the '*Indian-subcontinent*' cluster. The predictions produced under the safe range of environmental resemblance (see text for explanation) are represented in red and green colours. The remaining colours correspond to those regions where no reliable predictions (i.e., out of the 'safe range') could be made.

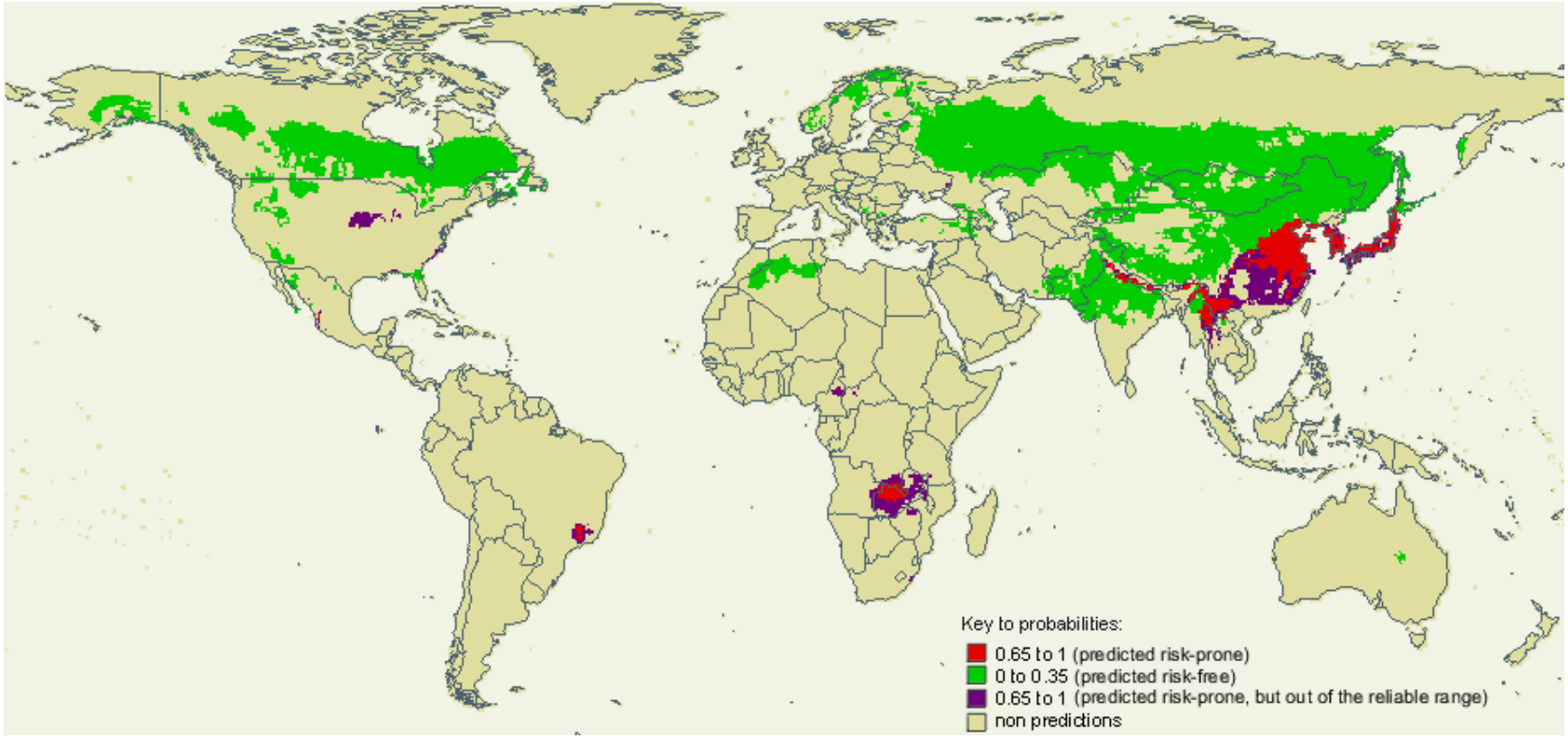


Figure 4.13 (2). P Partial maps of posterior probabilities for the Japanese encephalitis virus as based on the '*Nipo-Chinese*' cluster. The predictions produced under the safe range of environmental resemblance (see text for explanation) are represented in red and green colours. The remaining colours correspond to those regions where no reliable predictions (i.e., out of the 'safe range') could be made.

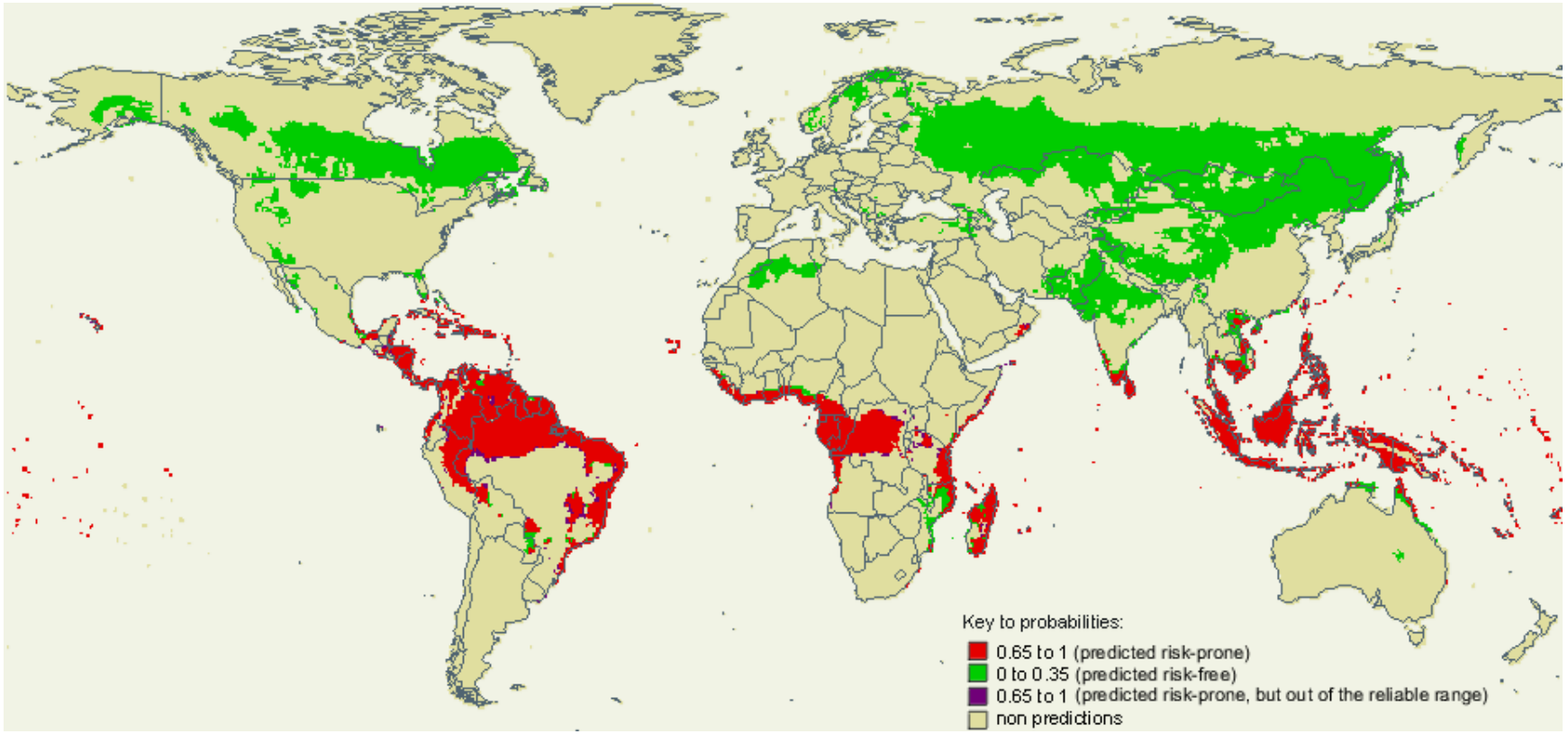


Figure 4.13 (3).Partial maps of posterior probabilities for the Japanese encephalitis virus as based on the '*Indo-Pacific*' cluster. The predictions produced under the safe range of environmental resemblance (see text for explanation) are represented in red and green colours. The remaining colours correspond to those regions where no reliable predictions (i.e., out of the 'safe range') could be made.

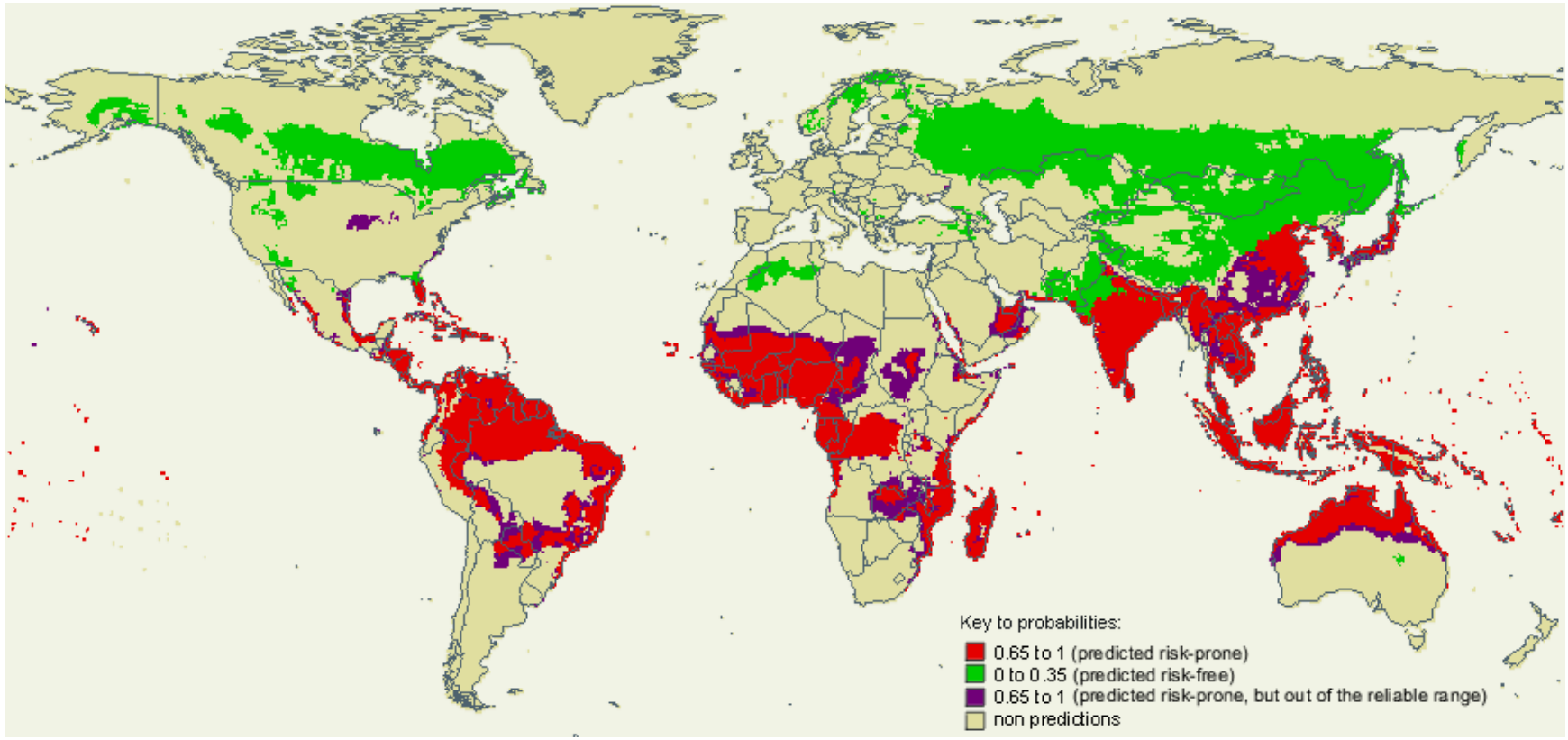


Figure 4.14 Final map of posterior probabilities. The predictions produced under the safe range of environmental resemblance (see text for explanation) are represented in red and green colours. The remaining colours correspond to those regions where no reliable predictions (i.e., out of the 'safe range') could be made.

Conclusions

Japanese encephalitis is not only a major health problem in most countries of Asia but, as a result of the wide range of its distribution, it also represents an important and potential threat for many areas around the world. Based on the assumption that regions at risk are those that are environmentally similar to that corresponding to the current distribution of this disease, the present analyses confirmed this assessment. Because of the relative environmental resemblance between the eco-climatic profile of various regions around the world and that where JE is currently present it was possible to forecast the risk of the disease in various areas outside the map produced by Tsai (1997), for which no records are available. The extent to which this potential risk will be realised in the future is impossible to forecast, given the mobility of such vector-borne viruses and increasing global connectivity (Rogers & Randolph 2003). Of the 32.1 millions of Km^2 included in the predictions, 67% were shown to be risk-positive. This area, inhabited by 556 million people (figures for 1995, CIESIN et al. 2000), corresponds to regions of the world where a large proportion of the human population still live below the poverty line (Thornton et al. 2002). Therefore, in some countries of Latin America and mainly in Africa, any measures to prevent the spread of yet another infectious disease and to provide medical care would most likely be hampered by a seriously deficient socio-economic and sanitary scenario. These countries are already unable to cope with the current pandemics (such as malaria and AIDS). Besides suitable eco-climatic conditions and a deficient socio-economic system, in Africa there is also a biological reason for concern: the most important vector of JE – *Cx. tritaeniorhynchus* – is not only present in this continent, but its distribution (WHO 1989) also overlaps to a great extent with that of predicted JE virus risk. Even other native mosquitoes species could potentially become effective vectors of the virus. In fact, it has been shown that seven mosquito species present in the United States (along with several domestic birds) are able to transmit the JE virus (Tsai 1997).

If on the one hand the predicted risk of the disease encompasses a very large area and its impact in terms of the proportion of the human population affected is worrying enough, on the other it is actually remarkable that such a surface is not even more extensive. In fact, because JE is adapted to a extremely broad range of

environmental conditions – from the eco-climates of marine Siberia to Australia, also reaching Pakistan in the west – it would have been reasonable to expect positive predictions to be widespread across much of the global land surface, with the exception of the deserts and high latitudes and/or elevations. Instead, besides the areas where no classification was possible, many regions were explicitly classified as unsuitable for the existence of the virus. This result suggest that, even though this disease is widespread across very different and diverse environmental spaces, key climatic factors might be involved and required in important stages of the transmission cycle.

Before proceeding, it is important to highlight that the results presented here should be viewed as an exploratory analysis and not as conclusive by themselves, given the impossibility of using independent data for assessing the model (Williams 1983). Although the number of records available was large enough to enable the use of half of them to test the model, these records could not be treated as independent units given their geographic proximity, as well as the lack of spatial alternation between the presence and absence records. Under such conditions, it is virtually impossible to distinguish the myriad of other factors, mainly anthropogenic (such as socio-economic, religious and historic factors) not accounted for by the present model that could underlie the high indexes of agreement observed. The prospect would be different if, instead, the presence and absence records were spatially dispersed in a mosaic formed by both of these classes. In such a case, the potential influence of confounding factors should be evenly distributed between these two categories, thus enabling the employment of a proper statistical test. However, the descriptions of the presence records currently available do not provide any indication as to possible locations (within JE current distribution; Figure 4.2) where the absence of the virus cannot be attributed to factors other than lack of climatic suitability (e.g., vaccination, absence of pig livestock, improvement of living conditions, etc). Inferring absence points within the current distribution of the virus would thus largely increase the likelihood of generating false-negative data (i.e., false predictions of absence). One alternative option would be to exclude from the modelling process those presence records located in distant sites (such as those contained in the 'Nipo-Chinese' cluster), and to use them in a subsequent stage to test the resulting model. Yet, this procedure would not be appropriate for the following reasons. First because it would provide us with only one

independent sample unit as a result of the same lack of interspersed spatial arrangement between the presence and absence points. Secondly, because the analyses presented here show that the eco-climatic variables interact in different ways in distinct regions (clusters) of the disease distribution (Rogers 2000, Brooker et al. 2002). Thereby we should not expect that the niche described on the basis of data collected in, for example, the 'Indo-Pacific' cluster would apply to the 'Nipo-Chinese' regions.

It is also important to consider the strategy used here to create the absence training-set data, as this is an essential aspect of the modelling process that can have a strong impact on the results. The definition of an absence belt around the nearest areas of presence areas in a scanned map, as originally implemented by Rogers and Randolph (2000), concentrates the collection of absence data near the boundaries of the disease distribution. By doing so, this procedure has the advantage of enabling the detection of the finer environmental differences defining the organisms' distribution and effectively preventing it from existing in a given location. As already mentioned, it also minimizes the risk of generating false absence predictions by not considering those sites that are located too far from the observed distribution, and which have not developed the disease due to historical reasons, such as lack of dispersers (otherwise, it would be the same case as having considered the USA as a proper absence site for the West Nile virus some years ago, before its spread in that country). Nevertheless, in case of species that are invading new habitats, or that live in habitats where some patches have become extinct, there is the serious danger of constraining unrealistically the definition of the niche of the species, thereby affecting the outputs produced by the model. Rogers and Randolph (2000) determined the sites of absence of *falciparum* malaria by defining a 10 degrees belt around the limits shown in a map of its global distribution. Because the distribution of the Japanese encephalitis in Asia might not yet be consolidated, I have increased this boundary to 20 degrees around the distribution map produced by Tsai (1997), but I did not project these absence data onto the Australian continent, where JE is clearly an invading species. The model therefore relies on the assumption that the boundaries of the distribution suggested by Tsai (1997) are stable to a certain degree. Nevertheless, as mentioned, there are reasons for suspecting that other non-climatic factors can be playing a vital role in the disease boundaries. Even though the variable representing pig distribution was not selected in the modelling process, one can

strongly suspect that the absence of this livestock in Pakistan due to religious beliefs might have a crucial role in disrupting the spread of the disease towards the west. Although the disease is not present or predicted in Pakistan, an exception is the occurrence of JE in the delta of the Indus River; its presence in this area might be associated with the high number of breeding sites for mosquitoes and the suitable environment for water birds provided by the extensive irrigation fields in this region. Furthermore, Pakistan and Australia have closely related species of flaviviruses that are well settled and may compete for available amplifying hosts (Figure 4.1; Hanna et al. 1999, Van den Hurk et al. 2001). In Australia there might be a reduced risk that JE becomes endemic due to socio-economic factors, the fact that few or no domestic pigs are kept, and the presence of a large number of poor hosts in the form of the marsupial fauna that may divert a high number of mosquito bites from any competent hosts (Van Den Hurk et al. 2003).

The model suggests that maximum vapour pressure is the most important factor underlying the current distribution of the JE virus in Asia. Humidity, as measured by vapour pressure, is high only where rainfall and temperatures are both high, conditions that are understood to promote the existence of breeding sites for the vector and to promote the replication of arboviruses (Reiter 1988, Mogi 1996, Mellor & Leake 2000, Solomon et al. 2000, Hales et al. 2002). The fact that the model selected maximum vapour pressure as the principle predictor variable may indicate that extreme values of humidity are indeed crucial for the circulation of JE. In tropical regions, the peak in the transmission of JE is concentrated at the end of the hot season, when high temperatures enable rapid virus replication and subsequent rains provide the vectors with suitable breeding sites (Mellor & Leake 2000, Solomon et al. 2000).

Still related to humidity, high levels of precipitation (pre-mx) are also associated with disease presence. Rainfall is indeed a major hydrological input for the maintenance of many ecosystems, and is therefore paramount in determining which ecosystems and species will be present in a given region. For transmission cycles that are dependent on mosquitoes, rainfall acts (together with other factors) by limiting the presence, absence, size and persistence of breeding sites (Mellor & Leake 2000). In general, dramatic hydrological changes – such as the extensive irrigation associated

with the beginning of the rice cropping cycle, or a massive rainfall in a semi-arid area – provoke an explosion in the density of the relevant *Culex* species (WHO 2001). By regressing climatic variables over annual incidence rates in a JE epidemic county for the period between 1980 and 1996, Bi *et al.* (1996) showed that elevated mean, maximum or minimum temperatures, together with high monthly rainfall, were associated with the transmission of the virus (Bi *et al*, 1996).

Maximum daily temperature range was also selected to give a better classification of the records corresponding to the Pakistanis cluster, with lower daily variations in temperature favouring the occurrence of the disease (Figure 4.11-2, table 4.1). In general, higher minimum temperatures were also associated with JE presence. Because temperature is unquestionably a critical factor in all stages of the transmission cycle, it is reasonable to assume that mild daily temperatures are required for the maintenance and completion of the cycle. The observation that the distribution of Japanese encephalitis is strongly associated with this climatic variable would have, if confirmed, implications in face of global warming forecasts: temperatures are expected to increase more during the night than during the day (IPCC 2001), thus leading to an overall reduction in the values of daily temperature range and consequently an expansion of the geographic areas presenting suitable conditions for the existence of this disease.

Although a large extent of the boundaries of the current distribution of JE is located in the transition to the colder temperatures of the high altitudes of the Himalayas (Figures 4.4 and 4.7), it is interesting to notice that the third selected variable represents an indirect measure of the number of frost days. The actual variable selected was that relating to the triannual oscillation in the number of frost days (frs-a3) – a factor that might be interacting with the other variables selected and indeed allowing the circulation of the virus mainly in some areas of Nepal. The reason why no direct measurements of frost days were selected by the model might be related to the presence of JE across a wide range of climates, including some where frost days are not uncommon in winter (e.g. northern records of presence shown in Figure 4.11-8). The following example illustrates how the transmission cycle of the disease is not impeded by the cold conditions in winter. The distribution of the main vector of JE, the mosquito

Cx. tritaeniorhynchus, covers most of the distribution of the virus (WHO 1989). Because in the cold winters of maritime Siberia and north of Japan adults of this vector hibernate before having their first blood meal, temperature could act by limiting the time available for the incubation of the virus in the vectors, thus restricting their ability to transmit the disease. However, in these cold regions other vector species, such as *Ae. japanicus,* probably maintain the virus over winter by transovarial transmission in drought resistant eggs (Mitchell 1988, Mellor & Leake 2000), enabling it to start the following summer as an already infected vector.

Given the multivariate nature of the analyses, it is less clear how the remaining variables influence the distribution of the disease due both to the multiple interactions with the other variables already selected and to the very specific role that each further variable included in the model has in improving the classification of only a few sample sites. The impact of the variation in the number of variables included in the model on the predictions about the distribution of the disease is explored in detail in Chapter 5.

Finally, the modelling technique employed here allowed me to investigate the major eco-climatic domains underlying the spatial distribution of the JE virus. The partition of the data into environmentally similar clusters is also useful to examine the phenotypic and genotypic plasticity of this species, allowing it to adapt to local conditions. For instance, based on a well-established genetic distribution of JE, Solomon et al. (2003) proposes the Indonesia-Malaysia region as the most likely candidate for the origin of the JE virus (from an African ancestral virus), as it is in this area where most genotypes, including the most ancient ones, are represented. If we take the distribution of genotypes IV and V (the most ancient) together with that of genotype II, the resulting region coincides with that shown for the '*Indo-Pacific*' cluster defined here. Conversely, the recently evolved genotypes would have their distribution extending mainly across the areas occupied by the '*Indian-subcontinent'* and '*Nipo-Chinese'* clusters, with genotype III present in both areas and the distribution of genotype I overlapping mainly with that of the latter cluster. The fact that the distribution of these serotypes is still not consolidated (Solomon et al. 2003) does not allow me, at this stage, to test whether the phenotypic plasticity associated with each genotype is defined by the combination of the environmental profiles of the clusters,

but it is at least encouraging that the definition of the clusters obtained here is so far not in conflict with the genetic data.

Because one of my concerns here was to explore the existence of potential causal relationships, the methodology was restricted to the use of a technique that allows me to explain the contribution of each environmental variable to the output produced, as well as the interactions between variables and all the assumptions that have to be accepted (Rogers 2000). The predictions generated are thereby explicit and measurable, enabling the biological interpretation of the results. It is important to highlight that this transparency of the model should allow its refinement and further confirmation when more data become available. The quest for a deeper validation of the results and the exploration of the method led to the next chapter, in which the results obtained in the present study will play a pivotal role.

Chapter 5

THE ROBUSTNESS OF DISCRIMINANT ANALYSIS TO THE OPEN CHOICES INVOLVED IN SPATIAL MODELLING: A CASE STUDY WITH THE JAPANESE ENCEPHALITIS VIRUS

Introduction

Over the last decade, discriminant function analyses (DFA) have been used increasingly in the production of predictive maps of the potential distribution of diseases in spatial epidemiological studies (Rogers et al. 1996, Robinson 2000, Rogers 2000). Specifically, non-linear discriminant analysis based on maximum likelihood classifications has proved to be an extremely flexible and useful tool, since the predictive power of this technique is not limited by the assumption of common covariances and multivariate normality between the multivariate spaces defining both the presence and absence data set (hereafter referred to as 'training-set data'). This is specially important because organisms might not live in a random subset of the entire environmental space, but in portions with specific conditions that not only differ from where the species is absent, but also differ between the various localities inhabited by its various populations, each of which requires a separate multivariate description of its climatic conditions (Rogers et al. 1996, Rogers 2000).

Such a technique has been used for investigating the distribution of the African trypanosomiases and their vectors (Williams et al. 1992, Rogers et al. 1996, Robinson et al. 1997, Rogers 1997, Hendrickx et al. 2000, Rogers 2000), the African and worldwide distribution of malaria (Rogers & Randolph 2000, Rogers et al. 2002), the transmission of tick-borne encephalitis in Europe (Randolph & Rogers 2000) and the presence of blue tongue disease in the Mediterranean basin (Baylis et al. 2002). Its use had been criticized because not because the qualities of the method per se, but because its application with "*lack of independent validation of model predictions, uneven interpretation of kappa statistics, and over-reliance on assumptions of absence based on data not well suited to establish absences with confidence*" (Levine et al. 2004). I consider these problems "external" (the method can be applied avoiding this problems), and I want to focus on the possible "internal" problems of DA.

The advantages defended by this method is that allows a straightforward and transparent test of the biological assumptions underlying each stage of the modelling process, as well as the opportunity of expert input. However, this latter feature may also be a source of setbacks, as researchers are confronted with several choices during the process of data analysis that unwittingly insert a degree of subjectivity in the model.

Given the importance of this technique for the interpretation of results in this field, in the present chapter I investigate its sensitivity and robustness to changes in the input values of those parameters and routines open to choice.

Geographic epidemiological models generally use maps containing information about environmental variables obtained from remote sensing or interpolated data collected *in situ* (e.g. from ground-based meteorological stations) as a data source. Reliable distribution maps of the species investigated are then used to extract from those sources the environmental parameters describing the localities where the organisms are present and absent (Hay et al. 2000, Rogers & Randolph 2003). Other biological, physical or socio-economic variables may also be included in the analyses. In discriminant function analyses the best predictors of the spatial distribution of a species are selected using a stepwise procedure that maximizes a certain function that best describes these training-set data. Such a function might be based on some measure of accuracy, selecting those variables that improve the ability of the model to predict the classification of the training-set data correctly into the proposed categories (i.e. presence and absence) at each step of the modelling process (henceforth I refer to methods maximizing this function as 'predictive accuracy' methods). Alternatively, this function might be based on the ability of the model to define a multidimensional space with those environmental variables that best segregate the datapoints belonging to the presence and absence categories (referred to as 'habitat discrimination' methods here). In this latter case, the appropriate metric for measuring these environmental distances at each step of the process is the so-called 'Mahalanobis distance' between the centroids (means of multivariate distributions) of the different categories, as it takes into account the correlations among the environmental variables selected and is independent of their scale of measurement. In a second step the process is reversed; the predictors selected by either method are used to determine the likelihood that a species occurs in geographic locations where it has not necessarily been observed before, by comparing its environmental similarity relative to the presence and absence data available and using maximum likelihood methods to ascribe the probability that the organism is present in that area, or can colonize it if introduced there (Rogers 2000; see also Chapter 4). Such procedures are, however, not completely pre-determined, giving the opportunity for input of biological knowledge at several stages of the analytical process

(Rogers 1998). In fact, it is upon this transparency and flexibility that the advantages of DFA mostly rest. But it is also here where its weaknesses may also lie, because if these decisions are based on uncertain biological evidence or result in different (but not necessarily 'wrong') interpretations about the optimal choices to be made, the outcome of the model might be affected in ways not easily measurable by the non-expert reader, such as public administrators and health authorities.

In addition to the requirement of a homogeneous matrix of covariance, in its classic form DFA also requires multivariate normality (Williams 1983, Williams & Titus 1988). Environmental variables, however, are seldom normally distributed and linearly related, preventing the mathematical description of the training set of species with parameters such as means, variances and covariances. One solution to this problem is the use of non-linear discriminant analysis, which allows the partition of the data into environmentally similar clusters of presence and absence, so that the above assumptions are met and the different environmental spaces inhabited by the organisms are better defined (Rogers 1997, Rogers 2000). The clustering process is in itself very flexible, hinging on a number of decisions throughout its implementation. For instance, several clustering methods can be employed, leading to very different results (Milligan & Cooper 1987, Legendre & Legendre 1998). However, this issue will not be dealt with here since the criterion of minimization of the within-cluster sum of squares between the objects and cluster centroids (K-means in partitioning methods where the number of clusters is defined in advance; or Ward linkage algorithm in hierarchical methods; Milligan & Cooper, 1987, Legendre & Legendre, 1998) has been the widespread choice in the literature.

Instead, I will analyse the effects of choosing a higher or lower number of clusters to describe the data once the clustering process is completed and the classification tree has been obtained. In general, this decision is based upon the visual inspection of bivariate scatterplots to determine how well the clusters account for non-linearity in the data set. The choice of the best number of clusters to use can also be assisted by determining the number which (1) best normalizes the distributions of the variables as shown in univariate histograms, (2) distinguishes the absence and presence records into the most cohesive groups in the classification tree, and (3) corresponds to

the point where the inclusion of additional clusters does not considerably contribute to a reduction in the variability in the multivariate description of each category (i.e. the sum of squares within groups).

Another crucial decision faced by the researcher refers to the number of climatic variables to be included in the model. The 'stopping rule' in these cases can be the stabilization of some measure of model accuracy which informs how well the model classifies the training-set data, such as the kappa statistic, Robinson, 2000).

In all these stages the results of the decisions open to choice are still unstable due to the unavoidable compromises behind each judgment. Partitioning the data into a higher number of clusters improves the multi-normality of the data, but might result in some clusters having an insufficient number of independent data points, therefore producing ill-defined covariance matrices. The inclusion of more variables increases not only the accuracy of the model, but it might also lead to 'over-fitting' of the model, namely the fact that with the inclusion of each new variable the environmental profile of the training-set data is more precisely defined, thereby reducing the likelihood that a new data point is considered 'similar' to it, thus reducing the generality of the results. Such decisions are therefore dependent on how a researcher trades-off the consequences associated with each choice, thus depending to some extent on subjective (presumably expertise based) criteria. If this were not the case, it would be possible to computerize the process by incorporating these analyses as stopping rules in the algorithm performing the computations. Even so, the possible mechanization of such procedure would still leave the researcher with the need to check the coherence of the results, to try to justify the chosen partition of the data and the causal mechanisms by which the selected variables act on the biological system under investigation.

My objective here is to examine how sensitive the modelling process is to variations in these decisions to be made by the researcher. Are the predictions generated by the model substantially affected by these choices? Do different decisions lead to similar conclusions about the environmental factors involved in the distribution of an organism? To answer these questions, I simulated a number of predictions about the geographic distribution of this group of viruses by systematically changing the

input values of the parameters investigated. The various predictions were subsequently compared to a template through pair-wise post classification comparisons. In addition, all previous comparisons were investigated for two 'templates', generated under two different frameworks underlying the variable selection process: 1) a method that maximizes the predictive accuracy of the model to classify the training-set data (the 'predictive accuracy' method), and 2) a method that maximizes the segregation of the different categories in the multivariate space (the 'habitat discrimination' method). As a common data source, I used the data employed to describe the distribution of Japanese encephalitis in Chapter 4.

Methods

Two reference models were built, each one based on a specific variable selection method, a certain number of clusters and a certain number of environmental variables. These models and resulting predictive maps were used as appropriate 'templates', to which all the predictions derived from models simulated with the same method were compared. To this end, each of the distribution maps generated by each 'simulation' was overlaid on a pixel-by-pixel basis to the template, resulting in a new matrix consisting of four categories. Two categories contained those pixels for which there was agreement between the template and the simulation (virus presence-risk and virus absence-risk predicted both in the template and in the simulation), and the other two contained those pixels for which a match between the template and the simulation was not observed (virus presence predicted in the template but not in the simulation and vice-versa).

'Predictive accuracy' method

Several measures of agreement can assist the selection of environmental variables by 'predictive accuracy' methods: sensitivity (proportion of observed presences correctly predicted), specificity (proportion of observed absences correctly predicted), and the kappa index (proportion of agreement once random agreement is removed; Cohen, 1960). These measures – chiefly kappa – have been used in epidemiological studies with non-linear discriminant analysis (Rogers et al. 2002). Recently, Cumming (1999) proposed the use of yet another measure of model accuracy in the spatial study of infectious diseases, referred to as the area under the curve (AUC)

of receiver operating characteristic (ROC) plots. ROC curves plot the proportion of presences predicted correctly (true positive cases) versus the proportion of absences predicted incorrectly (false positive cases) at a number of classification thresholds. By determining both the sensitivity and specificity of the model when a range of probability thresholds is considered, it generates a measure of agreement based on an optimum compromise between these two parameters (Brooker et al. 2002). It was used in Chapter 4 in its non-parametric version to generate predictive maps for the global distribution of Japanese encephalitis, and it is used in the present chapter as a representative of the class of 'predictive accuracy' methods.

In Chapter 4 I described and predicted the geographical distribution of the Japanese Encephalitic (JE) virus. In that Chapter, the decisions about the number of variables selected were based on an extensive analysis of the data, with the correspondent justification about such choices being found in that chapter. The model used in Chapter 4 was therefore used as the 'template' for the comparisons between the results yielded by the 'predictive accuracy' method. The values used in the template model as well as in the simulations for each of the parameters investigated are described below:

Number of clusters within the presence data set

Template = 3;
Simulated = from 1 to 5.

Number of clusters within the absence data set

Template = 3;
Simulated = from 1 to 5.

Number of selected variables

Template = 8;
Simulated = from 1 to 10.

'Habitat discrimination' method

'Habitat discrimination' methods can also employ different criteria to compute the distance between the several centroids of the clusters within the different classes proposed (of presence and absence, or of different abundance classes). Mahalanobis distances between all pair of presence and absence clusters can be summed to provide a final value, they can be averaged, or the minimum value observed can be retrieved

(Rogers 1998). The methods are not equivalent and the decision on which one to choose relies on the biological assumptions about the processes constraining the species or populations environmentally within their habitats. In their different forms they have been used in several studies in the past (Hendrickx et al. 2000, Williams et al. 1992, Robinson et al. 1997, Rogers et al. 1996, Rogers 1997, 2000, Randolph & Rogers 2000, Rogers & Randolph 2000, Baylis et al. 2002).

As a representative of the 'habitat discrimination' method here I chose to select the variable that at each step of the modelling process maximizes the distances between the nearest cluster centroids linking the presence and absence data. This choice was made on the understanding that the most important environmental factors determining whether a species will or will not inhabit a place are those that discriminate better between the most similar places where the species is, and is not, present.

Using this method I built a second template model using the same JE data set and the same partition of the data (3 presence and 3 absence clusters) as in the previous method. Nevertheless, as in the previous case, variables were selected until there was no further significant improvement in the accuracy of the training set classification. This decision on when to stop the selection of variables was assisted by the plot shown in Figure 5.1.

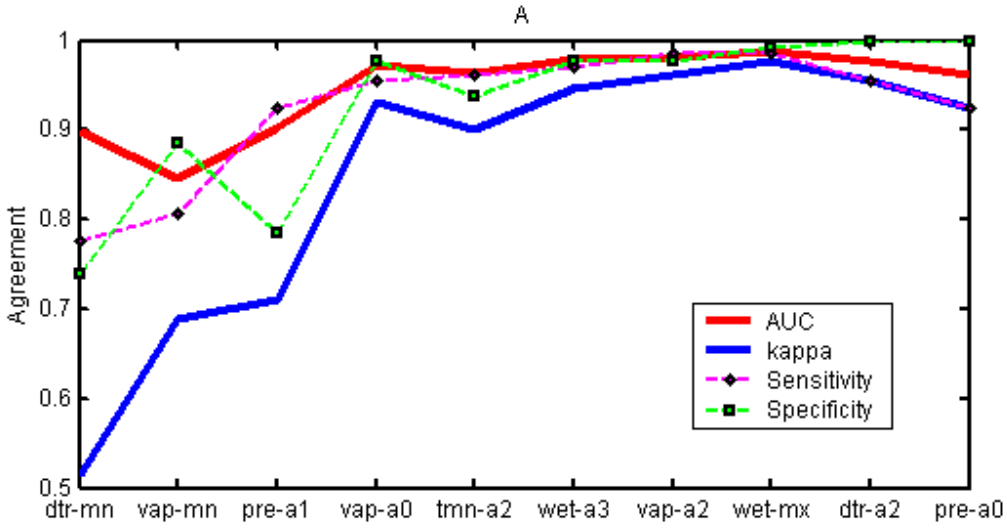


Figure 5.1 Curves used to assist the decision about the number of variables to be included in the model. Values reflect the effect of the model when the variable in the horizontal axis and all the others in its left are considered. The AUC (area under the curve) index is a threshold independent measure of model accuracy and it was the function that was maximized in the variable selection process. Sensitivity:

proportion of presence points of the training-set correctly predicted. Specificity: proportion of absence points of the training-set correctly predicted. Kappa: proportion of specific (non-random) agreement. All possible measures ranges from zero to one but kappa, which ranges from –1 to +1. A value equal to one in this case is – similarly to the other measures – an indication of perfect agreement, and values of zero or less suggest a performance neither better nor worse than random agreement.

The trend shown for most of the curves of Figure 5.1 is to reach a plateau of high agreement values with the inclusion of the fourth variable. Therefore, the template model representing the 'habitat discrimination' method was built with four variables. As a result, the values of the parameters used for the template and those that were systematically modified in the simulations were:

Number of clusters within the presence data set

Template = 3;
Simulated = from 1 to 5.

Number of clusters within the absence data set

Template = 3;
Simulated = from 1 to 5.

Number of selected variables

Template = 4;
Simulated = from 1 to 10.

Ranges of environmental resemblance used to produce the predictive maps

In Chapter 4, I mentioned the trade-off between the desired environmental breadth (and consequently geographic extension) for which the predictions of presence-risk or absence-risk of the disease will be made and the reliability of such predictions. In general, the larger the difference between the environmental profile of an unclassified area and that of the training-set data, the less reliable a prediction might be (given the low resemblance of the former with the environmental values defining both the presence and absence of a species). One way to deal with this is to set cut-off values, or inclusion criteria, which determine the maximum allowed distance between the multivariate centroids of the training set and the unclassified data point (Rogers 1998, 2000). Here I use three different inclusion criteria to determine the effects that changes in the parameters chosen by the researcher have on the model output under each of these three types of restriction. The implementation of these inclusion criteria

applied both to the template and to each simulated model simultaneously (i.e., no comparisons were made between models built with different inclusion criteria). The analysis was restricted to those geographic areas where both the template and the simulation maps produced predictions.

1) 'Safe' ranges

To set a cut-off value that maximized the likelihood of producing appropriate predictions I used the following procedure. First the Mahalanobis distances between the centroids of each cluster and its sample points were calculated, both for the presence and for the absence data set, and the largest within-cluster distance ($D_{sample\text{-}centroid}$) was registered for each cluster (excluding 5% of the extreme data). Secondly, the distances between the unclassified pixel and the centroids of the clusters of the training-set ($D_{new\text{-}centriod}$) were calculated to determine to which category (presence or absence), and each cluster within category, this new data point belonged. Finally, a prediction was produced if the latter distance ($D_{new\text{-}centroid}$) was shorter than the within-cluster distance ($D_{sample\text{-}centroid}$) of the cluster to which the unclassified point was closest.

2) 'Intermediate' ranges

The procedure followed here was a relaxed version of the above. The only difference is that a prediction was allowed if the distance between the unclassified pixel and the centroids of the nearest cluster of the training-set was shorter than twice the maximum within-cluster distance for the latter cluster (that is, if $D_{new\text{-}centroid} < 2 * D_{sample\text{-}centroid}$)

3) 'Wild' ranges

No limit was imposed. Forecasts of virus presence-risk/absence-risk were generated for all pixels within the land surface.

Data analysis

Since the objective is to compare distribution predictions generated with different parameters, it would be tempting to use agreement statistics such as the kappa index to describe the degree of consistency between the different results. The use of the kappa statistic in this stage of the spatial analysis, however, would be problematic as it requires independence of data. This assumption would be dramatically violated when

using continuous cells of grid-data given the likely spatial autocorrelation within the data. Furthermore, the scale employed here encompasses the entire globe, and therefore the traditional utility of significance testing, confidence intervals, etc., is lost (Zaslavsky 1995). I therefore opted to compare the overlaid categorical maps by estimating the proportions of areas for which correct predictions of both presence-risk and of absence-risk overlapped.

The resulting maps were converted into bar charts to allow a clearer quantitative analysis. Risk maps were also used, but only to facilitate the qualitative comparison between the template model and the simulations involving similar parameters (i.e. number of clusters and variables similar to those used in the template). Finally, since the clustering process is conducted prior to the variable selection process, I have also examined whether variations in the number of clusters used affected which environmental variables were selected as predictors of presence-risk/absence-risk in the simulations.

Units of Measurement

The layers containing the climatic information (derived from interpolated climate data) used to build the model are represented in a cylindrical equidistant (Plate Carre') projection. I have therefore re-projected them and used as unit of measurement the percentage of land surface relative to the total land area considered.

Software

All analyses, graphics and statistical tests were conducted using programs I developed in Matlab specifically for this study.

Results

Overall analysis

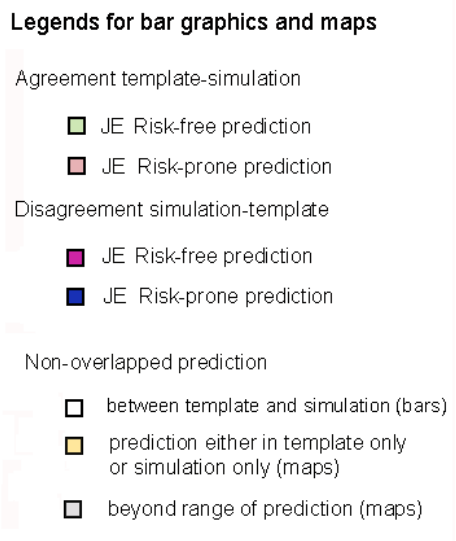
The relative positions, sizes and colours of the bars shown in the graphical outputs (Figures 5.2-5.8) show several different kinds of information. Firstly, **the total**

length of each bar represents the area (relative to the total size of the area analysed and measured in percentage) for which the inclusion criteria discussed above allowed predictions. For example, in those cases where predictions were limited by specified degrees of similarity between the training-set and the unclassified areas ('safe range' and 'intermediate range'), this length shortens progressively as the inclusion criteria become more strict and as more variables are added (Figure 5.2a,b; Figure 5.5a,b). This pattern arise from the fact that with the inclusion of each new variable the environmental profile of the training-set data is more precisely defined, thereby reducing the likelihood that a new data point is considered 'similar' to it. For obvious reasons, in those cases where making predictions was not limited by any criteria ('wild range') the size of the bars is never reduced, with predictive maps encompassing the entire geographic surface regardless of the number of variables included in the model (Figure 5.2 c; Figure 5.5 c).

Within any one set of inclusion criteria, no general trend could be identified as the numbers of clusters were varied (Figures 5.3, 5.4, 5.6, 5.7). Different numbers of clusters in fact led to the selection of different climatic variables (tables I - IV), and also produced changes in the values of the training-set centroids associated to each cluster. Since the new centroids may be more or less distant from the new data points to which they will be compared independently of whether more or less clusters are available, there is no consistent relationship between 'proportion of the land area for which predictions can be made' and 'cluster number'.

The second source of information readily available from a visual inspection of the figures relates to the relative proportion of predictions of absence-risk and presence-risk. This is indicated by the **total extension of the bar to each side of the vertical axis** positioned at the centre of the figure. The bars to the left side of the axis represent the absence-risk prediction area, while the bars on the right of the axis represent the presence-risk predictions. In some cases, the graphic is displaced to one of the sides. In such instances I did not rescale it on purpose to enable a quick visual presentation of how differently the predictions for both classes were affected by changes in the analytical parameters (numbers of clusters and variables used).

Finally, the most important information for the present analysis is **the proportion of disagreement (dark colours in the centre of the bar) relative to the sum of agreement and disagreement (that is, total coloured segment of the bar)**. The following colour code is used for all figures (Figure 5.2-5.8):



The areas in pale green and pink represent, respectively, the predicted risk-free and risk-positive areas that match the distribution of the template. Areas of disagreement between the template and the simulation are highlighted with darker colours: purple for areas that are risk-positive in the template and risk-free in the simulation (referred to operationally as 'false-negatives'), and dark blue for the opposite case ('false-positives'). For ease of comparison, these two latter regions were placed together in the centre of the bars.

All the above colour-coded information also encodes the same information when displayed in the maps positioned to the right of the bar charts. These maps show how the predictions produced in the simulations differ geographically from those produced by the template, allowing a qualitative assessment of the impact of the changes in the parameters under investigation. For example, it is possible to verify whether the disagreements in the predictions generated tend to correspond to those regions that are geographically more distant to the records of the training-set (southeast

Asia), or correspond to more densely populated areas, or even a particular habitat type, etc. The maps show the effects produced by the addition (upper map) or subtraction (lower map) of one variable (Figures 5.2, 5.5) or cluster (Figures 5.3, 5.4, 5.6, 5.7) relative to the template.

'Predictive accuracy' method

Comparison against simulations performed with different number of variables

For the 'safe range', changes in the number of variables included in the model do not have a noticeable impact upon the predictions produced (Figure 5.2a). In those cases where the number of variables is higher than in the template, the match is in fact almost perfect (i.e. there are no darker colours in the centre of the bars), with only 1.1% disagreement between the models when 8 (template) or 9 variables were selected. The percentage disagreement is even lower (0.6%) with one variable fewer than the template. It is only when the template is compared against simulations using five or fewer variables that the model presents some noticeable instability, with disagreements of 10-15%. . For the 'intermediate' range (Figure 5.2b), the percentage disagreement between the template and the simulation with one less variable is still within single figures (5.5%), but this increases in all the other cases to 10-20%. Similar, but lower, levels of disagreement were found for the wild range (e.g., 4.6% and 10.9% when the template is compared to simulations selecting seven and nine variables, respectively; Figure 5.2c).

It is worth noticing that in comparisons to simulations with a higher number of variables, both in the case of the 'intermediate' and 'wild' ranges, the regions of disagreements result completely from the attribution of risk-free areas to those regions where the template model predicted the presence of the disease (purple area in the bars). –This is especially worrying for epidemiological studies (Rogers 1998),especially as these 'false negative' areas include regions of high human density (such as the south of China or north of India, see upper maps in Figure 5.3b and 5.3c).

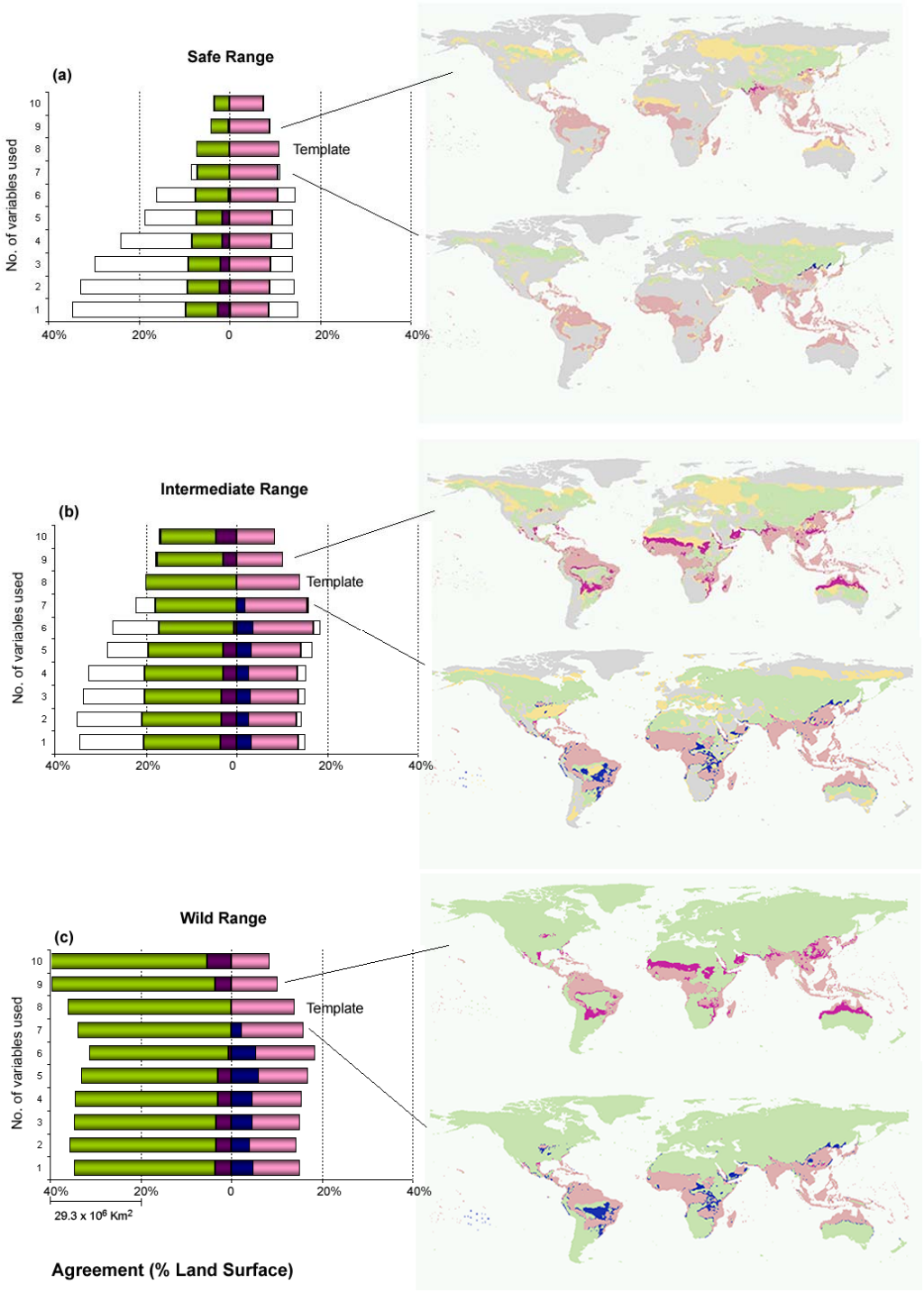


Figure 5.2. Level of agreement (as measured in land surface) for the predictions produced by the predictive accuracy model regarding the distribution of the Japanese encephalitis virus between a template model and simulated models for which **the number of climatic variables selected was systematically varied**. Range of environmental resemblance considered: a) 'safe', b) 'intermediate' and c) wild (see 'Methodology' for details).

Comparison against simulations performed with different number of clusters

The impact of differences in the number of clusters selected by the researcher on the predictions is again highly dependent upon the criteria of environmental similarity used to produce the risk maps.

When the number of presence clusters is systematically varied, the predictions produced by the simulations differ again only slightly in comparison to those associated to the template model in the 'safe range' (Figure 5.3a), with a maximum of only 3.3% disagreement (when the simulation was conducted with only one cluster of presence). Under the 'intermediate range', the level of disagreement between the template and its 'neighbouring' simulations (i.e., those considering one cluster of presence more and one less than the template) rises to 13.8 and 18.0%, respectively (Figure 5.3b).

In the wild range these figures drop slightly to 9-15% (Figure 5.3c). This reduction results from the fact that, although the areas of disagreement are basically the same for the 'intermediate' and 'wild' ranges, in the wild range the additional area for which predictions are produced corresponds mostly to predictions of low-risk, which do not change with variations in the number of clusters (since the area for which predictions are produced is larger, the proportion of disagreement is reduced; maps of Figures 5.2c, 5.3c and 5.4c).

The differences in overall level of disagreement between the template and simulations performed with different number of absence clusters (Figure 5.4) reach a maximum of 3.1% in the 'safe range', range from 13.8 to 18.0% in the 'intermediate range' and 10.7 to 11.4 % in the 'wild range'. Therefore they follow the same trend as for variations in the number of presence clusters (Figure 5.3). The only difference between varying the numbers of presence and absence clusters is as follows: when the number of presence clusters increase, the area corresponding to risk-positive predictions decreases (segment of the bars at the right side of the y-axis in Figures 5.3b and 5.3c); therefore, the number of 'false absences' tend to diminish with an increase in the number of clusters, whist the opposite happens in the case of 'false presences', which reach their maximum with the highest number of presence clusters. Such a phenomenon is not observed when the template is compared to simulations with a different number of absence clusters.

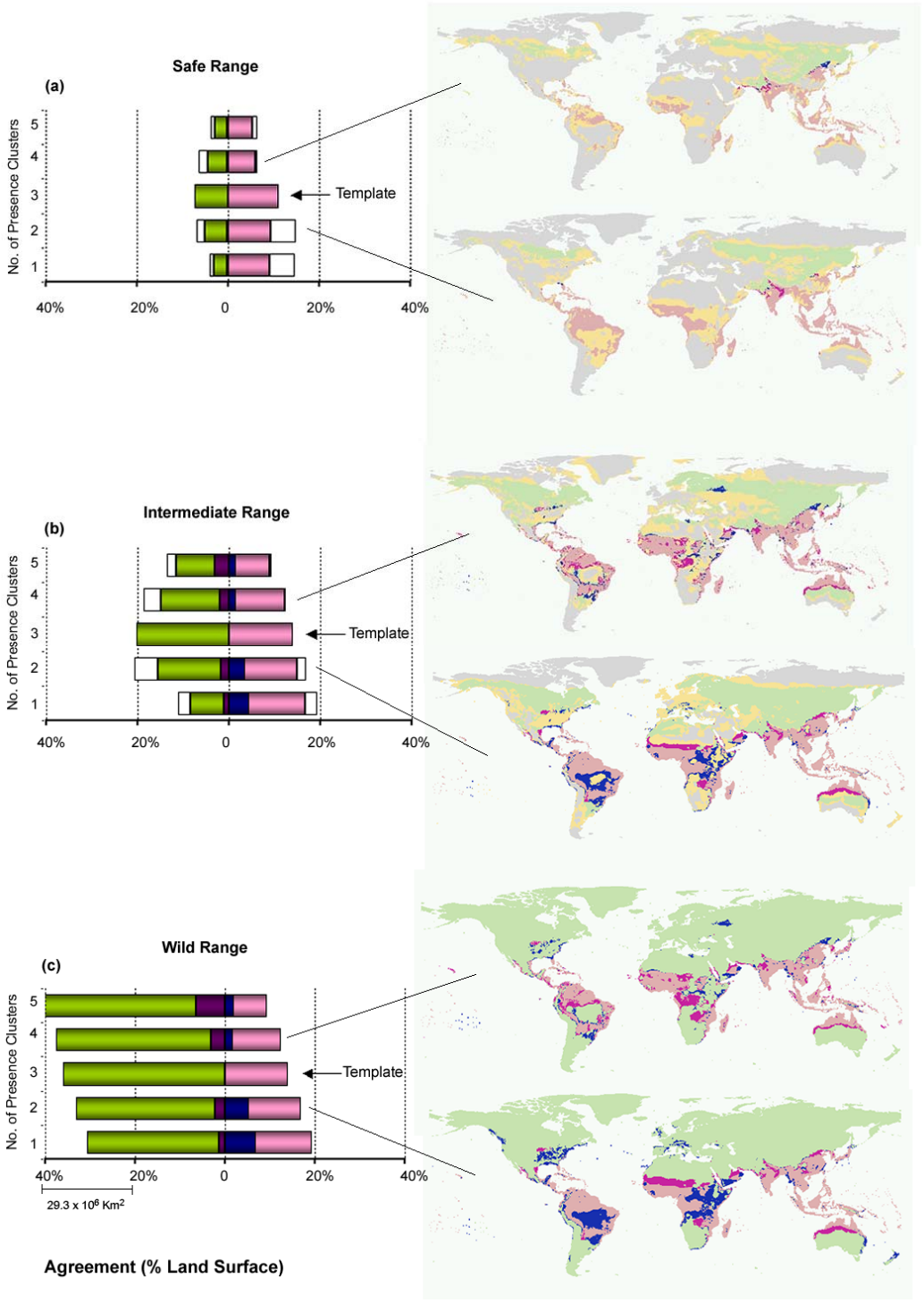


Figure 5.3. Level of agreement (as measured in land surface) for the predictions produced by the predictive accuracy model regarding the distribution of the Japanese encephalitis virus between a template model and simulated models for which **the number of clusters defining the presence data was systematically varied**. Range of environmental resemblance considered: a) 'safe', b) 'intermediate' and c) wild (see 'Methodology' for details).

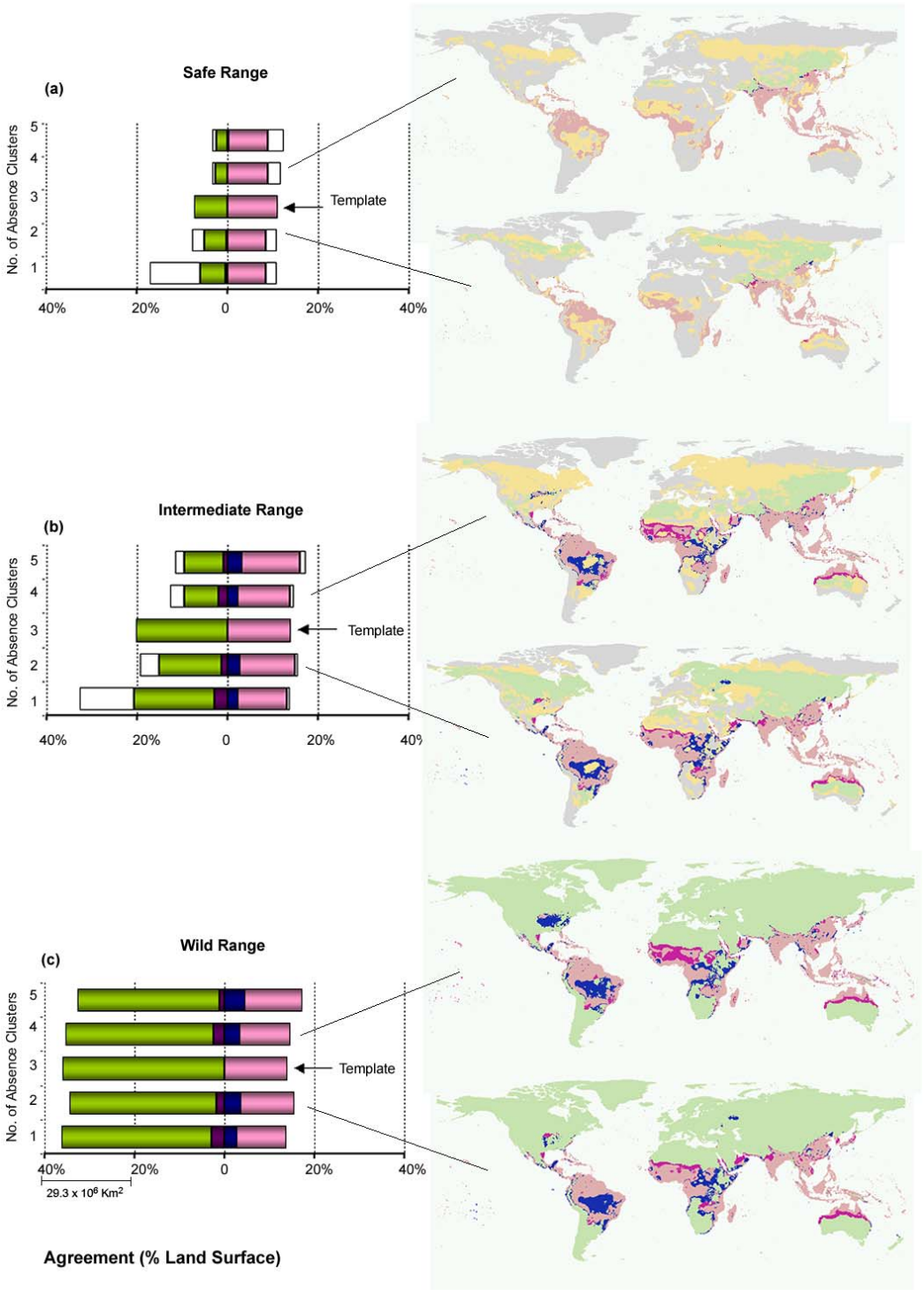


Figure 5.4 Level of agreement (as measured in land surface) for the predictions produced by the predictive accuracy model regarding the distribution of the Japanese encephalitis virus between a template model and simulated models for which **the number of clusters defining the absence data was systematically varied**. Range of environmental resemblance considered: a) 'safe', b) 'intermediate' and c) wild (see 'Methodology' for details).

Impact of differences in the number of clusters on the variables selected

As shown in tables 5.1 and 5.2, there was considerable variation in which climatic variables were selected when the simulated models employed different numbers of presence or absence clusters. In fact, if we consider the eight first variables used to build the template, each simulation selects, on average, 2.5 variables that are common to those selected by the template (this average was calculated by considering Tables 5.1 and 5.2 together).

Table 5.1 Climatic variables selected when the number of clusters defining the data set of virus **presence** was varied. The number of absence clusters was kept constant for all models, being equal to three. Colours of letters are used to assist in the identification of coincidence in variable selection.

		Template	Simulations under changed n° of **presence** clusters			
		3	1	2	4	5
Variables selected	1st	**vap-mx**	vap-mn	vap-a0	**vap-mx**	**vap-mx**
	2nd	**dtr-mx**	tmp-mx	tmx-mx	**dtr-mx**	vap-mn
	3rd	**frs-a3**	**tmn-mx**	dtr-a0	vap-a3	dtr-mn
	4th	**tmn-mx**	tmn-a2	tmp-mx	**tmn-mx**	pre-mn
	5th	**pre-mx**	tmx-a1	pre-a0	**wet-a1**	tmn-a0
	6th	**wet-a1**	tmx-mn	frs-a0	tmp-a3	tmn-a2
	7th	***tmx-a0***	tmp-a0	tmn-a0	dtr-a3	tmx-mn
	8th	frs-mx	dtr-a0	tmx-a1	***tmx-a0***	tmx-a1
	9th	tmn-a0	tmn-a1	dtr-a2	tmn-mn	dtr-a3
	10th	tmx-mn	tmn-a0	**pre-mx**	tmx-a2	frs-mn

Table 5.2 Climatic variables selected when the number of clusters defining the data set of virus **absence** was varied. The number of presence clusters was kept constant for all models, being equal to three. Colours of letters are used to assist in the identification of coincidence in variable selection.

		Template	Simulations under changed no. of **absence** clusters			
		3	1	2	4	5
Variables selected	1st	**vap-mx**	**vap-mx**	**vap-mx**	**vap-mx**	**vap-mx**
	2nd	**dtr-mx**	wet-mx	**dtr-mx**	frs-mn	vap-mn
	3rd	**frs-a3**	dtr-a0	**wet-a1**	tmn-a1	**wet-a1**
	4th	**tmn-mx**	**pre-mx**	tmp-a3	vap-mn	tmp-a0
	5th	**pre-mx**	***tmx-a0***	**pre-mx**	**pre-mx**	tmn-a3
	6th	**wet-a1**	tmp-a1	tmp-mx	**wet-a1**	frs-mn
	7th	***tmx-a0***	dtr-a3	tmp-a1	tmn-a0	**pre-mx**
	8th	frs-mx	frs-a2	dtr-a3	pre-mn	tmx-mn
	9th	tmn-a0	frs-a3	dtr-a0	tmn-a3	vap-a2
	10th	tmx-mn	pre-a2	tmx-a3	vap-a1	tmp-a3

The first variable selected in the template model, maximum vapour pressure (vap-mx), was in fact selected as the first variable by all clusterings of the absence data (Table 5.2) and when the presence data were partitioned into three or more clusters (Table 5.1). Maximum precipitation (pre-mx) is the only other variable that was consistently selected, first in all simulations with a different number of absence clusters

(Table 5.2) and in one simulation with a different partitioning of the presence data (Table 5.1).

'Habitat discrimination' method

Comparison against simulations performed with different number of variables

Under the 'habitat discrimination method' the patterns of disagreement between the template and the simulations with different numbers of variables differed from that shown for the 'predictive accuracy' method. In the safe range, there were again almost no differences between the template and most of the simulations (0.3-2.0% disagreement with one more and one less variable, respectively; Figure 5.5a). Highest levels of disagreement were, however, obtained in the wild range, with 9.4-13.0% disagreement with one fewer and one more variable, respectively, increasing even further when the number of variables used to generate the simulations depart further from the template model (Figure 5.5c).

It is worth noting that, for all similarity ranges, the percentage of 'false positives' shoots up when only one variable is selected. This effect results from the type of 'habitat discrimination' method chosen, which selects those variables that best segregate the most environmentally similar clusters at each round of the modelling process. Each variable can define very accurately the most similar clusters in most instances, but will not necessarily discriminate well the totality of the presence and absence training set. Models produced by this method and involving only one variable are therefore likely to have poor predictive power.

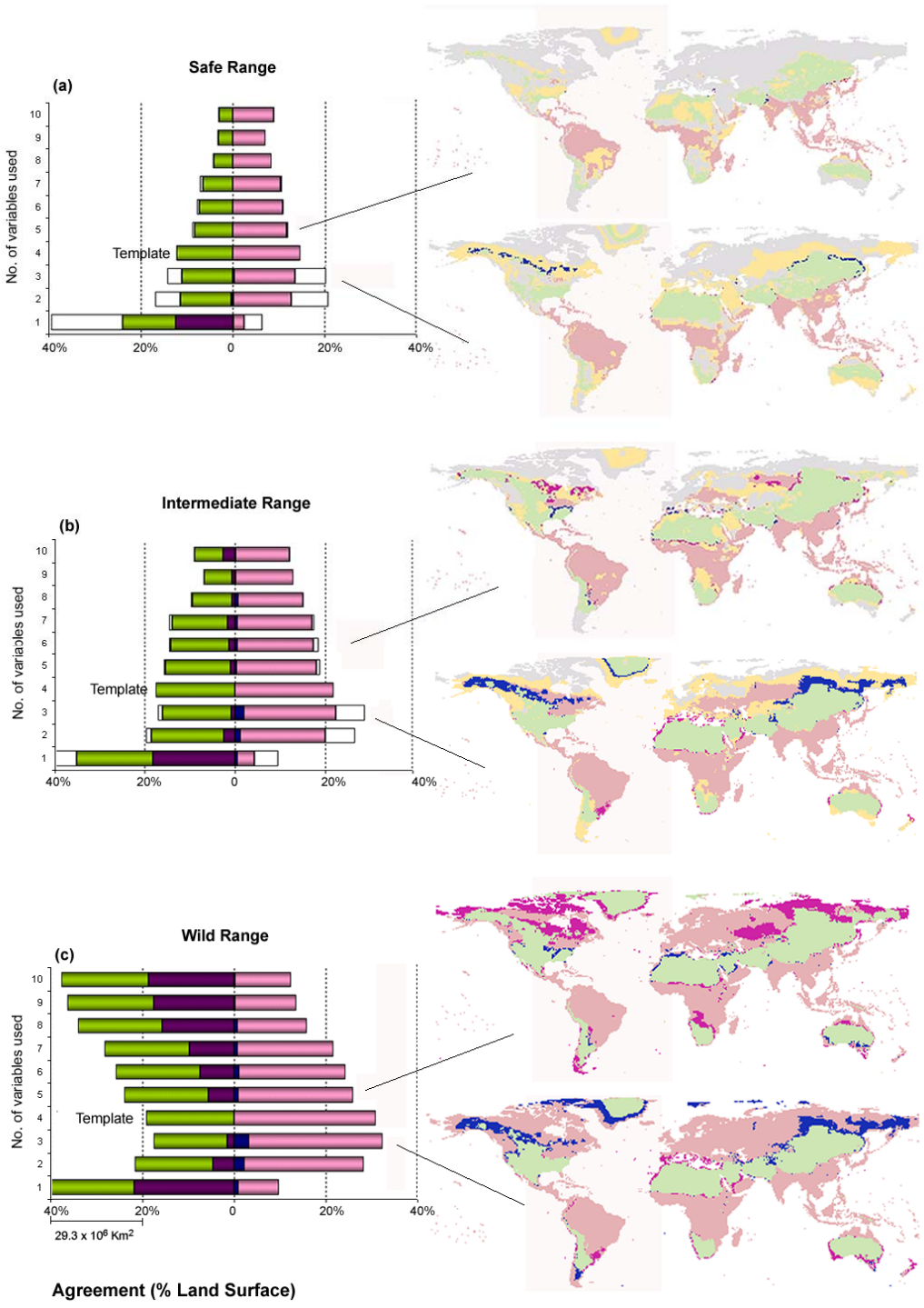


Figure 5.5. Level of agreement (as measured in land surface) for the predictions produced by the habitat discrimination model regarding the distribution of the Japanese encephalitis virus between a template model and simulated models for which **the number of climatic variables selected was systematically varied. Range of environmental resemblance considered**: a) 'safe', b) 'intermediate' and c) wild (see 'Methodology' for details).

Comparison against simulations performed with different number of clusters

The impact of differences in the number of clusters selected with this method is similar to that observed with the previous one. Under the 'safe range', there was 2.1-4.0% disagreement with one less and one more presence cluster, respectively, than in the template (Figure 5.6a). Equivalent figures rose to 9.7-15.0% for the 'intermediate' range (Figure 5.6b) and 21.0-23.4% for the 'wild range' (Figure 5.6c).

When the number of absence, rather than presence clusters is varied, lower levels of agreement are observed. For the safe range, there was 6.3-9.5% disagreement with one less and one more presence cluster, respectively, than in the template (Figure 5.7a). Equivalent figures rose to 16.4-20.8% for the 'intermediate' range (Figure 5.7b), and 29.6-32.4% for the 'wild range' (Figure 5.7c).

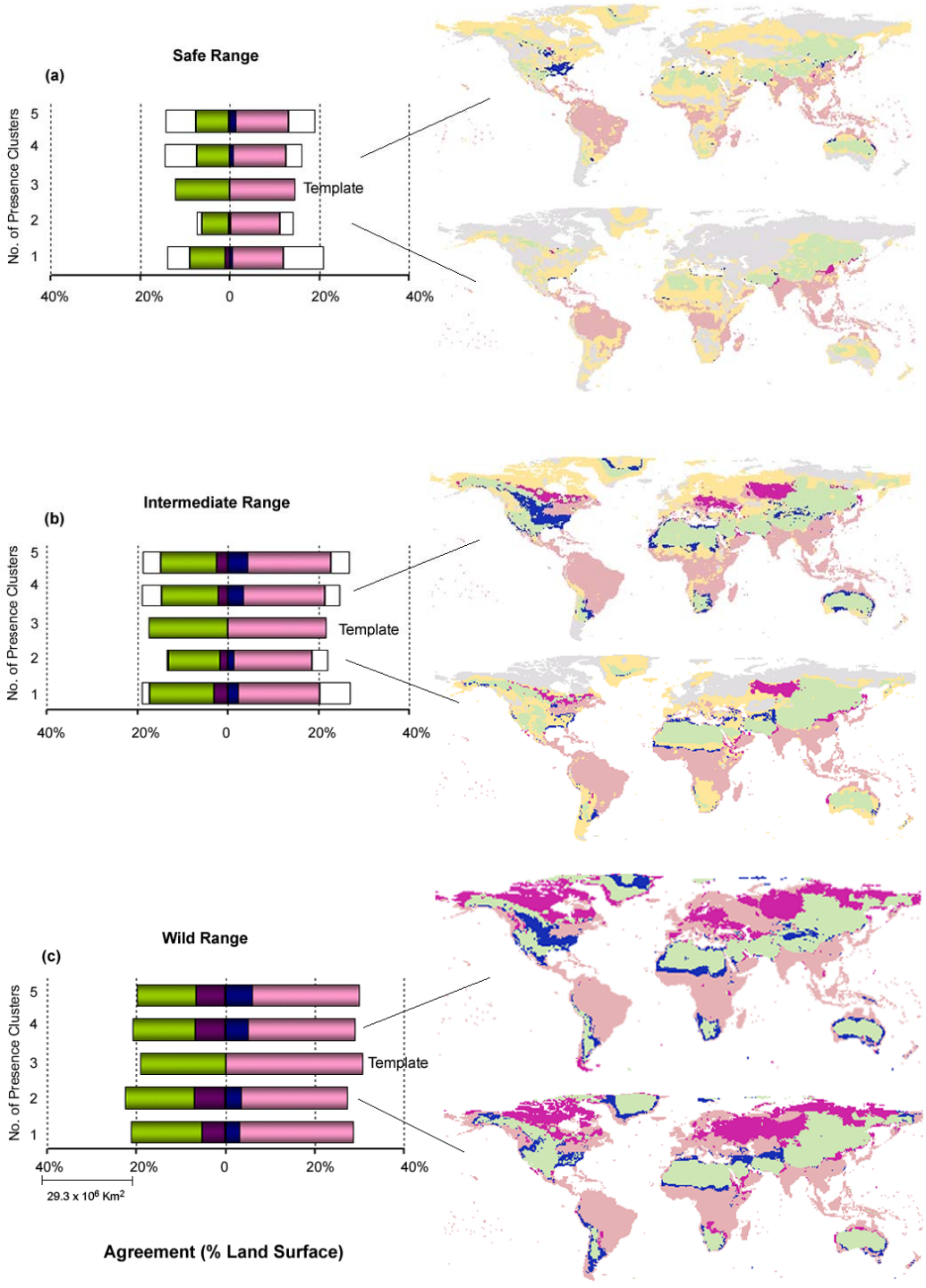


Figure 5.6. Level of agreement (as measured in land surface) for the predictions produced by the habitat discrimination model regarding the distribution of the Japanese encephalitis virus between a template model and simulated models for which **the number of clusters defining the presence data was systematically varied**. Range of environmental resemblance considered: a) 'safe', b) 'intermediate' and c) 'wild' (see 'Methodology' for details).

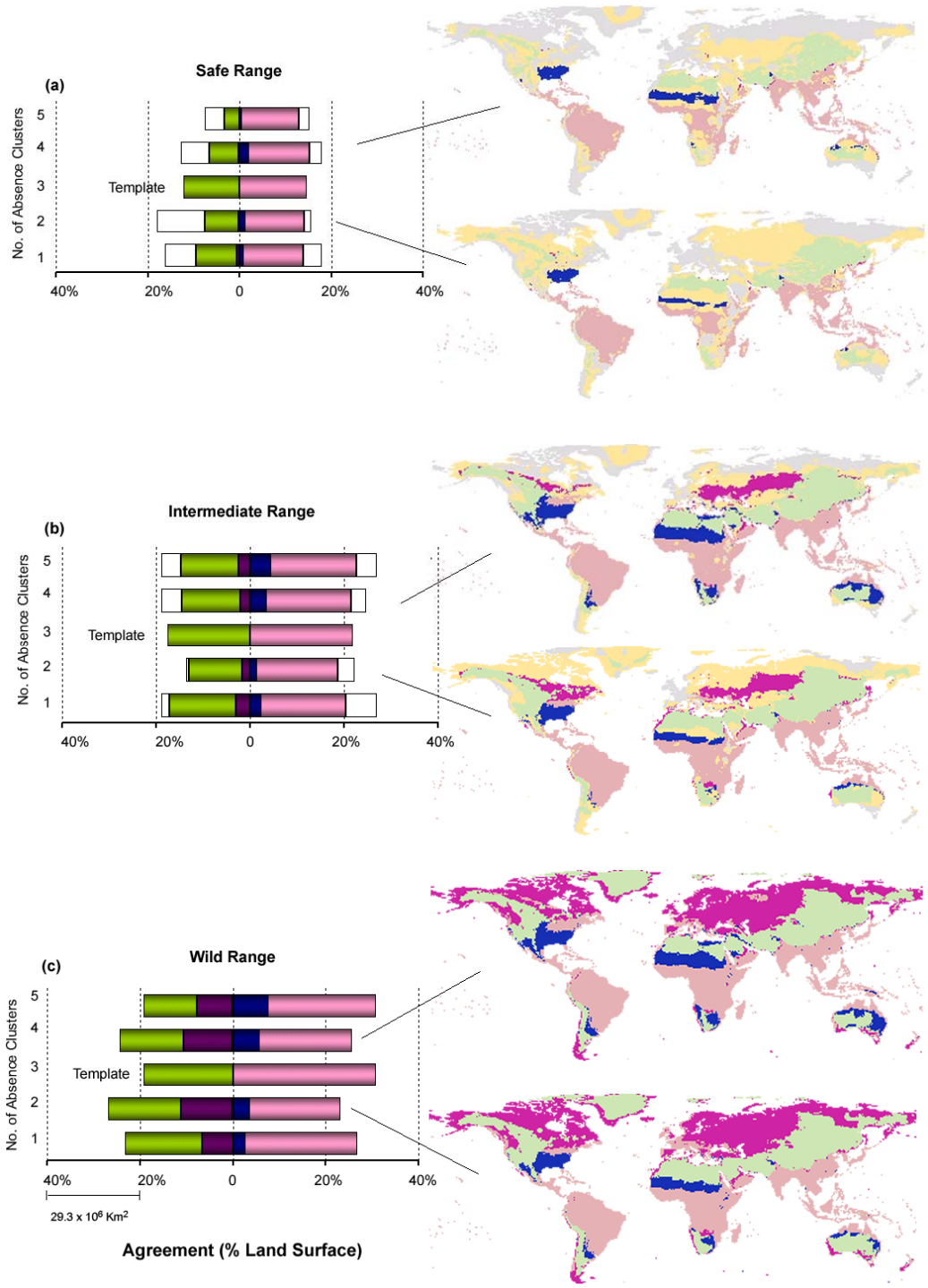


Figure 5.7 Level of agreement (as measured in land surface) for the predictions produced by the habitat discrimination model regarding the distribution of the Japanese encephalitis virus between a template model and simulated models for which **the number of clusters defining the absence data was systematically varied**. Range of environmental resemblance considered: a) 'safe', b) 'intermediate' and c) 'wild' (see 'Methodology' for details).

Impact of differences in the number of clusters on the variables selected

Similar to the results observed for the 'predictive accuracy' model, the climatic variables selected by models with different clustering of the data did not match those selected by the template model very closely. On average, each simulation selected less than one variable of the first four used to build the 'template' model (considering Tables 5.3 and 5.4 together).

When the simulations varied the number of clusters defining JE presence (Table 5.3), the variable "minimum values of diurnal temperature range" (dtr-mn) was in fact the first one to be selected whether the data set consisted of one, two or three clusters. The only other variable to be selected in more than one simulation, however, was the mean values of vapour pressure (vap-a0, Table 5.4).

Table 5.3. Climatic variables selected when the number of clusters defining the data set of virus **presence** was varied. The number of absence clusters was kept constant for all models, being equal to three. The first four variables selected in the template model are highlighted in the table.

		Template	Simulations under changed no. of **presence** clusters			
		3	1	2	4	5
Variables selected	1st	***dtr-mn***	***dtr-mn***	***dtr-mn***	pre-mx	pre-a0
	2nd	***vap-mn***	pre-a0	pre-a0	***vap-mn***	tmp-a1
	3rd	***pre-a1***	dtr-a3	tmn-a1	dtr-a3	pre-a2
	4th	***vap-a0***	wet-a1	tmx-a1	tmn-a1	tmn-a3
	5th	tmn-a2	dtr-a2	***pre-a1***	pre-mn	dtr-a3
	6th	wet-a3	vap-a2	dtr-a2	vap-a2	pre-mn
	7th	vap-a2	dtr-a1	wet-mx	wet-a3	dtr-a2
	8th	wet-mx	frs-a0	vap-a2	wet-mx	vap-a2
	9th	dtr-a2	tmn-a1	wet-a3	dtr-a0	wet-a3
	10th	pre-a0	vap-a1	dtr-a1	dtr-mx	tmx-a2

Table 5.4. Climatic variables selected when the number of clusters defining the data set of virus **absence** was varied. The number of presence clusters was kept constant for all models, being equal to three. The first four variables selected in the template model are highlighted in the table.

		Template	Simulations under changed no. of **absence** clusters			
		3	1	2	4	5
Variables selected	1st	***dtr-mn***	pre-mx	pre-a0	tmp-mn	***vap-a0***
	2nd	***vap-mn***	frs-a0	tmn-mn	pre-a0	wet-a1
	3rd	***pre-a1***	wet-mn	wet-a2	wet-a2	tmn-mn
	4th	***vap-a0***	***vap-a0***	vap-mx	tmn-a1	pre-mx
	5th	tmn-a2	tmn-a1	tmn-a2	***pre-a1***	***vap-mn***
	6th	wet-a3	tmp-a2	tmx-mn	tmx-mn	tmx-a3
	7th	vap-a2	tmx-a3	dtr-a1	dtr-a1	vap-a2
	8th	wet-mx	tmn-a3	vap-a2	vap-a2	wet-a3
	9th	dtr-a2	tmx-a2	wet-a3	dtr-a2	tmx-mx
	10th	pre-a0	tmp-mn	frs-a1	wet-a3	pre-a0

‘Predictive accuracy’ versus ‘habitat discrimination’ methods

In Figure 5.8 the predictions of the templates of the ‘predictive accuracy’ and ‘habitat discrimination’ methods are compared. In this safe range, only 5.3% disagreement is observed. This figure is dramatically increased (27.3%) in the ‘intermediate range’ and even further in the ‘wild range’ (40.0%). Beyond the safe range, these mismatches resulted from the fact that the ‘predictive accuracy’ method produced a much larger area of risk-negative cases than the ‘habitat discrimination’ method.

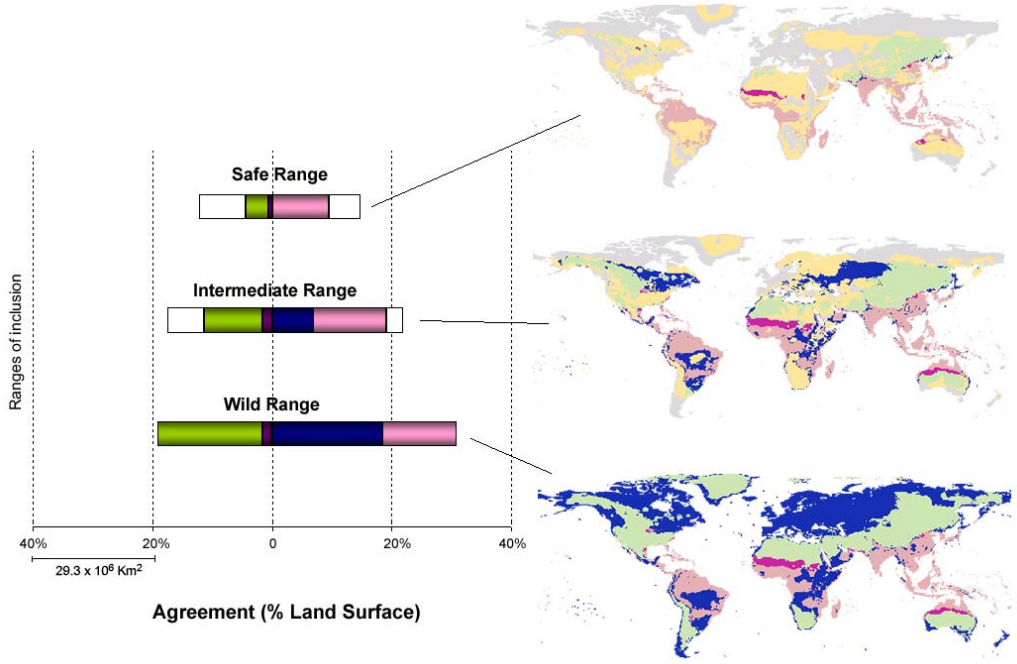


Figure 5.8. Level of agreement (as measured in land surface) for the predictions regarding the distribution of the Japanese encephalitis virus **between the template model of the ‘Predictive accuracy’ method and the template of the ‘Habitat discrimination’ method**. Range of environmental resemblance considered: a) 'safe', b) 'intermediate' and c) ‘wild’ (see 'Methodology' for details).

Discussion

The statistical modelling of vector-borne diseases can be an extremely useful tool for generating forecasts of great epidemiological importance, especially in those cases where biological approaches are limited by the lack of appropriate quantitative descriptions of crucial parameters and their relations with environmental factors

(Rogers & Packer 1993, Hay et al. 2000, Rogers & Randolph 2000). Yet predicting the distribution of organisms requires a deep knowledge about relevant environmental factors, the ecology of the species, statistical modelling, geographical information systems and remote sensing for it to be cost-effective (Hay et al. 2000, Austin 2002). Here I explored some of these latter aspects, by examining how the predictions produced with non-linear discriminant techniques can be affected by variations in the decisions made during the modelling process derived, for example, from different assumptions about the ecology (or biology) of the species investigated.

The present results have shown that the decisions made during the implementation of the discriminant function analysis can have a significant impact on the conclusions about the environmental factors underlying an organism's distribution, as well as on the final predictions regarding its distribution in unclassified areas. Nevertheless, I have also shown that the impact of possible differences on the results obtained will depend not only on the parameter values chosen by the researcher, but also – and most significantly – on the limits imposed on the differences between the environmental profile of the new locations and that of the training-set data for the production of predictions. In those cases where no restrictions are imposed – as in the situation referred to as 'wild range' here – the mismatch among the predictions produced with different parameter values is high in almost all of the situations analysed, but especially under the 'habitat discrimination' method of variable selection.

The finding that the predictive power of the model is dependent upon subjective decisions made by the researcher whenever the climatic similarity between the training set and the new geographic area is low ('wild range') adds to the general argument against the production of predictions in such conditions. As highlighted by Rogers (1998, 2000), the absolute value of the Mahalanobis distance between a new data point and the training set (i.e. their environmental distance) must be considered when defining the probabilities that this point belongs to a particular category (in this case virus presence or absence). Otherwise, one would be forecasting the risk of a disease in geographic areas with climatic conditions very different from those used to discriminate between the presence and absence of the disease in the training-set data. To make it clearer, it would be similar, for instance, to predicting the risk of a disease

in the Amazon rainforest based on data collected in the United Kingdom and Spain. Here, in addition to the above argument that those analyses that violate this requirement of 'environmental similarity' might produce inappropriate predictions, I have also shown that the predictions produced under these conditions are themselves unstable with regard to the parameters used in the analysis.

At the other extreme, when predictions are limited to a 'safe range', the level of agreement between the different outputs is almost complete, independent of the number of climatic variables selected (provided that, in the specific type of 'habitat discrimination' method used here, at least two variables are selected). The predictions of the model are also, to some extent, still robust to changes in the number of clusters. Nevertheless, care should be taken when determining this number, because, although the levels of mismatch were generally <5%, for the 'habitat discrimination' method up to 9.5% disagreement occurred with a difference of just one absence cluster. Still, the reliability of the predictions produced under the safe range and their robustness to variations in the number of variables, clusters and method of variable selection are reassuring for the predictions of the distribution of Japanese encephalitis presented in Chapter 4, which were produced for this range of environmental similarity.

One obvious problem, however, associated with the use of the 'safe range' is the level of restriction imposed upon the geographical extent of possible predictions of the distribution of an organism, especially in those cases for which the training-set data does not include a large variety of environmental conditions. The 'intermediate' range seemed to provide a good compromise in the case of the 'habitat discrimination' method. The addition or subtraction of climatic variables to build the model in such a case led to mismatches of only approximately 5% (Figure 5.5b) – indeed a reasonable margin to allow for oscillations in the model output. But once more, the levels of disagreement with variations in the number of clusters used ranged from 10% to 20%, therefore affecting a relatively high proportion of the land surface and human population considered. The lowest observed percentage of mismatch between the template and the simulation – 9.7% (Figure 5.6b) – represents, in geographic units, an area of approximately 9 million Km^2 (roughly the size of the USA), or an area (retrieved from the map Figure 5.6b) inhabited by 536 millions of people according to

estimates from 1995. It is also worth noticing that such figures represent only those areas for which both the template and the simulation maps produced predictions, therefore not including the regions signalled in yellow - and grey - in the upper map of Figure 5.6b. An even stronger argument against the use of the 'intermediate' range comes from the comparison between the two different variable selection methods. While the use of the 'safe range' once more justified its name (levels of disagreements were kept low, at around 5%), the proportion of disagreement reached values of more than 25% in the intermediate range, a figure beyond any reasonable tolerance limit for most applications. It is thus possible that these sources of instability could lead different researchers to obtain different and sometimes opposing results with the same set of data, method, but different (and probably all mathematically valid) decisions. The likelihood that different results are generated is also increased if one considers that the climatic variables selected also depend on the number of cluster chosen. Therefore, unless this latter choice is assisted by a strong biological/ecological justification and/or the model is independently tested, the partitioning of the data should be conducted with caution. However, if such constraints are overcome, the clustering process can be useful not only to achieve the multi-normality and linearity necessary for the analyses, but also serve a very useful role on the understanding of the phenotypic and genotypic plasticity of a species to adapt to local conditions.

By considering the figures and results previously discussed, the conclusion and recommendation that emerges from this study is that predictions beyond the limits imposed by the 'safe range' should preferably be excluded, or made available only if a strong warning about the accuracy and reliability of the resulting forecast is provided. For example, different set of colours could be used for these predictions, as illustrated (in purple and beige) in the map showing the distribution of the Japanese encephalitis virus in Figure 4.14 (Chapter 4). Once these steps are taken, the technique can provide us with an extremely useful tool to explore the potential mechanisms underlying the current (and future) distribution of an organism.

Chapter 6

CLIMATIC CORRELATES OF PHYLOGENY:

A STUDY WITH VIRUSES OF THE JAPANESE ENCEPHALITIS SEROCOMPLEX

Introduction

In Hutchinson's definition of a "*species' fundamental niche*" (Hutchinson 1957, Hutchinson 1965) each species occupies an hypervolume within a space defined by coordinates along axes representing environmental gradients. Under this logic, the speciation process would imply changes in the 'coordinates' of one or more of those axes, so a new species would occupy a different subset of this hypervolume. But to what extent could climatic conditions act as a selective pressure in the cladogenesis of organisms like arboviruses (*ar*thropod-*bo*rne *viruses*), which could be thought of as being shielded from their external environmental surroundings due to their condition as obligate parasites?

It has already been shown that the distribution, abundance and population dynamics of the arthropod vectors of vector-borne pathogen systems can indeed be predicted and understood through the description and analysis of their spatial and temporal eco-climate spaces (Martens et al. 1999, Hay et al. 2000, Hales et al. 2002, Kuhn et al. 2002, Zhou et al. 2004). In addition to determining the distribution of vectors, abiotic factors may also directly influence the dynamics of interaction between hosts, vector and pathogens. For instance, in the case of arthropod-borne viruses, a group that relies on complex transmission cycles involving ectothermic vectors and vertebrate hosts, the presence of appropriate environmental conditions is essential to guarantee their successful transmission and completion of key stages of their transmission cycles (Reiter 1988, Martens et al. 1999, Hay et al. 2000, Randolph 2000, Hales et al. 2002, Randolph & Rogers 2002). This leads to the logical inference that the fundamental ecological niche of these pathogens would tend to be narrower than that defined by the range of environmental conditions where both vectors and hosts coexist. Currently, the most widely accepted concept of species used for viruses is defined as "*a polythetic class of viruses constituting a replicating lineage and occupying a particular ecological niche*" (Van Regenmortel 1990). This definition stresses the cohesive forces present in ancestral-descendant biological populations sharing a common biotic niche (Van Regenmortel & Mahy 2004), hence opening the possibility to investigate statistically the eco-climate context for the evolution of vector-borne pathogens.

Degallier *et al*. (1998) pioneered the use of multifactorial data to describe the ecological niche of virus species and compare it to known phylogenies. Nevertheless, the use of geographical systems of information (GIS) and spatial statistics for this aim has only recently been proposed (Randolph 2000, Randolph & Rogers 2002). The possibilities offered by this approach are indeed promising, as it potentially provides a basis upon which to explore the environmental forces shaping – or not – the evolution of pathogen species. However, because the approach is pioneering, basic data collection is still required, as well as the development of further statistical tools that are appropriate to tackle these questions. Here, I deal with both of these aspects, first by focusing the present investigation on a group of viruses of great epidemiological importance, and secondly by assembling a set of techniques both to build and test a model representing the environmental conditions involved in the evolutionary history of the virus species considered.

The genus *Flavivirus* is a geographically dispersed group consisting of approximately 70 antigenically related species, many of them with an enormous significance to human health, being responsible for many serious and widespread diseases such as dengue haemorrhagic fever, yellow fever (from which the name of the family and genus is derived: *flavus* is the Latin for yellow), Japanese encephalitis and tick-borne encephalitis. It constitutes a group of enveloped, positive sense, single stranded RNA viruses that radiated globally from a single ancestral virus existent 10,000 to 20,000 years ago, evolving rapidly to fill new ecological niches (Gould et al. 1997). In fact, the evolution, dispersal patterns and epidemiological characteristics of this group seem to be shaped by a set of biological, environmental and historic constraints, chiefly those given by the arthropod vectors, vertebrate hosts, their associated ecology and the influence of human commercial activities (Zanotto et al. 1996, Gould et al. 1997, Gaunt et al. 2001, Gould et al. 2001). As described by Gould *et al.* (2001): "*the flaviviruses are an excellent example of the influence of arthropods and their vertebrate hosts on virus evolution. They display a very wide range of epidemiological characteristics and the phylogenies reflect this as an arrangement of viruses that correlates very closely with their epidemiology and ecology*". If their evolution is so highly associated to their ecology, finding those eco-climate factors potentially related to the cladogenesis of this group could be extremely valuable from

an evolutionary, ecological and epidemiological perspective.

It is important to notice, however, that causal explanations are difficult to infer from correlative studies on spatial patterns, usually demanding a great deal of biological justification and, ideally, appropriate confirmatory experiments. For instance, even when species are adapted to local environments, we cannot say that the observed genetic differences were involved in the processes that led to the organisms' speciation, since these differences could have been accumulated due to genetic drift or as a result of adaptations to selective pressures other than those presently found (Gould & Vrba 1982). Caution is also needed when using this approach with special reference to one particular limitation: the spatial distribution of an organism does not necessarily reflect the whole range of environmental conditions to which it is adapted (the 'ideal niche'). The distribution does provide us with information about some 'slices' of this ideal hypervolume (or its 'realized niche' as defined by Hutchinson), but there are many reasons to expect that these 'slices' are not random representations of that entire space. Some of the reasons are methodological –biased 'sampling' is very common in studies that depend on available records. Other factors relate to the history and idiosyncrasies of the organism. For instance, organisms might never have had the opportunity to be exposed to the environment where they could actually thrive. In the case of viruses, cross-immunity prevents some species from occupying the same locations where other species are present, even when the ideal climatic conditions occur (Work 1971, Karabatsos 1985). On the other hand, geographic records of virus presence provide us with an immense source of data that, if well employed, can be extremely useful in understanding the processes underlying the genesis and distribution of the species studied.

Although the life cycle and biology of the vectors associated with the virus species belonging to the *Flavivirus* genus are highly variable, such variability is highly structured phylogenetically (Gaunt et al. 2001). Two major branches originated from the putative ancestor of the genus, one comprising non-vector-borne, and the other vector-borne virus species (Kuno et al. 1998). Within the vector-borne clade, the subsequent division leads once more to a major division of the virus cycle, separating those species relying on either ticks or mosquitoes as vectors. Those belonging to the

mosquito-borne clade are then grouped, to a large extent, depending on which of two major groups of mosquitoes (*Culex* spp. and *Aedes* spp. mosquitoes) act as potential vectors (Gould et al. 1997, Gaunt et al. 2001, Gould et al. 2001, Gould 2002, Holmes & Twiddy 2003). Here, I studied a sub-group of the *Flavivirus genus,* the Japanese Encephalitis serocomplex (JEV-serocomplex; Heinz et al. 2000, Gould et al. 2001) – that consists of closely related species all utilising mosquitoes as a vector. Besides my particular interest in this group of vectors as a consequence of the approach adopted in this thesis, studying species with a common kind of vector is important since other vectors, such as ticks, have very different biological, evolutionary and epidemiological dynamics (Zanotto et al. 1995, Randolph 1998, Gould et al. 2001), something which inserts a great deal of background variability into the data.

The phylogeny upon which I based this investigation changed during its development. In previous studies conducted by Gould *et al* (1997, 2001) and Gaunt *et al*. (2001), the JEV-serocomplex (also called *Culex* or JE/WNV serocomplex) was characterized by the presence of some species able to cause neurotropic diseases mainly in birds, but also possibly in rodent hosts, with vectors represented basically by *Culex* spp. mosquitoes. However, a review of the phylogeny conducted by Holmes and Twiddy (2003) incorporated species that also cause hemorrhagic diseases in primate hosts (such as dengue fever – although there are rare cases of dengue encephalitis; references in Gaunt *et al.* 2001), transmitted mainly by *Aedes* spp. mosquitoes. The present study was therefore expanded to consider the latter classification.

Non-linear discriminant analysis has been advocated as the most appropriate technique of multivariate analysis to describe the eco-climate of vector-borne diseases because it allows a biological interpretation of the transparent results, and the assumptions of common covariances between the multivariate spaces describing the different classes of the dependent variable are relaxed (Rogers et al. 1996, Rogers 2000). Therefore the technique was specially appropriated for addressing the present question, as each species certainly needs (at least) one particular multivariate description of its eco-climate space, and because the biological transparency offered by the technique is paramount for the exploration of possible causal mechanisms. In fact, Randolph and Rogers (2002) describe how this technique can be used to characterize

the eco-space of the western species of tick-borne flaviviruses based on records of virus presence and inferred virus absence. From these data, a matrix of ecological distances (as measured by the Mahalanobis metrics) between the species can be established, and compared with a matrix of genetic distances. Since the same environmental variables ought to be used to describe the eco-climate spaces corresponding to each species – otherwise the comparison between species is not possible – one method is to opt for the variables most frequently selected as predictors of the presence and absence of each virus, and used them to build the former matrix. Nevertheless, when looking for the overall environmental forces underlying the observed phylogenetic classification of a group, a sensible alternative procedure is to select directly those variables that best explain the relative genetic distances among a group of species, and subsequently measure the predictive fit of the model with independent data, instead of selecting those variables that explain the distribution of viruses relative to records of absence. Therefore, here I propose this latter technique, whereby a single stepwise process of variable selection is conducted for all species at once, aimed at maximizing the correlation between genetic distances and environmental distances described by non-linear discriminant analysis. To my knowledge such a method has not been proposed or used before.

To implement this approach, out of the set of species analysed, half (or another sensible figure) are used to build the model, and the other half are used to test it. This technique presents several immediate advantages over the previous one. For instance, it offers the inherent advantage of not involving the use of records of virus absence – an important advance for studies that depend on available records and where an overall sample design is not feasible (Hirzel et al. 2002). Instead, because the model is tested with species that were not used to build it, this method allows the use of all the records of presence for each species in the definition of its 'niche'. This procedure makes the model flexible, allowing its use in situations where some species are represented by only a few records. It is important to notice, however, that the results obtained in the test stage cannot be used to modify the model previously built, otherwise it would no longer constitute a test but only part of the modelling process, and the model itself would be left without a means of independent validation.

Finally, I also incorporate a control for the presence of spatial autocorrelation in the data by adding to the analyses matrices of geographic distances through the use of partial correlative methods. This is necessary because, other conditions being the same, those species that are genetically closer also tend to be geographically closer due to the presence of (even loose) constraints on the organisms' dispersal and the way that speciation occurs within a once continuous population. If this is the case, and the abiotic factors measured are spatially autocorrelated, more closely related species of viruses could occupy – on average – more similar eco-climate spaces simply due to their geographic proximity. More than controlling purely for a spurious association, by explicitly analysing sources of spatial correlation in the data we gain an additional tool to understand, measure and model spatial patterns (Legendre & Fortin 1989, Liebhold & Gurevitch 2002).

Methods

Genetic data

Dr. Edward C. Holmes (Zoology Department - Oxford University) kindly provided me with the matrix of phylogenetic distances (unpublished at the time) employed in his study (generated with the software PAUP 4.0b10 for Macintosh; PPC/Altivee – see Holmes & Twiddy 2003 and Appendix 3 of this thesis). The very selection of the species analysed in the present chapter was based upon these genetic data (Holmes & Twiddy 2003). Specifically, my investigation was focused in the branch of the phylogenetic tree reproduced in Figure 6.1.

Figure 6.1. Phylogeny of the JE Serocomplex (modified from Holmes & Twiddy, 2003). Colour key: viruses used for creating (red), testing (black) and not included (grey) in the model.

The virus species belonging to this phylogenetic were then allocated to two distinct sets, one consisting of those species that would be used to build the model (hereafter the training data-set), and the other comprising those used to test its predictive fit (hereafter the testing data-set). Because of the relatively small number of species available in this clade, using a full randomisation procedure to assign the species to either training or test set would lead to the omission of some branches of the clade, and to the overrepresentation of other ones. This would not be a problem if the environmental variables had a uniform effect across all groups and in every speciation event. Yet, this is probably not the case, and therefore the model has to be as representative as possible of the entire phylogeny analysed. Hence I have used a stratified randomisation: both for the training and testing data sets the species were selected as far as possible from the terminal points of each monophyletic branch of the phylogenetic tree, while nevertheless randomising the attribution of each pair of

neighbouring species to these sets . The resulting groups are shown with a bi-coloured key in Figure 6.1.

Environmental Data

The environmental information matching the distribution of the viruses was obtained by searching for records of the viruses' presence that were geographically well defined (at city, county, or other small territories units), and using these records to extract the relevant information from climatic metadata.

The records of presence were obtained mainly through an extensive search in the bibliography and in other reliable sources of information available in the Internet - such as Promed[3] and the software Gideon[4] - for the last 40 years. I did not distinguish (despite keeping this information in the data set) whether the presence of the virus was detected from mosquito samples or serum samples from humans and other domestic and wild vertebrates. This was based on the assumption that the presence of a virus in a given location would indicate that the climatic conditions in that particular location were favourable for its existence, regardless of the specificity of its association with a certain hosts and/or vectors. Each 'presence record' was geo-referenced using gazetteers available in the Internet[5]. Climate data were extracted for each grid-cell of viral presence record according to the methods described in Chapter 4 (page 58-9)

To obtain the matrices of environmental distances for all the variables to be considered in the model it was necessary to decide which parameters would be compared across species. Since only a few records were available for some virus species, it was not possible to make any inference or use any information related to the variance and covariance of the climatic data for each species. I therefore used Euclidean distances to calculate the environmental distances among the environmentally defined centroids of the species (Figure 6.2). Because the distribution of spatial data can depart substantially from normality and variable means are the very

[3] http://www.promedmail.org
[4] http://www.gideononline.com
[5] http://fat-albert.alexandria.ucsb.edu:8827/gazetteer/
http://www.getty.edu/research/tools/vocabulary/tgn/index.html
http://164.214.2.59/gns/html/cntry_files.html

parameters used to obtain the distances, all data were clustered for species with broad and diverse environmental profiles, with each cluster representing a separate multivariate description of environmentally similar fractions of a species distribution range (same methodology as that used in Chapter 4). Since Euclidean distances were used instead of Mahalanobis distances, a previous standardization of the data was conducted for each variable across all sample data, to avoid possible biases due to discrepancies in the magnitudes and scales of measurement.

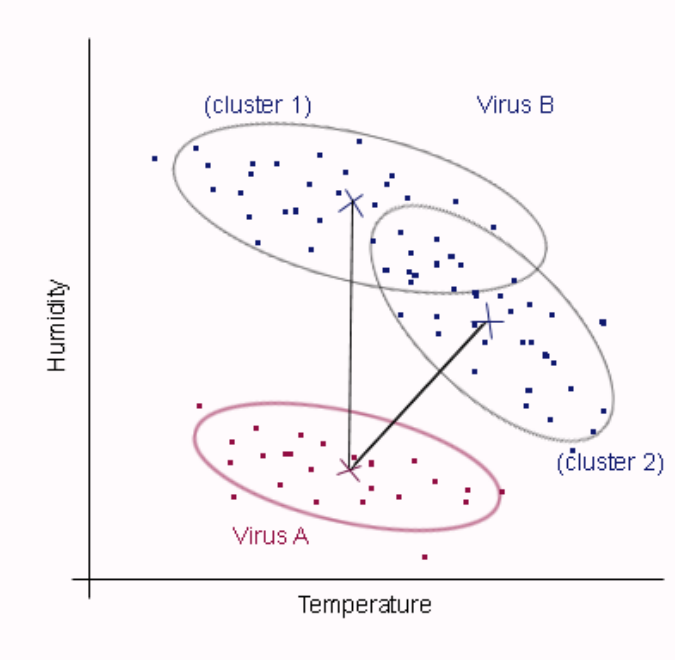


Figure 6.2 Empirical eco-climate spaces (area inside ellipses) defined for two hypothetic species of virus (A and B) and for two illustrative variables based on sample data (points inside and near the ellipses). There is one Euclidean distance between virus A and each cluster of virus B. The final distance representing the environmental distance between viruses A and B is the shortest one (in this case between the centroid of virus A and the centroid of cluster 2 of virus B).

Geographic Distance Data

As stated, to examine whether more closely related species also live under more similar environmental conditions it is also necessary to control for the differences in the geographic distances between the species, because it is possible that the more ancient the divergence of two species the greater their geographic separation. As longer geographic distances might correlate positively with a higher degree of environmental difference, a spurious correlation between genetic and environmental distances might be generated.

The present investigation deals, at a worldwide scale, with parasitic species that, as far as is known, base their cycles and transmission modes fully on terrestrial organisms. Before the increasingly intense movement of humans, livestock and commodities around the world, the dispersal of arboviruses had to rely most probably exclusively on the movement of their vertebrate hosts and/or vectors. Large surfaces of water should have imposed even greater barriers to this movement. For this reason, I considered the relative 'friction' (see below) associated with these movement patterns when establishing the geographic distances between the points of distribution of the arboviruses.

I used two matrices of 'geographic distances' to detect – within a range of different degrees of grain resolution – whether some spatial factors were affecting the observed environmental distances between species, as follows.

Cost distances between presence data

A least-cost path analysis technique was employed here to find the shortest geographic distance between records of presence of different species. The pushbroom algorithm (Eastman 1989) was used to create cost surfaces, where distances between points of presence are measured taking into account a friction component, which in the present case had values of 1-grid cell equivalents for land pixels, and 10-grid cell equivalents for sea pixels. In a second stage, the pathway algorithm establishes the least-cost route between the presence sites of a virus species 'A' and a species 'B' for which a cost surface has been calculated. An example of the results is shown in Figure 6.3, which shows the least-cost route linking the nearest presence points between the Australian Kunjin virus and the South American Rocio virus. The distances were calculated by summing the pixel values from the friction surface of the resulting route between each pair of viruses. To my knowledge, this technique has never been used before for establishing meaningful matrices of geographic distances in biological problems.

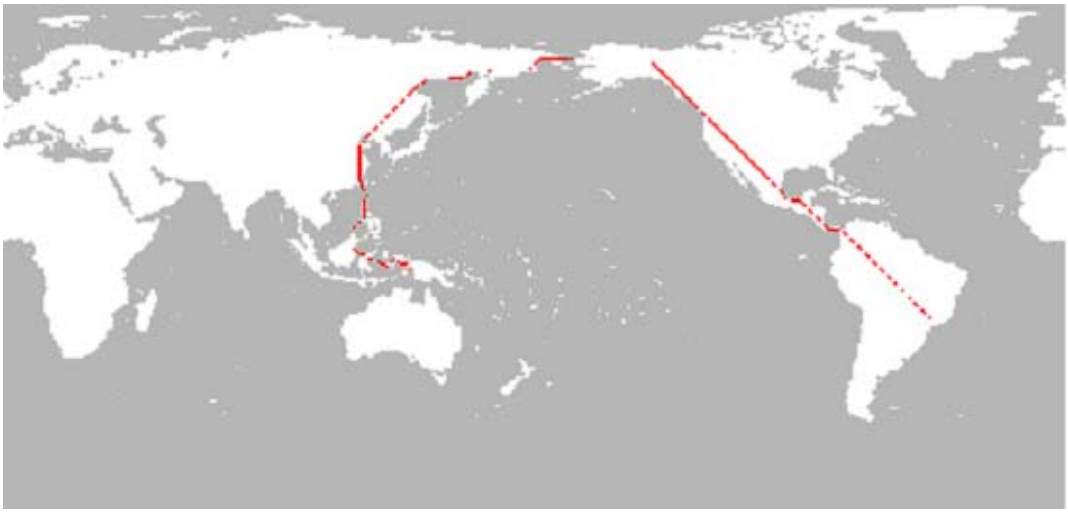


Figure 6.3. An example of a least-cost path linking two viruses, here Kunjin (Australia) and Rocio (South America), as revealed by the Pushbroom algorithm (see text).

Continental-share distances

The previous measure of distance was obtained directly using individual records of virus presence, and selecting those records for which the shortest distances between virus species were observed. It was therefore highly dependent upon the sampling effort shown for each virus. Here I created a matrix of Boolean distances based on the continental distribution of the viruses. Pairs of viruses belonging to the same continent were assigned a distance value of zero. When the species were not in a common continent a distance equal to one was assigned. The geographic information is summarized in Table 6.1.

Table 6.1. Continental masses where each virus was recorded. For illustrative purposes the species are disposed in the same order and following the major divisions as shown in the branches of their phylogeny (Figure 6.1).

Virus species	Continent
Alfuy	Australia
Japanese encephalitis	Asia
Murray Valley encephalitis	Australia - Asia
Usutu	Africa
Koutango	Africa
Kunjin	Australia - Asia
West Nile	Asia - Africa
Yaounde	Africa
Cacipacore	Americas
Bagaza	Africa
Israel-Turkey meningoencephalatis	Asia - Africa
Ntaya	Asia - Africa
Tembusu	Asia
Ilheus	Americas
Rocio	Americas
Kedougou	Africa
Spondweni	Africa
Zika	Asia - Africa
Saint Louis encephalitis	Americas
Dengue 1	Asia – Africa - Americas
Aroa	Americas
Iguape	Americas
Bussuquara	Americas
Naranjal	Americas
Kokbera	Australia - Asia
Stratford	Australia

Building the model

The generation of the matrix of environmental distances is a gradual process, where in each step all possible candidate variables are tested to obtain the one that, when included in the set of variables already selected, produces a matrix of Euclidean distances that better maximizes the correlation with the matrix of genetic distances.

Before explaining how this function is maximized, it is important to observe that the recurring calculation of the matrix of distances encompasses in itself a 'selection process'. As already described, the data set of presence of those species that had their distribution extended across very different climatic environments was submitted to a clustering process, with different parts of a species distribution range represented by separate clusters. This method generates a number of multivariate centroids (one corresponding to each cluster) for each species, and consequently a number n of measures of distance between pairs of species (where n is equal to the product of the numbers of clusters described for each of a species A and a species B).

The distance between each pair of species is taken as the minimum distance between the clusters of the different species – as proposed by Rogers (1998) for presence and absence classes. On the one hand, the method described here is computationally demanding, for it requires the calculation of all distances among all cluster centroids of all possible pairs of species in the selection of each variable. On the other hand, it is biologically most sensible, because the critical environmental differences (as revealed by the discriminant function), that effectively prevent a species A from occupying the same eco-climate space as a species B, should be captured more accurately if we compare ranges of the two species' distribution where the environmental conditions are most similar.

The variable selection process proceeds by calculating partial correlations between the matrices of phylogenetic distances (P) and environmental distances (E), while controlling for the effect of geographical distances (G), $r_{PE.G}$, using the formula (6.1) of first-order partial correlation coefficients (Smouse et al. 1986, Legendre & Legendre 1998, Legendre 2000), such that:

$$r_{PE.G} = \frac{r_{PE} - r_{PG}\, r_{EG}}{\sqrt{1 - r_{PG}^2}\ \sqrt{1 - r_{EG}^2}} \qquad [6.1]$$

where 'r' stands for the simple correlations (Pearson's coefficient) between the distance matrices P, E and G defined above. Decisions on i) the type of geographic control and ii) the variables to be included in the final model are based on a comparison of the curves obtained by plotting the simple and partial correlation coefficients against the number of variables selected at each step of the variable selection process (see Figure 6.6 below). The selected curve, as well as the number of variables included, must be the one that maximizes the value of the correlation coefficients with a minimum number of variables.

Testing the model

The predictive fit of the model was assessed with independent data (the testing set of species), by testing for a statistically significant correlation between the matrices of genetic and environmental distances built with the variables selected at the model

building stage. Spatial autocorrelation was controlled for using the same type of geographic matrix used to build the model.

The null hypothesis of no association between the genetic and environmental distances was assessed using partial Mantel tests (Mantel 1967, Smouse et al. 1986, Manly 1997, Legendre & Legendre 1998, Méot et al. 1998). Statistical distributions were generated by a Monte Carlo method, randomising rows (and corresponding columns) in the matrix of phylogenetic distances 10,000 times, and calculating for each one of these permutations the partial correlation coefficient between the two matrices, controlling for the third (geographic) matrix. The one-tailed probability of a type I error (i.e. rejection of a true null hypothesis) was taken as the proportion of correlation coefficients – sorted in ascending order – that were higher than or equal to the obtained correlation coefficient. I accepted as statistically significant those probabilities below $\alpha = 0.05$.

Software

Geographic distances obtained with the pushbroom algorithm were calculated using Idrisi Software. All other spatial analyses, graphics and statistical tests were conducted mostly using scripts written on Matlab specifically for this study. The relevant algorithms and programming codes developed are presented in appendix 4.

Results

Overall I collected a total of 1,782 records of *Flavivirus* presence. As shown in Figure 6.4, the number of records is highly imbalanced amongst the virus species, and amongst the world regions. For many species, presence records at the geographic resolution needed for the present study were rarely available, because in most cases publications only mention the names of larger administrative units where the samples were collected (such as provinces or countries, which represent too coarse a scale to be used in the analysis), with the name of villages, towns or counties only anecdotally provided. Some of the species are better registered, like West Nile and St. Louis encephalitis, especially in the USA. Out of the four Dengue serotypes, only Dengue 1 (for which a higher number of records was available) was considered. Although the

four distantly related subtypes of Dengue diverged a long time ago (apparently the result of a separate zoonotic transfer from non-human primates to humans) and have coexisted ever since (Zanotto et al. 1996, Holmes & Burch 2000, Twiddy et al. 2003), at present it is still not possible to distinguish the extent to which the differences and similarities between the eco-climate spaces occupied by these strains are due to historical or environmental factors (e.g. human intervention following the identification of clinic cases resulting from the presence of one serotype, and the very recent and continuing history of rapid occupation of different environments by this emergent disease - Rico-Hesse et al. 1997, WHO 2002). In addition, I have not considered the American records for the West Nile virus due to its recent occupation of the American continent (Campbell et al. 2002).

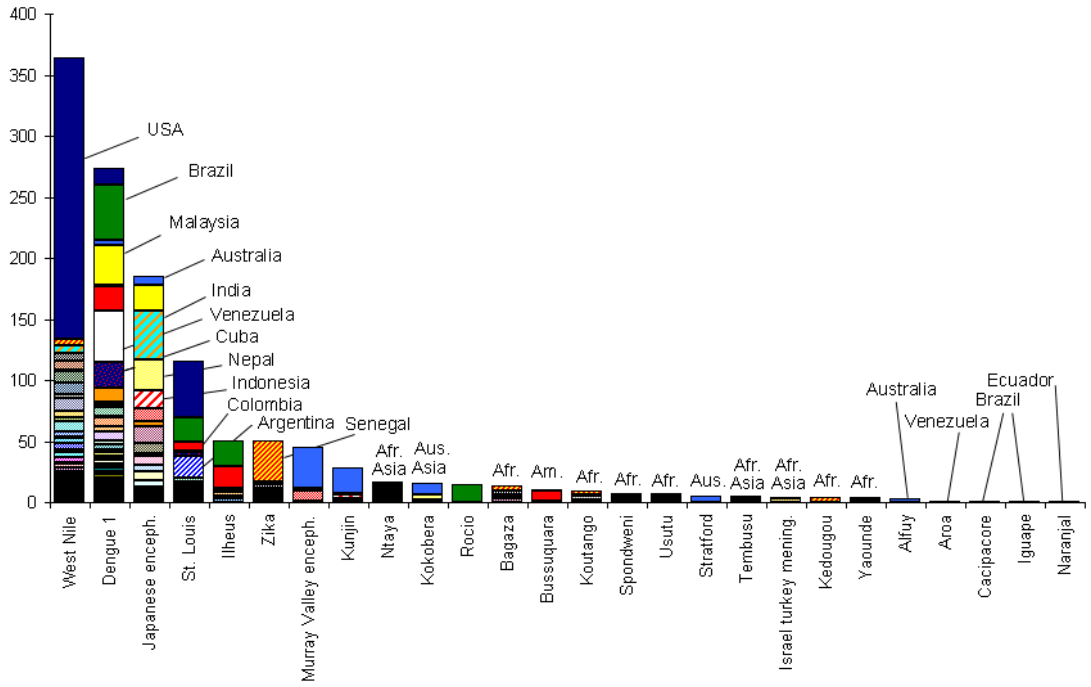


Figure 6.4 Number of records for each JE serocomplex species. The names of some countries or continents of origin of the records are displayed.

Since no information about the abundance of a species is used in the present study, repeated records of the same virus concurring at a same pixel (on the scale defined by the environmental surfaces employed) were ignored, leaving a final data set of 741 records.

The resulting mean values of each cluster for all the species and variables considered are shown in Appendix 2.

Geographic distance data

The Boolean matrices of distances resulting from the existence (or not) of common landmasses for pairs of species are easily inferred from the data shown in Table 6.1. The distances representing the geographical cost of movement between the nearest locations between two species are, however, better understood with the dendrogram of Figure 6.5, from which it is possible to distinguish three intuitively obvious major blocks. The first one includes those species with an Australian (Alfuy, Kokbera, Murray Valley and Stratford) and Asian (Japanese Encephalitis and Tembusu) distribution. The second block includes those species with a purely African distribution (Bagaza, Kedougou, Spondweni, Usutu, Yaounde), or distributed in Africa and other continents (Dengue 1, Israel-Turkey meningoencephalitis, Ntaya, Zika, West Nile). Finally the third group – the one also further apart from the previous two groups – includes the American species (Aroa, Bussuquara, Cacipacore, Ilheus, Iguape, Naranjal, St. Louis and Rocio). Therefore, the clusters produced by this method brings together species with close or overlapping geographic distribution, keeping apart those species separated by large distances and sea surfaces.

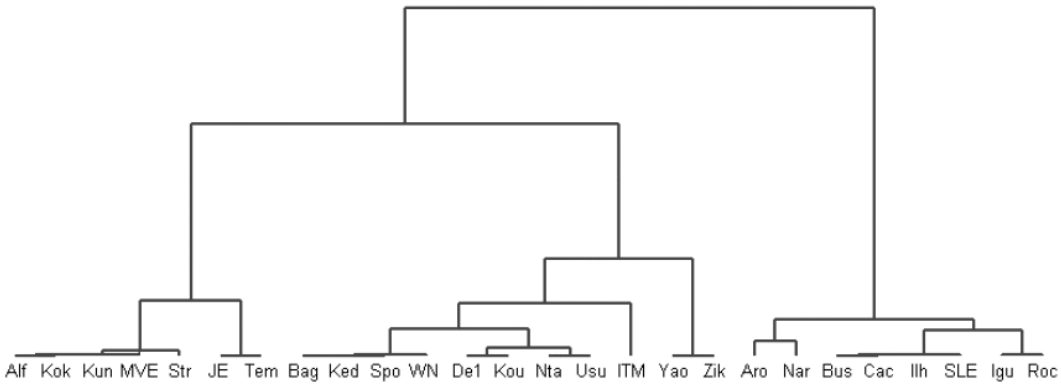


Figure 6.5. Hierarchical clusters of geographical cost distances. Linkage algorithm: 'complete'.

Environmental Data

Figure 6.6 shows the coefficients of correlation between the matrices of environmental and genetic distances obtained in the stepwise process of variable selection with the two geographic controls. The same first seven variables were

selected in all models, and the significance levels were in all cases <0.001. It is possible to see in the figure that the selection of climatic variables produced a reasonable model with correlation coefficients close to r = 0.63 in all controlled and non-controlled situations. The resulting asymptotic curve reaches stability with the inclusion of the fourth variable onwards. The model was thereby built with four variables: the amplitude of the annual cycle of the monthly frequency of wet days (wet-a1), the amplitude of the annual cycle of amount of precipitation per day (pre-a1), the minimum values of amount of precipitation per day (pre-mn) and the amplitude of the bi-annual cycle of the daily maximum temperatures (tmx-a2).

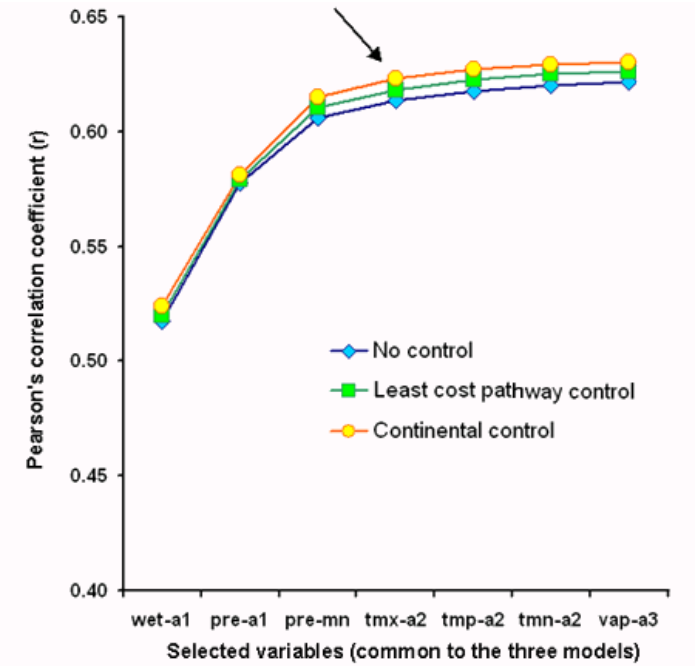
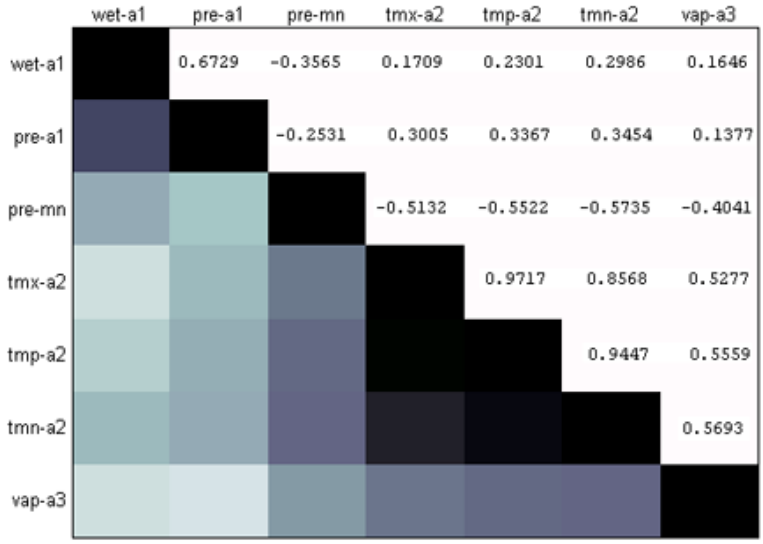


Figure 6.6. Coefficients of correlation between matrices of environmental and genetic distances for an increasing number of variables included during the stepwise process of variable selection. The different curves correspond to the different geographic controls.

Since the geographic controls did not have an influence on the structure of the model in this study, I decided not to use partial correlative methods and worked instead with simple correlations. Table 6.2 shows the correlation coefficients among the selected variables. The first two variables are well correlated (r = 0.67) – not surprisingly since both of them relate to different measures of the annual variation of rainfall. The third and fourth selected variables show weaker correlations with the previous ones. In contrast, the fifth and sixth variables are highly correlated (0.94-97)

with the fourth and fifth variables. This confirms the trend shown in Figure 6.6, that there would not be any considerable gain in information by including more than the first four variables.

Table 6.2. Matrix of correlation coefficients amongst the 7 variables first selected by the model. The lower triangular sub-matrix provides the same information, in a grey-tone scale, ranging from white (no correlation) to black (complete correlation).



The model

Based on these results, the matrix of environmental distances (Table 6.3) contains the nearest Euclidean distances between the clusters of the species belonging to the training data set. Using this matrix and the complete linkage algorithm (Legendre & Legendre 1998), I generated a dendrogram, and added histograms showing the distributions of the variables for each cluster of each species (Figure 6.7). This visual setting provides us not only with an overview of the relative distances between the species, but also direct access to the data from which the distances were calculated. For instance, the top cluster of the two shown for the virus Kunjin ('Kun' in the graph) is responsible for its positioning close to the viruses Usutu ('Usu'), since for this cluster the variable pre-mn is very similar to that shown for Usutu. This top cluster corresponds to the Australian distribution of Kunjin, where mean precipitations are lower than in the Indonesian cluster but similar to those found in the African distribution of the Usutu virus within the savannah region. The same kind of reasoning,

just changing the variables and values, can be extended to the other species and branches of the dendrogram.

Table 6.3. Matrix of environmental distances among the virus species considered during the model building process. Euclidean distances considering the four selected standardized variables were used to calculate the distances between the species. The lower triangular sub-matrix provides the same information, in a grey-tone scale, ranging from white (zero distance) to black (maximum distance in the matrix).

	Kun	Usu	JE	SLE	Zik	ITM	Kou	Bus	Roc	Tem	Aro	Ked	Kok
Kun		0.6	1.0	1.1	0.9	2.0	1.5	2.5	2.2	3.6	2.5	5.1	3.3
Usu			1.0	1.0	0.9	1.7	1.1	2.1	2.1	3.3	2.8	4.7	3.0
JE				1.1	0.9	2.5	1.8	1.7	1.2	2.9	3.1	4.7	2.1
SLE					0.6	2.1	1.8	0.7	1.2	1.9	1.5	4.7	2.2
Zik						1.3	0.9	2.0	1.6	3.2	2.4	3.0	2.7
ITM							1.0	3.1	3.0	3.0	3.6	4.0	3.7
Kou								2.5	2.9	3.2	3.7	3.8	2.8
Bus									1.9	2.3	4.3	4.4	2.3
Roc										2.5	2.8	5.7	4.0
Tem											4.8	4.3	3.9
Aro												7.4	3.4
Ked													3.2
Kok													

A visual comparison of this dendrogram with the phylogenetic tree shown in Figure 6.1 also enables us to infer the extent to which each species contributed to the degree of correlation obtained in the model. For instance, the environmental model grouped together the viruses Kunjin, Usutu and Japanese encephalitis, which also share the same phylogenetic branch. However, although phylogenetically the virus Koutango is grouped together with these three species, it is assigned to a different environmental branch.

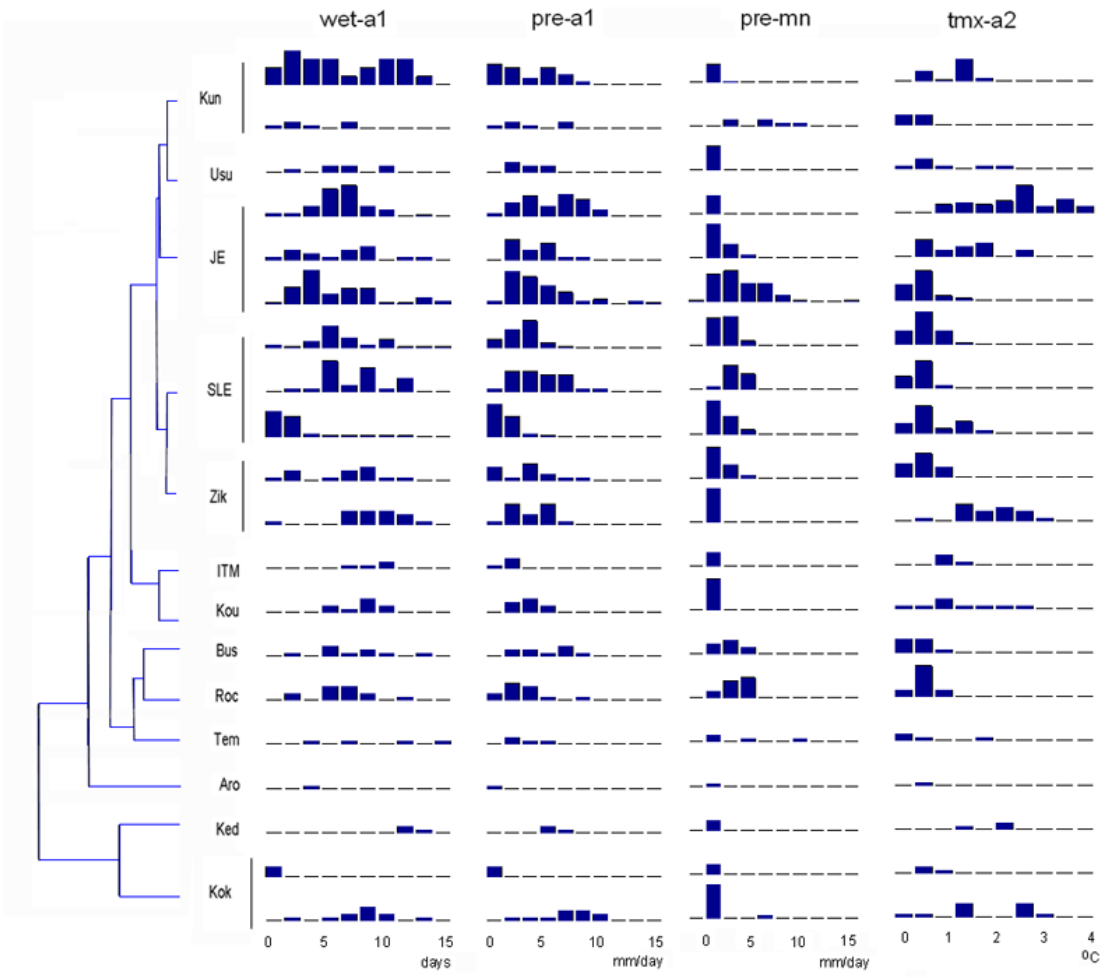


Figure 6.7. The left hand side of the figure represents the hierarchical cluster tree of Euclidean distances amongst the training-set of species, based on the four selected climate variables (linkage algorithm: complete distance). The right hand side of the figure shows the distribution of each selected variable for each species or cluster (when the data were partitioned).

Testing the model

When the model was applied to the testing set of species, no correlation was found between the genetic and the environmental distances among these species ($r = -0.03$, $p = 0.56$). Given the relatively low number of species available for testing the model, its power was too low (power = 0.28) to allow any definitive conclusions about the association at this stage. Nevertheless, the extremely low value obtained for the correlation coefficient indicates that any possible association between phylogenetic and environmental distances would be very low even with larger data sets. To confirm these results further, the model was run again, but this time the data were not clustered.

Instead, the Box-Cox normalization method (Box & Cox 1964, Sokal & Rohlf 1995) was applied to improve the normality of variable distributions. The model again selected the same four variables that were selected before, with the plateau of the correlation reached at a coefficient r = 0.61. As with the previous analyses, this test of the model also did not yield a significant correlation between the genetic and environmental distances for the testing set of species (r = -0.08, p = 0.3, randomised estimate of post-hoc power = 0.34).

Discussion

Beaty et al (1997) consider that, given the evolutionary potential of arboviruses resulting from the several available mechanisms for generating genetic diversity (e.g. high mutation rates consequent from being largely RNA viruses, establishment of new host opportunities through vectorial transmission, and development of persistent long-lasting and sometimes dual infections in vectors), it is remarkable that "*they have not taxonomically degenerated into a confusing array of antigenically distinct viruses. Rather, genetic stability frequently occurs over time and geographic space* (…)*. Apparently, constraints exist on the extent of genetic variation tolerated*". These constraints could operate at the level of phenotypic interaction of the virus with vertebrate and invertebrate host receptors, replication and maturation specificities, host immunity, vector feeding preferences (Beaty et al. 1997) and the very viability limit imposed by the amount of deleterious mutations that can be afforded by the genome at each replication cycle (Holmes 2003). If environmental factors – temperature, rainfall and other climatic variables, as well as their seasonal variations – are also a significant and directional force (i.e., leading to higher levels of differentiation) acting upon some of those components, then examining the environmental parameters underlying the distribution of these virus species could be a useful tool in understanding the evolution of a clade, complementary to genome sequence analysis.

Randolph (2000) foresaw the opportunities offered in the study of the environmental forces shaping the evolution of vector-borne pathogens through the use of GISs and remotely sensed (RS) data. The same success achieved in the study of the ecology of these species through the description of their spatial distribution on a

temporal scale comprising the life-span of vectors and hosts (Hay et al. 2000), could perhaps be achieved when addressing questions about the common evolutionary background of related species on a time scale relating to the cladogenesis of the group. Thereby the mapping of environmental profiles onto the phylogenetic tree of the group could be a pathway for the investigation of the combination of eco-climatic conditions guiding the evolutionary history of these organisms. Preliminary results from Randolph & Rogers (2002) suggest that some environmental variables captured by meteorological satellite imagery led to an eco-climatic classification of another group of *Flavivirus* (tick-borne encephalitis – TBE) consistent with their clinal evolution (Zanotto et al. 1995). In fact, the distribution of each of these viruses across Western Europe could be explained by the same 10 satellite-derived seasonal climatic variables, showing the evolution of the clade along a spatial-environmental cline (Randolph & Rogers 2002).

Here I followed the same pathway by studying the JEV-serocomplex, a group with a worldwide distribution consisting mainly of mosquito-borne viruses. However, although I was able to build environmental models representing the phylogenetic classification of the group with relatively high levels of correlation, no evidence of an association between these two sorts of classification could be detected when the model was tested with independent data. The correlation coefficients obtained for the test of the model were indeed so low that even though the power of the test was not high, they indicate that the strength of any possible association between phylogenetic and environmental distances would be very low even with larger data sets.

This is not to say that as yet I completely rule out the possibility of finding a combination of eco-climate variables that consistently correlates with the genetic distances in the group, since methodological limitations may have prevented the detection of an association. For instance, sampling biases affecting the observed distribution of the species may have led to distortions in the definition of their environmental profiles. Given the impossibility of implementing a sampling design at a global scale (a common shortcoming in this kind of study; Reutter et al., 2003), it was inevitable that for some species the available data showed a high bias: sampling effort was concentrated in some geographic regions, leaving an almost complete absence of

data in large areas of the globe. Symptomatic of this was the paucity of records for some of the species studied. As this lack of records of virus presence probably reflects these sampling problems more than the narrowness of the species' niche, I constrained the parameters used to compare the species in this analysis. For example, instead of using covariance matrices to define the niche of a species (Legendre & Legendre 1998, Rogers 2000, Reutter et al. 2003), I had to use only the mean value of each variable to prevent misrepresenting the extent of the range occupied by each species in the multidimensional space. Although average values are important parameters, they are probably insufficient to reflect accurately the subtleties of the differences between the species' climatic habitats. Also, the smaller the sample size, the greater the likelihood that mean values do not reflect the centroid of the ideal definition of the climatic habitat of a species - although this latter problem is attenuated when viewed from a statistical perspective, since each species corresponded to only one replicate within the data set used.

A second methodological limitation relates to the spatial scale used in the present study. On the one hand, a finer spatial scale would not be realistic, since many records mention the name of the districts, cities or hospitals where patients were diagnosed, and not the actual geographic locations (sometimes several kilometres distant from the diagnosis centres) where these individuals were infected and the virus circulated. In this sense, the granulation of the metadata used in this study (0.5 x 0.5 arc degree, New et al. 1999), giving pixels with a maximum area of 50x50 Km (at the Equatorial line) does seem to be a proper scale to deal with this lack of precision in the reports. On the other hand, these limitations might prove crucial for the type of investigation conducted here, since it is possible that the ideal climatic conditions for persistent virus transmission cycles might be only recognizable at micro-climate levels observed on a scale of fewer kilometres (Reiter 1988, Coetzee 2004). The large spatial scale employed in this study also constrains the use of the technique employed to process the climatic data, since studies conducted on a global scale cannot benefit from the use of the phase data generated by the Fourier transform of variables. This means that an important seasonal aspect relating to the time of the year when some climatic conditions change or reach their peak is lost. This aspect might be important for the specialization of viruses in different climates.

If, on the one hand, I may have lost accuracy by using a coarser spatial scale, on the other, by expanding the analysis to a worldwide scale and by controlling for geographic autocorrelation through the use of partial correlative methods, one can be confident that the relative geographic distances between the species were not underlying any potential correlation between eco-climatic and phylogenetic data. The phylogenetic tree shows some level of geographical clustering (Gaunt et al. 2001), making it possible to identify some cases of closely related species that are still found in the same continent, such as the South American Iguape, Aroa, Bussuquara and Naranjal. But while such spatial proximity is found at this level, it fades quickly when we expand, even if only slightly, the number of nodes considered – e.g. the group of South American viruses is closely related to the group of Australian Kokbera and Stratford viruses. As a result, spatial correlation at this scale did not in fact affect the analysis presented here to any noticeable extent. Certainly, the fact that the species studied are parasites or pathogens of flying creatures, having also recently evolved by exploiting human movement across continents, and that their vectors (*Culex* and *Aedes* mosquitoes) are globally dispersed, all contribute to their ability to disperse and occupy distant regions (Gould et al. 1997, Gaunt et al. 2001).

Although the above methodological shortcomings may have reduced the likelihood of detecting a correlation between the eco-space of the JEV-serocomplex viruses (at the scale studied) and their evolutionary history, the results presented here might indeed reflect a true lack of association between these two factors. In this case, we can conjecture why climatic conditions – important as they are for the population dynamics and biology of both vectors and hosts – would not have a noticeable impact upon the cladogenesis of these pathogens. One first possibility is that speciation processes in arboviruses might indeed lead to the gradual colonization of different environments, but a large fraction of these environments cannot be defined through the use of GIS techniques incorporating purely climatic factors. For instance, some of the ecological parameters underlying the separation of arbovirus species might be related to the vertical positioning of their vectors and hosts in the forest profile, with some virus species depending on hosts and vectors preferentially inhabiting the forest canopy, while others are found closer to the ground. Alternatively, the speciation process might

simply imply a change in the hosts' and vectors' habits from nocturnal to diurnal, or vice-versa (Woodall 1979, Degallier et al. 1998). No mean climate database could seize the relevant information involved in these cases.

Another possibility is that climatic components are indeed part of the selective pressures involved in the genesis of new virus species, but each speciation episode does not lead new organisms to occupy ecological niches progressively different from those occupied by their ancestors (e.g., they could instead sometimes reverse habits and occupy environments already inhabited by distant ancestors). Therefore, if a reversal to ancestral climatic profiles is a common event during speciation, we may face a very difficult – if not impossible – task of tracing a 'memory' of the cladogenesis of a group by looking at the current differences in the climates associated with its extant species. In fact, since the potential number of combinations of climatic factors in which a species circulate is not as high as the potential number of combinations available in the sequence of nucleotides, it is possible that the environmental information currently available to track a group's evolutionary history might prove to be disappointingly less enlightening than that provided by its genetic profiles.

Finally, it is important to consider the hypothesis of niche exclusion, namely the possibility that the lower the degree to which the viruses are serologically related, the higher the likelihood that they can share similar ecological niches (Degallier et al. 1998). This implies that, when ecological niches are identified and defined through the use of spatial distributions, similar species should be less likely to occupy the same region. In the case of Flaviviruses, a particular mechanism related to this phenomenon might be the presence of heterologous immunity, i.e. the acquired immune protection against other species of pathogens elicited by a previous infection from a related species. In fact, this immunological 'cross-reaction' to different species promotes 'competition' for hosts, so that a newly arriving virus may be confronted with hosts that have already developed immunity (in those regions where related virus species circulate). As a consequence, under contact with a new pathogen, some hosts could develop only low levels of viraemia, or even be completely refractory to infection (Goverdhan et al. 1992, Ilkal et al. 1994, Richman et al. 1997, Tesh et al. 2002). In both

cases, the chances of successful colonization by the newly arriving virus species are hampered.

Tesh et al. (2002) specifically point out that the unique and largely non-overlapping geographic distribution of various members of the JEV-serocomplex may be better understood by considering the interference from heterologous antibodies in birds and other vertebrate hosts. If this phenomenon is taken '*in extremis*', species belonging to the JEV-serocomplex are not specifically adapted to the particular climatic factors present in the environments they inhabit; instead, stochastic factors involved in the sequence of arrival of the species to a region could be primarily responsible for the observed distributions. In the context of the present study, this possibility mitigates against a positive association between environmental and genetic factors, therefore being supported by the findings reported here. Yet, given the '*post hoc*' nature of this justification, it should be further evaluated with independent data.

In terms of the methodological contribution of the present work, it was possible to assemble a series of analytical and statistical tools into a methodology that seems to be adequate for dealing with the research questions discussed. The potential of this 'phenetic discriminant analysis' is open to a deeper exploration when more extensive data sets become available. For instance, Mahalanobis distances could be used instead of Euclidean distances when the number of samples used allows a good definition of environmental matrices of covariance for all species. In addition, the model might be built liberally – allowing one to choose among different types of autocorrelation scales and hypotheses, select variables in the stepwise process described, analyse distributions and follow expert guidelines. The only requirement is that, once the model has been built, refined and complemented, a group of species should be used only once to test it. All the relevant programming algorithms and functions developed here were transformed in user-friendly Matlab (The Mathworks, Inc.) functions. These programs might prove useful for future studies dealing with the same questions with other groups of organisms – or even with species of the *Flavivirus* group when more extensive records of their distribution become available. All algorithms and functions developed will be made available in the Internet as an open code to be used by other researchers.

Chapter 7

CONCLUDING REMARKS

Few scientific activities can be said to hold as much responsibility in determining the fate of human lives as the study of infectious diseases. Even those scientists who, despite a primary interest in the study of the evolution, behaviour or ecology of some organisms, fortuitously verge upon the epidemiological field cannot avoid sensing the colossal proportions of the human tragedy behind the figures mentioned routinely in studies dealing with the vectors or pathogens involved in illnesses such as malaria, Trypanosomiases or arboviral encephalitis.

The exploration of novel paradigms in this area is, as in any other scientific enterprise, open to great joys, but also commonly to disappointment when the expectation of groundbreaking results is not substantiated by empirical data. Yet, for the very reasons exposed above, there can be no temptations in this area: results have to be described as clearly and completely as possible and the limitations associated with the methods used and possible shortcomings in their implementation always highlighted. In part due to circumstances set out in the Foreword (Chapter 1), and in part as a result of a personal choice to learn pioneering techniques in this field, the majority of the studies presented in this thesis were based on approaches still nearly unexplored, hence highly exposed to the inherent difficulties often associated with journeys on unsettled domains.

The study of vector preferences is widely acknowledged to be of fundamental epidemiological importance. But the possibility that some vector species possess cognitive abilities such as learning, that would affect their potential in the transmission of diseases, seems to be much less clear. The first part of this thesis was developed precisely within this field. Based on the assumption that the ability of vectors to memorize and learn features of their environment might itself lead to heterogeneity in vector-host contact, I investigated whether the mosquito vector of dengue and yellow fever viruses, *Aedes aegypti* was able to associate unconditional stimuli (blood feeding, human breath, vibration and electrical shock) with a wide range of specific odours and visual patterns to which it had previously been observed to be responsive (Chapter 2). The mosquitoes responded consistently to the experimental set-up and conditions, but the results did not indicate the existence of associative learning abilities and memory in

this species. While the recurrent failure to find evidence of associative learning under a broad range of conditions casts doubt on the possibility that adult of this species can be conditioned, it is nevertheless not possible to rule out the hypothesis that these conditions did not reproduce entirely the precise context in which such an ability would be observed. Future studies should in this sense prove fruitful in confirming (or not) these findings.

A careful review of the literature on learning and memory in mosquitoes (Chapter 3) revealed that, although several published reports suggested that the observed heterogeneity in mosquito behaviour resulted from the existence of memory and learning abilities, in fact only a few studies provided conclusive evidence of this. By critically highlighting the shortcomings and successes of the investigations conducted in this area, I believe that the review can contribute to a better assessment of the extent to which plasticity in behaviour is known to be due to learning in this important group of insects. I also provide directions and notes of caution in the design of future experiments. As argued in the Introduction, such investigations are a fruitful ground for interdisciplinary collaboration and, from an epidemiological perspective, tackle interesting questions concerning the evolution of the behaviour of hosts, vectors and even parasites.

In the second part of the thesis a different approach to explaining the epidemiology of mosquito-borne diseases was undertaken. Specifically, I explored the environmental factors potentially involved in the distribution, and also evolution, of pathogens through the use of geographic information systems (GIS). In Chapter 4, I investigated the eco-climatic conditions underlying the current distribution of the Japanese encephalitis virus and produced a risk-map for its global distribution by using the analytical tools of non-linear discriminant analysis. The predictive map showed that, of the hitherto unaffected regions of the world, an area of at least ~21.5 million Km^2 possesses eco-climatic conditions similar to those in the regions of Southeast Asia where this virus has been recorded. These regions are therefore environmentally suitable for the potential establishment of Japanese encephalitis. These locations correspond mostly to equatorial regions in Latin America and Africa currently inhabited by 556 million people. Furthermore, poor socio-economic conditions there

would constitute another factor facilitating the spread of the virus if it were introduced, potentially adding to the human burden given the shortage of resources needed to deal even with current public-health problems.

By investigating the robustness of non-linear discriminant function analysis (DFA) to the open choices involved in its implementation (Chapter 5), a number of effects became apparent. Unless predictions are produced only for unsurveyed regions that are environmentally very similar to the area of known distribution of the disease, I have shown just how much the predictions resulting from the use of DFA can differ depending upon the modelling parameters chosen by the researcher. Overall, relatively high levels of disagreement between spatial models involving different number of environmental variables or different clustering of the data can be observed. In addition, such effects can depend on the methodology used to select the variables that best describe the environmental profile of range of the organism(s) under study. These results highlight the importance of considering very carefully and justifying biologically such modelling choices when using these techniques either for predictive purposes or to infer the causal mechanisms underlying an organism's distribution, especially when the generation of predictions is not restricted by using the most rigorous criteria.

Finally, I explored those eco-climate factors potentially related to the cladogenesis of the Japanese Encephalitis serocomplex (a sub-group of the *Flavivirus* genus comprising mosquito-borne viruses) (Chapter 6). I assembled a set of techniques enabling me to make a comparison between the evolutionary history of the species as given by their phylogeny and their environmental profiles of their ranges as described through the use of GIS. The adjustments of the GIS technique proposed in this chapter allowed its direct use to search for the eco-climatic variables maximizing the correlation between environmental and phylogenetic distances, and provided a means for independently validating the final model produced. Although climatic conditions are critical for the population dynamics and biology of both vectors and hosts, my results indicate that current eco-climatic factors distinguishing the environmental profile of these virus species do not seem to be associated with the events leading to speciation in the JE serocomplex group.

The two different lines of research developed in this thesis both had implications for the epidemiological study of mosquito-borne pathogens. With their inherent limitations and promising potential, they represent two pathways that are worth taking in the important journey to understand these deadly, but also fascinating, evolutionary companions.

REFERENCES

Abramson, C.I. 1990. Invertebrate Learning: A Source Book. American Psychological Association.

Abramson, C.I. 1994. A Primer of Invertebrate Learning: The Behavioral Perspective. Washington, D.C.: American Psychological Association.

Alekseev, A.N. & Dubinina, H.V. 2002. Stability of parasitic systems under conditions of anthropogenic pressure. Contributions from the Zoological Institute St. Petersburg, 6.

Alonso, W.J. 2002. Epidemiology and ethology in the study of the infectious agents: a desirable mutual contamination. In: Anais XX Encontro Anual de Etologia (Ed. by Albuquerque, F. d. S.), pp. 221-229. Natal, Brazil: Universidade Federal do Rio Grande do Norte.

Anderson, R.M. & May, R.M. 1991. Infectious Diseases of Humans: Dynamics and Control. Oxford: Oxford University Press.

Aron, J.L. & May, R.M. 1982. The population dynamics of malaria. In: Population Dynamics of Infections Diseases (Ed. by Anderson, R. M.), pp. 139-179. London: Chapman and Hall.

Arredondo-Jimenez, J.I., Bown, D.N., Rodriguez, M.H., Villarreal, C., Loyola, E.G. & Frederickson, C.E. 1992. Tests for the existence of genetic determination or conditioning in host selection by Anopheles albimanus (Diptera, Culicidae). Journal of Medical Entomology, 29: 894-897.

Austin, M.P. 2002. Case studies of the use of environmental gradients in vegetation and fauna modelling: theory and practice in Australia and New Zeeland. In: Predicting Species Occurrence: Issues of Accuracy and Scale (Ed. by Scott, J. M., Heglund, P. J., Hauffler, J. B., Raphael, M. G., Wall, W. A. & Samson, F. B.), pp. 73-82. Washington: Island Press.

Baylis, M., Mellor, P.S., Wittmann, E.J. & Rogers, D.J. 2002. Prediction of areas around the Mediterranean at risk of bluetongue by modelling the distribution of its vector using satellite imaging (vol 149, pg 639, 2001). Veterinary Record, 150: 404-404.

Beaty, B.J., Borucki, M., Farlan, J. & White, D. 1997. Arbovirus-vector interactions: determinants of arbovirus evolution. In: Factors in the emergence of arbovirus diseases (Ed. by Saluzzo, J. F. & Dodet, B.), pp. 23-35. Paris: Elselvier.

Bi, P., Tong, S., Donald, K., Parton, K.A. & Ni, J. 2003. Climate variability and transmission of Japanese encephalitis in eastern China. Vector-Borne and Zoonotic Diseases, 3: 111-5.

Box, G.E. & Cox, D.R. 1964. An analysis of transformations. Royal Statistical Society B, 26: 211-252.

Brooker, S., Hay, S.I. & Bundy, D.A.P. 2002. Tools from ecology: useful for evaluation infection risk models? Trends in Parasitology, 18: 70-74.

Burke, D.S. & Leake, C.J. 1988. Japanese encephalitis. In: The arboviruses: Epidemiology and ecology (Ed. by Monath, T. P.), pp. 63-92. Florida: CRC Press.

Campbell, G.L., Marfin, A.A., Lanciotti, R.S. & Gubler, D.J. 2002. West Nile virus. The Lancet Infectious Diseases, 2: 519-529.

Canyon, D.V., Hii, J.L.K. & Muller, R. 1998. Multiple host-feeding and biting persistence of Aedes aegypti. Annals of Tropical Medicine and Parasitology, 92: 311-316.

Canyon, D.V., Hii, J.L.K. & Muller, R. 1999. The frequency of host biting and its effect on oviposition and survival in Aedes aegypti (Diptera : Culicidae). Bulletin of Entomological Research, 89: 35-39.

Carter, R., Mendis, K.N. & Roberts, D. 2000. Spatial targeting of interventions against malaria. Bulletin of the World Health Organization, 78: 1401-1411.

Charlwood, J.D., Graves, P.M. & Marshall, T.F.D. 1988. Evidence for a memorized home range in Anopheles farauti females from Papua New Guinea. Medical and Veterinary Entomology, 2: 101-108.

Charlwood, J.D., Pinto, J., Sousa, C.A., Madsen, H., Ferreira, C. & do Rosario, V.E. 2002. The swarming and mating behaviour of Anopheles gambiae s.s. (Diptera: Culicidae) from Sao Tome Island. Journal of Vector Ecology, 27: 178-83.

Charlwood, J.D., Thompson, R. & Madsen, H. 2003. Observations on the swarming and mating behaviour of Anopheles funestus from southern Mozambique. Malaria Journal, 2: 2.

CIESIN, Columbia University, IFPR & WRI. 2000. Gridded Population of the World (GPW), Version 2. Columbia University

Clements, A.N. 1992. The Biology of Mosquitoes. Development, Nutrition and Reproduction. London: Chapman & Hall.

Clements, A.N. 1999. The Biology of Mosquitoes. Sensory Reception and Behaviour. London: Chapman & Hall.

Coetzee, M. 2004. Distribution of the African malaria vectors of the Anopheles gambiae complex. American Journal of Tropical Medicine and Hygiene, 70: 103-104.

Cohen, J. 1988. Statistical power analysis for the behavioral sciences. New York: Erlbaum Associates.

Combes, C. 2001. Parasitism: the ecology and evolution of intimate interactions. Chicago; London: University of Chicago Press.

Cumming, G.S. 1999. The evolutionary ecology of African ticks. Oxford: Oxford University. D.Phil thesis.

Curtis, C.F. & Lines, J.D. 1987. Insecticides in the management of insect vectors of tropical disease. Insect Science and Its Application, 8: 709-714.

Davies, C. 1990. Interrupted feeding of blood-sucking insects - causes and effects. Parasitology Today, 6: 19-22.

Degallier, N., Travassos da Rosa, A.P.A., Vasconcelos, P.F.C., Sa Filho, G.C., Travassos da Rosa, E.S., Rodrigues, S.G. & Travassos da Rosa, J.F.S. 1998. Evolutionary aspects of the ecology of arboviruses in Brazilian Amazonia, South America. In: An overview of arbovirology in Brazil and neighbouring countries (Ed. by Travassos da Rosa, A. P. A., Vasconcelos, P. F. C. & Travassos da Rosa, J. F. S.), pp. 42-60. Belem: Instituto Evandro Chagas.

Dietz, K. 1988. Density-Dependence in Parasite Transmission Dynamics. Parasitology Today, 4: 91-97.

Dietz, K. 1993. The estimation of the basic reproduction number for infectious diseases. Statistical methods in Medical Research, 2: 23-41.

Dye, C. & Hasibeder, G. 1986. Population dynamics of mosquito-borne disease - effects of flies which bite some people more frequently than others. Transactions of the Royal Society of Tropical Medicine and Hygiene, 80: 69-77.

Eastman, J.R. 1989. Pushbroom algorithms for calculating distances in raster grids. In: Proceedings AUTOCARTO, pp. 288-297.

Edman, J.D. & Scott, T.W. 1987. Host defensive behavior and the feeding success of mosquitoes. Insect Science and Its Application, 8: 617-622.

Edman, J.D., Scott, T.W., Costero, A., Morrison, A.C., Harrington, L.C. & Clark, G.G. 1998. Aedes aegypti (Diptera : Culicidae) movement influenced by availability of oviposition sites. Journal of Medical Entomology, 35: 578-583.

Edman, J.D. & Spielman, A. 1988. Blood-feeding by vectors: physiology, ecology, behaviour and vertebrate defense. In: The arboviruses: Epidemiology and ecology (Ed. by Monath, T. P.), pp. 153-189. Florida: CRC Press.

Endy, T.P. & Nisalak, A. 2002. Japanese encephalitis virus: Ecology and epidemiology. In: Japanese Encephalitis and West Nile Viruses, pp. 11-48.

Ernst, R. & Heisenberg, M. 1999. The memory template in Drosophila pattern vision at the flight simulator. Vision Research, 39: 3920-3933.

FAO. 1996. Livestock Geography. A Demonstration of GIS techniques applied to Global Livestock Systems and Populations. Consultancy Report prepared by Environmental Research Group Oxford Ltd for the Animal Health Service of the Animal Production and Health Division of the Food and Agriculture Organisation of the United Nations, Rome (Author W.Wint).

Fielding, A.H. & Bell, J.F. 1997. A review of the methods for assessment of prediction errors in conservation presence/absence models. Environmental Conservation, 24: 38-49.

Focks, D.A. 1988. Quantitative models of arboviruses infection. In: The arboviruses: Epidemiology and ecology (Ed. by Monath, T. P.), pp. 311-318. Florida: CRC Press.

Focks, D.A., Haile, D.G., Daniels, E. & Mount, G.A. 1993. Dynamic Life Table Model For Aedes aegypti (Diptera, Culicidae) - Simulation and Validation. Journal of Medical Entomology, 30: 1018-1028.

Fonseca, D.M., Keyghobadi, N., Malcolm, C.A., Mehmet, C., Schaffner, F., Mogi, M., Fleischer, R.C. & Wilkerson, R.C. 2004. Emerging Vectors in the Culex pipiens Complex. Science, 303: 1535-1538.

Frank, S.A. 1996. Models of parasite virulence. Quarterly Review of Biology, 71: 37-78.

Galvani, A.P. 2003. Epidemiology meets evolutionary ecology. Trends in Ecology & Evolution, 18: 132-139.

Galvão, C. 2002. Studies on the behaviour of triatomines (Hemiptera, Reduvidae), vectors of the Chagas disease. In: Anais XX Encontro Anual de Etologia (Ed. by Albuquerque, F. d. S.), pp. 212-220. Natal, Brazil: Universidade Federal do Rio Grande do Norte.

Gaunt, M.W., Sall, A.A., de Lamballerie, X., Falconar, A.K.I., Dzhivanian, T.I. & Gould, E.A. 2001. Phylogenetic relationships of flaviviruses correlate with their epidemiology, disease association and biogeography. Journal of General Virology, 82: 1867-1876.

Gibson, G. & Torr, S.J. 1999. Visual and olfactory responses of haematophagous Diptera to host stimuli. Medical and Veterinary Entomology, 13: 2-23.

Giglioli, M.E.C. 1964. The influence of irregularities in the bush perimeter of the cleared agricultural belt around a Gambian village on the flight range and direction of approach of a population of Anopheles melas. In: XII International Congress of Entomology (Ed. by Freeman, P.). London.

Giglioli, M.E.C. 1965. The age composition of Anopheles melas Theobald (1903) populations collected simultaneously by different methods in The Gambia, West Africa. Cahiers O.R.S.T.O.M., 3: 11-26.

Gillies, M.T. 1964. Selection for host preference in Anopheles gambiae. Nature, 203: 852-854.

Gillies, M.T. 1972. Some aspects of mosquito behaviour in relation to the transmission of parasites. In: Behavioural Aspects of Parasite Transmission (Ed. by Canning, E. U. & Weight, C. A.). London: Academic Press Inc.

Gillies, M.T. 1988. Anopheline mosquitoes: vector behavior and bionomics. In: Malaria (Ed. by Wernsdorfer, W. H. & Gregor, I. M.), pp. 453-485. Edinburgh: Churchill and Living-stone.

Ginzburg, L.R. & Jensen, C.X.J. Rules of thumb for judging ecological theories. Trends in Ecology & Evolution, In Press, Corrected Proof.

Gould, E.A. 2002. Evolution of the Japanese encephalitis serocomplex viruses. In: Japanese Encephalitis and West Nile Viruses (Ed. by Mackenzie, J. S., Barrett, A. D. T. & Deubel, V.), pp. 391-404. Berlin, London: Springer.

Gould, E.A., de Lamballerie, X., Zanotto, P.M.D. & Holmes, E.C. 2001. Evolution, epidemiology, and dispersal of flaviviruses revealed by molecular phylogenies. In: Advances in Virus Research, pp. 71-103.

Gould, E.A., Zanotto, P.M.D. & Holmes, E.C. 1997. The genetic evolution of flaviviruses. In: Factors in the emergence of arbovirus diseases (Ed. by Saluzzo, J. F. & Dodet, B.), pp. 51-63. Paris: Elselvier.

Gould, S.J. & Vrba, E.S. 1982. Exaptation - a missing term in the science of form. Paleobiology., 8: 4-15.

Goverdhan, M.K., Kulkarni, A.B., Gupta, A.K., Tupe, C.D. & Rodrigues, J.J. 1992. Two-way cross-protection between West Nile and Japanese encephalitis viruses in bonnet macaques. Acta Virologica, 36: 277-283.

Gubler, D.J., Reiter, P., Ebi, K.L., Yap, W., Nasci, R. & Patz, J.A. 2001. Climate variability and change in the United States: Potential impacts on vector- and rodent-borne diseases. Environmental Health Perspectives, 109: 223-233.

Hales, S., de Wet, N., Maindonald, J. & Woodward, A. 2002. Potential effect of population and climate changes on global distribution of dengue fever: an empirical model. Lancet, 360: 830-834.

Hanna, J.N., Ritchie, S.A., Phillips, D.A., Lee, J.M., Hills, S.L., van den Hurk, A.F., Pyke, A.T., Johansen, C.A. & Mackenzie, J.S. 1999. Japanese encephalitis in north Queensland, Australia, 1998. Medical Journal of Australia, 170: 533-536.

Hart, B.L. 1990. Behavioral adaptations to pathogens and parasites - 5 strategies. Neuroscience and Biobehavioral Reviews, 14: 273-294.

Hart, B.L. 1992. Behavioral Adaptations to Parasites - an Ethological Approach. Journal of Parasitology, 78: 256-265.

Hasibeder, G. & Dye, C. 1988. Population-Dynamics of Mosquito-Borne Disease - Persistence in a Completely Heterogeneous Environment. Theoretical Population Biology, 33: 31-53.

Hay, S.I., Randolph, S.E. & Rogers, D.J. 2000. Remote Sensing and Geographical Information Systems in Epidemiology. In: Advances in Parasitology (Ed. by Baker, J. R., Muller, R. & Rollinson, D.), pp. 357. London: Academic Press.

Heinz, F.X., Collett, M.S., Purcell, R.H., Gould, E.A., Howard, C.R., Houghton, M., Moorman, T.B., Rice, C.M. & Thiel, H.J. 2000. Family Flaviviridae. In: Virus taxonomy: classification and nomenclature of viruses (Ed. by MHV, v. R., CM, F., DHL, B., EB, C., MK, E. & Lemon SM, e. a.), pp. 859-878. San Diego: Academic Press.

Hellriegel, B. 2001. Immunoepidemiology - bridging the gap between immunology and epidemiology. Trends in Parasitology, 17: 102-106.

Hendrickx, G., Napala, A., Slingenbergh, J.H.W., De Deken, R., Vercruysse, J. & Rogers, D.J. 2000. The spatial pattern of trypanosomosis prevalence predicted with the aid of satellite imagery. Parasitology, 120: 121-134.

Hii, J.L.K. 1985. Evidence for the existence of genetic variability in the tendency of Anopheles balabacensis to rest in houses and to bite man. Southeast Asian Journal of Tropical Medicine and Public Health, 16: 173-182.

Hii, J.L.K., Chew, M., Sang, V.Y., Munstermann, L.E., Tan, S.G., Panyim, S. & Yasothornsrikul, S. 1991. Population genetic-analysis of host seeking and resting behaviors in the malaria vector, Anopheles balabacensis (Diptera, Culicidae). Journal of Medical Entomology, 28: 675-684.

Hii, J.L.K. & Vun, Y.S. 1987. The influence of a heterogeneous environment on host feeding behaviour by Anopheles balabacensis (Diptera: Culicidae). Tropical Biomedicine, 4: 67-70.

Hildebrand, J., Bock, G.R. & Cardew, G. 1996. Olfaction in Mosquito-Host Interactions. Chichester: John Willey & Sons.

Hirzel, A.H., Hausser, J., Chessel, D. & Perrin, N. 2002. Ecological-niche factor analysis: How to compute habitat- suitability maps without absence data? Ecology, 83: 2027-2036.

Holmes, E.C. 2003. Error thresholds and the constraints to RNA virus evolution. Trends in Microbiology, 11: 543-6.

Holmes, E.C. & Burch, S.S. 2000. The causes and consequences of genetic variation in dengue virus. Trends in Microbiology, 8: 74-7.

Holmes, E.C. & Twiddy, S.S. 2003. The origin, emergence and evolutionary genetics of dengue virus. Infection, Genetics and Evolution, 3: 19-28.

Hurlbert, S.H. 1984. Pseudoreplication and the design of ecological field experiments. Ecological Monographs, 54: 187-211.

Hutchinson, G.E. 1957. Concluding remarks. In: Population Studies: Animal Ecology and Demography, pp. 415-427: Cold Spring Harbor Symposia on Quantitative Biology.

Hutchinson, G.E. 1965. The ecological theatre and the evolutionary play. New Haven: Yale University Press.

Ilkal, M.A., Prasanna, Y., Jacob, P.G., Geevarghese, G. & Banerjee, K. 1994. Experimental studies on the susceptibility of domestic pigs to West Nile virus

followed by Japanese encephalitis virus infection and vice versa. Acta Virologica, 38: 157-161.

IPCC. 2001. Climate Change 2001: Impacts, adaptation and vulnerability. Contribution of working group II to the third assessment report of the intergovernmental panel on climate change. Cambridge: Cambridge University Press.

Jaenike, J. & Papaj, D.R. 1993. Behavioral plasticity and patterns of host use by insects. In: Insect Learning: Ecological and Evolutionary Perspectives (Ed. by Papaj, D. R. & Lewis, A. C.), pp. 245-264. New York: Chapman & Hall.

Jones, J.C. & Pilitt, D.R. 1973. Blood-feeding behavior of adult Aedes aegypti mosquitoes. Biological Bulletin, 145: 127-139.

Karabatsos, N. 1985. International catalogue of arboviruses including certain other viruses of vertebrates. Santo Antonio: American Society of Tropical Medicine and Hygiene.

Kaur, J.S., Lai, Y.L. & Giger, A.D. 2003. Learning and memory in the mosquito Aedes aegypti shown by conditioning against oviposition deterrence. Medical and Veterinary Entomology, 17: 457-60.

Kellogg, F.E. & Wright, R.H. 1962. The guidance of flying insects. V. Mosquito attraction. The Canadian Entomologist, 94: 1009-1016.

Kelly, D.W. 2001. Why are some people bitten more than others? Trends in Parasitology, 17: 578-581.

Kelly, D.W. & Thompson, C.E. 2000. Epidemiology and optimal foraging: modelling the ideal free distribution of insect vectors. Parasitology, 120: 319-327.

Krakauer, D.C. & Rodriguez-Gironez, M. 1995. Searching and learning in a random environment. Journal of Theoretical Biology, 177: 417-429.

Kuhn, K.G., Campbell-Lendrum, D.H. & Davies, C.R. 2002. A continental risk map for malaria mosquito (Diptera : Culicidae) vectors in Europe. Journal of Medical Entomology, 39: 621-630.

Kuno, G., Chang, G.-J.J., Tsuchiya, K.R., Karabatsos, N. & Cropp, C.B. 1998. Phylogeny of the Genus Flavivirus. Jounal of Virology, 72: 73-83.

Legendre, P. 2000. Comparison of permutation methods for the partial correlation and partial Mantel tests. Journal of Statistical Computation and Simulation, 67: 37-73.

Legendre, P. & Fortin, M.-J. 1989. Spatial pattern and ecological analysis. Vegetatio, 80: 107–138.

Legendre, P. & Legendre, L. 1998. Numerical ecology. Amsterdam; Oxford: Elsevier.

Levin, B.R., Lipsitch, M. & Bonhoeffer, S. 1999. Population Biology, Evolution, and Infectious Disease: Convergence and Synthesis. Science, 283: 806-809.

Levine, R.S., Peterson, A.T. & Benedict, M.Q. 2004. Geographic and ecologic distributions of the Anopheles gambiae complex predicted using a genetic algorithm. American Journal of Tropical Medicine and Hygiene, 70: 105-109.

Liebhold, A.M. & Gurevitch, J. 2002. Integrating the statistical analysis of spatial data in ecology. Ecography, 25: 553–557.

Macdonald, G. 1957. The epidemiology and control of malaria. London: Oxford University Press.

Mackenzie, J.S., Barrett, A.D.T. & Deubel, V. 2002. The Japanese encephalitis serological group of flaviviruses: a brief introduction to the group. In: Japanese Encephalitis and West Nile Viruses, pp. 1-10.

Mackintosh, J. 1974. The Psychology of Animal Learning. New York: Academic Press.

Maclure, M. & Willett, W.C. 1987. Misinterpretation and misuse of the kappa statistic. Am J Epidemiol, 126: 161-9.

Mahalanobis, P.C. 1936. On the generalized distance in statistics. Proceedings of the Mathematical Institute of Science of India, 2: 49-55.

Manly, B. 1997. Randomization, bootstrap and Monte Carlo methods in biology. Chapman and Hall.

Mantel, N. 1967. The detection of disease clustering and a generalized regression approach. Cancer Research, 27: 209-220.

Martens, P., Kovats, R.S., Nijhof, S., de Vries, P., Livermore, M.T.J., Bradley, D.J., Cox, J. & McMichael, A.J. 1999. Climate change and future populations at risk of malaria. Global Environmental Change, 9: S89-S107.

May, R.M. & Anderson, R.M. 1983. Epidemiology and genetics in the coevolution of parasites and hosts. Proceedings of the Royal Society of London Series B-Biological Sciences, 219: 281-313.

McCall, P.J. & Eaton, G. 2001. Olfactory memory in the mosquito Culex quinquefasciatus. Medical and Veterinary Entomology, 15: 197-203.

McCall, P.J. & Kelly, D.W. 2002. Learning and memory in disease vectors. Trends in Parasitology, 18: 429-433.

McCall, P.J., Mosha, F.W., Njunwa, K.J. & Sherlock, K. 2001. Evidence for memorized site-fidelity in Anopheles arabiensis. Transactions of the Royal Society of Tropical Medicine and Hygiene, 95: 587-590.

Mellor, P.S. & Leake, C.J. 2000. Climatic and geographic influences on arboviral infections and vectors. Revue Scientifique et Technique de l'Office International des Epizooties, 19: 41-54.

Méot, A., Legendre, P. & Borcard, D. 1998. Partialling out the spatial component of ecological variation: questions and propositions in the linear modeling framework. Environmental and Ecological Statistics, 5: 1-27.

Milligan, G.W. & Cooper, M.C. 1987. Methodology Review - Clustering Methods. Applied Psychological Measurement, 11: 329-354.

Milligan, G.W. & Cooper, M.C. 1988. A Study of Standardization of Variables in Cluster-Analysis. Journal of Classification, 5: 181-204.

Mitchell, C.J. 1988. Occurrence, biology, and physiology of diapause in overwintering mosquitoes. In: The arboviruses: Epidemiology and ecology (Ed. by Monath, T. P.), pp. 191-217. Florida: CRC Press.

Mogi, M. 1996. Overwintering strategies of mosquitoes (Diptera:Culicidae) on warmer islands may predict impact of global warming on Kyushu, Japan. Journal of Medical Entomology, 33: 438-44.

Monath, T.P. 1990. Flaviviruses. In: Virology (Ed. by Fields, B. N. & Knipe, D. M.), pp. 763-814. New York: Raven Press.

Moore, J. 2002. Parasites and the behavior of animals. Oxford: Oxford University Press.

Muir, L.E., Kay, B.H. & Thorne, M.J. 1992a. Aedes aegypti (Diptera, Culicidae) vision - response to stimuli from the optical environment. Journal of Medical Entomology, 29: 445-450.

Muir, L.E., Thorne, M.J. & Kay, B.H. 1992b. Aedes aegypti (Diptera, Culicidae) vision - spectral sensitivity and other perceptual parameters of the female eye. Journal of Medical Entomology, 29: 278-281.

Murkaya, L. 1977. Genetic control of feeding preferences in mosquitoes Aedes (Stegomya) simpsoni and aegypti. Physiological Entomology, 2.

Mwandawiro, C., Boots, M., Tuno, N., Suwonkerd, W., Tsuda, Y. & Takagi, M. 2000. Heterogeneity in the host preference of Japanese encephalitis vectors in Chiang Mai, northern Thailand. Transactions of the Royal Society of Tropical Medicine and Hygiene, 94: 238-242.

New, M., Hulme, M. & Jones, P. 1999. Representing twentieth-century space-time climate variability. Part I: Development of a 1961-90 mean monthly terrestrial climatology. Journal of Climate, 12: 829-856.

Nielsen, E.T. & Nielsen, A.T. 1953. Field observations on the habits of Aedes taeniorhynchus. Ecology, 34: 141-156.

Nutsathapana, S., Sawasdiwongphorn, P., Chitprarop, U., Cullen, J.R., Gass, R.F. & Green, C.A. 1986. A mark-release-recapture demonstration of host-preference heterogeneity in Anopheles minimus Theobald (Diptera, Culicidae) in a Thai village. Bulletin of Entomological Research, 76: 313-320.

Papaj, D.R. & Lewis, A.C. 1993. Insect Learning: Ecological and Evolutionary Perspectives. New York: Chapman & Hall.

Pearce, J.N. 1997. Animal Learning and Cognition. New York: Psychology Press.

Quinn, W.G., Harris, W.A. & Benzer, S. 1974. Conditioned behavior in Drosophila melanogaster. Proceedings of the National Academy of Sciences of the United States of America, 71: 708-712.

Randolph, S.E. 1998. Ticks are not insects: Consequences of contrasting vector biology for transmission potential. Parasitology Today, 14: 186-192.

Randolph, S.E. 2000. Ticks and tick-borne disease systems in space and from space. In: Remote Sensing and Geographical Information Systems in Epidemiology (Ed. by Hay, S. I., Randolph, S. E. & Rogers, D. J.), pp. 217-243. London: Academic Press.

Randolph, S.E. & Rogers, D.J. 2000. Fragile transmission cycles of tick-borne encephalitis virus may be disrupted by predicted climate change. Proceedings of the Royal Society of London Series B-Biological Sciences, 267: 1741-1744.

Randolph, S.E. & Rogers, D.J. 2002. Remotely sensed correlates of phylogeny: tick-borne flaviviruses. Experimental and Applied Acarology, 28: 231-237.

Rawlings, P. & Curtis, C.F. 1982. Tests for the existence of genetical variability in the tendency of Anopheles culicifacies species B to rest in houses and to bite man. Bulletin of the World Health Organization, 60: 427-432.

Reiter, P. 1988. Weather, vector biology, and arboviral recrudescence. In: The arboviruses: Epidemiology and ecology (Ed. by Monath, T. P.), pp. 245-280. Florida: CRC Press.

Reiter, P. 1992. Status of current Aedes aegypti control methodologies. In: Dengue: a worldwide problem, a common strategy (Ed. by Halstead, S. B. & Gomez-Dantes, H.), pp. 41-48. Mexico City: Ediciones Copilco.

Reiter, P., Amador, M.A., Anderson, R.A. & Clark, G.G. 1995. Dispersal of Aedes aegypti in an urban area after blood-feeding as demonstrated by rubidium-marked eggs. American Journal of Tropical Medicine and Hygiene, 52: 177-179.

Renshaw, M., Service, M.W. & Birley, M.H. 1994. Host finding, feeding patterns and evidence for a memorized home-range of the mosquito Aedes cantans. Medical and Veterinary Entomology, 8: 187-193.

Reutter, B.A., Helfer, V., Hirzel, A.H. & P., V. 2003. Modelling habitat-suitability on the base of museum collections: an example with three sympatric Apodemus species from the Alps. Journal of Biogeography, 30: 581-590.

Ribbands, C.R. 1949. Studies on the attractiveness of human populations to anophelines. Bulletin of Entomological Research, 40: 227-238.

Richman, D.D., Whitley, R.J. & Hayden, F.G. 1997. Clinical Virology. pp. 1355. New York, Edinburgh: Churchill Livingstone.

Rico-Hesse, R., Harrison, L.M., Salas, R.A., Tovar, D., Nisalak, A., Ramos, C., Boshell, J., de Mesa, M.T.R., Nogueira, R.M.R. & Rosa, A.T.d. 1997. Origins of dengue type 2 viruses associated with increased pathogenicity in the Americas. Virology, 230: 244-251.

Robertson, M.P., Caitheness, N. & Villet, M.H. 2001. A PCA-based modelling technique for predicting environmental suitability for organisms from presence records. Diversity and Distributions, 7: 15-17.

Robinson, T., Rogers, D. & Williams, B. 1997. Mapping tsetse habitat suitability in the common fly belt of Southern Africa using multivariate analysis of climate and remotely sensed vegetation data. Medical and Veterinary Entomology, 11: 235-245.

Robinson, T.P. 2000. Spatial statistics and geographical information systems in epidemiology and public health. In: Remote Sensing and Geographical Information Systems in Epidemiology (Ed. by Hay, S. I., Randolph, S. E. & Rogers, D. J.), pp. 81-128. London: Academic Press.

Rogers, D.J. 1988a. The dynamics of vector-transmitted diseases in human communities. Philosophical Transactions of the Royal Society of London Series B-Biological Sciences, 321: 513-539.

Rogers, D.J. 1988b. A general-model for the African Trypanosomiases. Parasitology, 97: 193-212.

Rogers, D.J. 1997. Satellite imagey and the prediction of tsetse distribuions in East Africa. In: Towards Livestock Disease Diagnosis and Control in the 21st Century, pp. 602. Vienna: International Atomic Energy Agency.

Rogers, D.J. 1998. Satellite imagery and the prediction of tsetse distributions in East Africa. In: Towards Livestock Disease Diagnosis and Control in the 21st Century, pp. 602. Vienna: International Atomic Energy Agency.

Rogers, D.J. 2000. Satellites, space, time and the African trypanosomiases. In: Remote Sensing and Geographical Information Systems in Epidemiology, pp. 129-171. London: Academic Press.

Rogers, D.J., Hay, S.I. & Packer, M.J. 1996. Predicting the distribution of tsetse flies in West Africa using temporal Fourier processed meteorological satellite data. Annals of Tropical Medicine and Parasitology, 90: 225-241.

Rogers, D.J. & Packer, M.J. 1993. Vector-Borne Diseases, Models, and Global Change. Lancet, 342: 1282-1284.

Rogers, D.J. & Randolph, S.E. 2000. The global spread of malaria in a future, warmer world. Science, 289: 1763-1766.

Rogers, D.J. & Randolph, S.E. 2003. Studying the global distribution of infectious diseases using GIS and RS. Nature Reviews - Microbiology, 1: 231-237.

Rogers, D.J., Randolph, S.E., Snow, R.W. & Hay, S.I. 2002. Satellite imagery in the study and forecast of malaria. Nature, 415: 710-715.

Rogers, D.J. & Williams, B.G. 1994. Tse-tse distribution in Africa: seeing the wood and the trees. In: Large-scale Ecology and Conservation Biology (Ed. by Edwards, P. J., May, R. M. & Webb, N. R.), pp. 249-273. Oxford: Blackwell Scientific.

Rosenheim, J.A. 1993. Comparative and experimental approaches to understanding learning. In: Insect Learning: Ecological and Evolutionary Perspectives (Ed. by Papaj, D. R. & Lewis, A. C.), pp. 273-307. New York: Chapman & Hall.

Ross, R. 1911. The prevention of malaria. London: Murray.

Sabesan, S. 2003. Forecasting mosquito abundance to prevent Japanese encephalitis. Current Science, 84: 1172-1173.

Service, M.W. 1993. Mosquito Ecology: Field Sampling Methods. London: Chapman & Hall.

Service, M.W. 1994. Male Swarming of the Mosquito Culex (Culex) torrentium in England. Medical and Veterinary Entomology, 8: 95-98.

Service, M.W. 1997. Mosquito (Diptera : Culicidae) dispersal - the long and short of it. Journal of Medical Entomology, 34: 579-588.

Silva, M.A.N. 2002. Environmental impacts and their implications on the populations of Culicidae. In: Anais XX Encontro Anual de Etologia (Ed. by Albuquerque, F. d. S.), pp. 240-247. Natal, Brazil: Universidade Federal do Rio Grande do Norte.

Skelly, C. & Weinstein, P. 2003. Pathogen survival trajectories: An eco-environmental approach to the modeling of human campylobacteriosis ecology. Environmental Health Perspectives, 111: 19-28.

Slingenbergh, J. & Wint, W. 1997. Livestock Geography And Landuse. Study commissioned by the Food and Agriculture Organisation of the United Nations. In: International Conference on Livestock and the Environment (Ed. by Nell, A. J.). Wageningen: The World Bank, the Food and Agriculture Organisation of the United Nations and the International Agriculture Centre.

Smouse, P.E., Long, J.C. & Sokal, R.R. 1986. Multiple-Regression and Correlation Extensions of the Mantel Test of Matrix Correspondence. Systematic Zoology, 35: 627-632.

Sneath, P.H.A. & Sokal, R.R. 1973. Numerical taxonomy: the principles and practice of numerical classification. San Francisco: W.H.Freeman.

Snow, W.F. 1970. An observation on the effect of vegetation on the flight pattern of Anopheles melas and Aedes chamboni in The Gambia, West Africa. Transactions of the Royal Society of Tropical Medicine and Hygiene, 64: 477.

Snow, W.F. & Wilkes, T.J. 1970. The height and direction of non-host oriented flight, and the age, of Anopheles melas in relation to wind direction in The Gambia, West Africa. Transactions of the Royal Society of Tropical Medicine and Hygiene, 64: 477.

Sokal, R.R. & Rohlf, F.J. 1995. Biometry. New York: W. H. Freeman and Company.

Solomon, T. & Cardosa, M.J. 2000. Emerging arboviral encephalitis. British Journal of Medicine, 321: 1484-1485.

Solomon, T., Dung, N.M., Kneen, R., Gainsborough, M., Vaughn, D.W. & Khanh, V.T. 2000. Neurological aspects of tropical disease: Japanese encephalitis. J Neurol Neurosurg Psychiatry, 68: 405-415.

Solomon, T., Ni, H., Beasley, D.W.C., Ekkelenkamp, M., Cardosa, M.J. & Barrett, A.D.T. 2003. Origin and evolution of Japanese encephalitis virus in southeast Asia. Journal of Virology, 77: 3091-3098.

Spielman, A. 1999. The role of surveillance in interventions directed against Vector-Borne disease. Ecosystem Health, 5: 141-145.

Spielman, A. & D'Antonio, M. 2001. Mosquito : a natural history of our most persistent and deadly foe. London: Faber and Faber.

Stephens, D.W. 1993. Learning and behavioural ecology: incomplete information and environmental predictability. In: Insect Learning: Ecological and Evolutionary Perspectives (Ed. by Papaj, D. R. & Lewis, A. C.), pp. 195-218. New York: Chapman & Hall.

Strauss, R.E. 2003. Matlab v6 library. Texas.

Suwanee, A., Nualachawee, K., Murai, S., Honda, K., Noomhorm, A. & Lertlum, S. 1999. Remote Sensing for monitoring the distribution of mosquitoes related to Japanese encephalitis. ASIAN-PACIFIC Remote Sensing and GIS Journal, 11: 57-65.

Swartzman, G. 1979. Evaluation of ecological simulation models. In: Contemporary Quantitative Ecology and Related Ecometrics (Ed. by Patil, G. P. & Rosenzweig, M.), pp. 295-318. Fairland, MD: Co-operative Publishing House.

Takken, W. 1996. Synthesis and future challenges: the response of mosquitoes to host odours. Ciba Foundation symposium, 200: 302-312.

Tesh, R.B., da Rosa, A., Guzman, H., Araujo, T.P. & Xiao, S.Y. 2002. Immunization with heterologous flaviviruses protective against fatal West Nile encephalitis. Emerging Infectious Diseases, 8: 245-251.

Thomson, M.C. & Connor, S.J. 2000. Environmental information systems for the control of arthropod vectors of disease. Medical and Veterinary Entomology, 14: 227-244.

Thornton, P.K., Kruska, R.L., Henninger, N., Kristjanson, P.M., Reid, R.S., Atieno, F., Odero, A.N. & Ndegwa, T. 2002. Mapping Poverty and Livestock in the Developing World., pp. 124. Nairobi: ILRI (International Livestock Research Institute).

Trpis, M. & Hausermann, W. 1978. Genetics of house-entering behaviour in East African populations of Aedes aegypti (L.) (Diptera: Culicidae) and its relevance to speciation. Bulletin of Entomological Research, 68: 521-532.

Tsai, T.F. 1997. Factors in the changing epidemiology of Japanese encephalitis and West Nile fever. In: Factors in the emergence of arbovirus diseases (Ed. by Saluzzo, J. F. & Dodet, B.), pp. 179-189. Paris: Elselvier.

Tsai, T.F. 2000. New initiatives for the control of Japanese encephalitis by vaccination:; Minutes of a WHO/CVI meeting, Bangkok, Thailand, 13-15 October 1998. Vaccine, 18: 1-25.

Tsai, T.F., Chang, G.W. & Yu, X.Y. 1999. Japanese encephalitis vaccines. In: Vaccines - 3rd edition (Ed. by Plotkin, S. A. & Orenstein, W. A.), pp. 672-710. Philadelphia: WB Saunders, Inc.

Tully, T. & Quinn, W.G. 1985. Classical-conditioning and retention in normal and mutant Drosophila melanogaster. Journal of Comparative Physiology A -Sensory Neural and Behavioral Physiology, 157: 263-277.

Turell, M.J., Rossignol, P.A., Spielman, A., Rossi, C.A. & Bailey, C.L. 1984. Enhanced arboviral transmission by mosquitos that concurrently ingested microfilariae. Science, 225: 1039-1041.

Twiddy, S.S., Pybus, O.G. & Holmes, E.C. 2003. Comparative population dynamics of mosquito-borne flaviviruses. Infection, Genetics and Evolution, 3: 87-95.

Ulloa, A., Arredondo-Jimenez, J.I., Rodriguez, M.H. & Fernandez-Salas, I. 2002. Mark-recapture studies of host selection by Anopheles vestitipennis. Journal of the American Mosquito Control Association, 18: 32-35.

Van Den Hurk, A.F., Johansen, C.A., Zborowski, P., Paru, R., Foley, P.N., Beebe, N.W., Mackenzie, J.S. & Ritchie, S.A. 2003. Mosquito host-feeding patterns and implications for Japanese encephalitis virus transmission in northern Australia and Papua New Guinea. Medical & Veterinary Entomology, 17: 403-411.

Van Regenmortel, M.H. 1990. Virus species, a much overlooked but essential concept in virus classification. Intervirology, 31: 241-54.

Van Regenmortel, M.H. & Mahy, B.W.J. 2004. Emerging issues in virus taxonomy. Emerging Infectious Diseases, 10: 8-13.

Waddell, S. & Quinn, W.G. 2001. Flies, genes, and learning. Annual Review of Neuroscience, 24: 1283-1309.

White, G.B. 1974. Anopheles gambiae complex and disease transmission in Africa. Transactions of The Royal Society Of Tropical Medicine And Hygiene, 68: 278-301.

WHO. 1989. Geographical distribution of arthropod-borne diseases and their principal vectors. Geneva: World Health Organization.

WHO. 2001. Report of the fourth WHOPES Working Group meeting. Geneva: World Health Organization.

WHO. 2002. Fact sheet: dengue and dengue haemorrhagic fever. Geneva: World Health Organization.

Williams, B.G., Rogers, D.J., Staton, G., Ripley, B. & Booth, T. 1992. Statistical modelling of georeferenced data: maping tsetse distribution Zimbabwe using climate and vegetation data. In: Modelling Vector-borne and other Parasitic Diseases. Nairobi: International Laboratory for the Research into Animal Diseases (ILRAD).

Williams, B.K. 1983. Some observations on the use of discriminant-analysis in ecology. Ecology, 64: 1283-1291.

Williams, B.K. & Titus, K. 1988. Assessment of sampling stability in ecological applications of discriminant-analysis. Ecology, 69: 1275-1285.

Windsor, D.A. 1998. Most of the species on Earth are parasites. International Journal for Parasitology, 28: 1939-1941.

Wood, R.J. 1977. Genetic aspects of "quality" in mass reared insects. In: The use of ionizing radiation in agriculture (Ed. by Communities, C. o. t. E.), pp. 301-309. Wageningen, The Netherlands.

Woodall, J.P. 1979. Transmission of group C arboviruses (Bunyaviridae). In: Artic and tropical arboviruses, pp. 123-128.

Woolhouse, M.E., Dye, C., Etard, J.F., Smith, T., Charlwood, J.D., Garnett, G.P., Hagan, P., Hii, J.L., Ndhlovu, P.D., Quinnell, R.J., Watts, C.H., Chandiwana, S.K. & Anderson, R.M. 1997. Heterogeneities in the transmission of infectious agents: implications for the design of control programs. Proceedings of the National Academy of Sciences of the United States of America, 94: 338-42.

Work, T.H. 1971. On the Japanese B--West Nile virus complex or an arbovirus problem of six continents. American Journal of Tropical Medicine and Hygiene, 20: 169-186.

Wyatt, T.D. 2002. Pheromones and Animal Behaviour: Communication by Smell and Taste. Cambridge: Cambridge University Press.

Yoshizawa, C.N. & Le Marchand, L. 1988. Re: "Misinterpretation and misuse of the kappa statistic". Am J Epidemiol, 128: 1179-81.

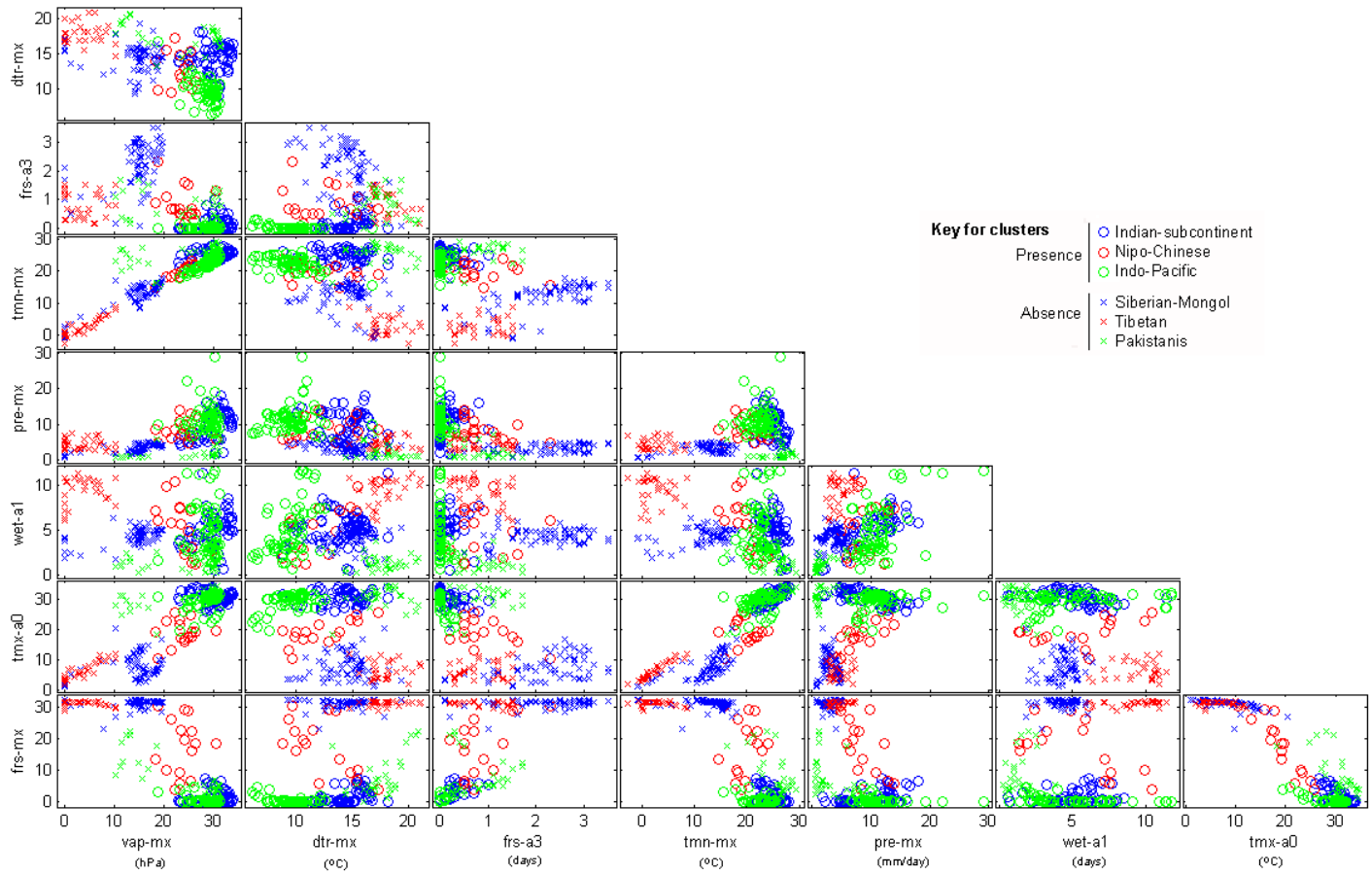
Zanotto, P.M., Gao, G.F., Gritsun, T., Marin, M.S., Jiang, W.R., Venugopal, K., Reid, H.W. & Gould, E.A. 1995. An arbovirus cline across the northern hemisphere. Virology, 210: 152-9.

Zanotto, P.M., Gould, E.A., Gao, G.F., Harvey, P.H. & Holmes, E.C. 1996. Population dynamics of flaviviruses revealed by molecular phylogenies. Proceedings of the National Academy of Sciences of the United States of America, 93: 548-53.

Zaslavsky, I. 1995. Analysis of association between categorical maps in multi-layer GIS. In: Gis/Lis 95 - Annual Conference and Exposition, pp. 1066-1074. Nashville, TN,: American Soc. Photogrammetry & Remote Sensing.

Zhou, G., Minakawa, N., Githeko, A.K. & Yan, G. 2004. Association between climate variability and malaria epidemics in the East African highlands. Proceedings of the National Academy of Sciences of the United States of America, 101: 2375-2380.

APPENDICES



Appendix 1. Bivariate scatterplots of the eight eco-climatic variables selected. Circles and crosses represent the records of presence and absence, respectively. Blue, red and green circles represent the presence clusters 'Indian-subcontinent', 'Nipo-Chinese' and 'Indo-Pacific'. Blue, red and green crosses represent the absence clusters 'Siberian-Mongol', 'Tibetan' and 'Pakistanis'. Main variable codes: dtr: diurnal temperature range; vap: vapour pressure; tmn: temperature; pre: precipitation; frs: number of frost days; wet: number of wet days. Secondary codes associated to each variable: a0: mean; mx: maximum, mn: minimum; a1: amplitude of the annual cycle; a3: amplitude of the triannual cycle.

Appendix 2. Observed values for each of the eco-climatic variables considered and each of the virus species. Key to codes: a0: mean; a1: amplitude of the annual cycle; a2: amplitude of the biannual cycle; a3: amplitude of the triannual cycle; mn: minimum; mx: maximum.

		Diurnal Temp. Range (oC)						Ground-frost Frequency (days)					
Species	**Cluster**	**a0**	**a1**	**a2**	**a3**	**mn**	**mx**	**a0**	**a1**	**a2**	**a3**	**mn**	**mx**
Alfuy	1	12.40	2.50	0.30	0.20	9.40	14.85	0.10	0.05	0.00	0.00	0.05	0.20
JE	1	11.29	3.33	0.99	0.31	7.27	14.41	0.67	0.63	0.33	0.16	0.01	1.65
JE	2	9.99	1.85	0.99	0.36	7.35	12.27	6.48	8.52	2.73	0.89	-0.09	17.54
JE	3	8.56	0.89	0.34	0.14	7.52	9.65	0.21	0.19	0.05	0.02	0.03	0.43
MVE	1	13.77	1.82	0.35	0.21	11.73	15.55	0.89	1.24	0.66	0.30	-0.02	3.04
MVE	2	10.25	1.62	0.23	0.21	8.51	12.01	0.31	0.21	0.01	0.01	0.09	0.50
Usutu	1	11.84	2.70	0.66	0.33	9.00	14.86	0.16	0.24	0.23	0.21	-0.10	0.84
Koutango	1	12.22	3.16	0.87	0.16	8.46	15.18	0.10	0.11	0.00	0.00	0.00	0.18
Kunjin	1	13.13	2.32	0.30	0.26	10.60	15.55	0.76	1.01	0.50	0.21	0.00	2.46
Kunjin	2	8.75	0.67	0.27	0.17	7.97	9.60	0.30	0.12	0.03	0.02	0.15	0.43
WN	1	12.45	2.32	0.77	0.26	9.56	14.83	1.51	2.27	1.10	0.44	-0.10	5.16
WN	2	9.40	2.58	0.61	0.27	6.51	11.82	7.77	10.58	2.98	0.86	-0.14	20.92
WN	3	10.24	1.98	0.49	0.25	8.18	12.55	0.12	0.13	0.06	0.05	0.00	0.36
Yaounde	1	12.48	2.98	0.80	0.48	9.03	15.68	0.20	0.35	0.33	0.33	-0.15	1.20
Cacipacore	1	8.80	1.10	0.30	0.00	7.40	9.70	0.00	0.00	0.00	0.00	0.00	0.00
Bagaza	1	14.17	4.00	0.98	0.33	9.08	17.55	0.08	0.08	0.00	0.00	0.00	0.15
Bagaza	2	10.32	2.48	0.64	0.32	7.96	13.54	0.20	0.34	0.30	0.28	-0.14	1.10
ITM	1	12.10	1.23	0.80	0.15	10.40	13.83	1.98	3.25	1.80	0.60	-0.20	7.30
Ntaya	1	8.75	1.05	0.48	0.23	7.55	10.26	0.22	0.20	0.13	0.12	0.00	0.64
Ntaya	2	11.72	1.62	0.46	0.30	10.16	13.70	0.26	0.28	0.14	0.08	0.00	0.74
Tembusu	1	9.55	2.10	0.38	0.25	7.65	11.88	0.08	0.05	0.00	0.00	0.03	0.13
Ilheus	1	9.46	0.84	0.39	0.15	8.46	10.64	0.40	0.46	0.13	0.05	0.03	0.97
Ilheus	2	9.11	1.16	0.51	0.16	7.91	10.68	0.10	0.03	0.02	0.00	0.05	0.14
Rocio	1	9.21	0.92	0.29	0.10	8.16	10.32	0.08	0.09	0.06	0.04	0.00	0.25
Kedougou	1	13.03	4.40	0.77	0.40	8.13	16.60	0.23	0.27	0.00	0.03	0.00	0.50
Spondweni	1	11.94	2.30	0.69	0.24	9.35	14.45	0.26	0.36	0.24	0.14	-0.04	0.91
Zika	1	10.01	1.57	0.54	0.21	8.37	12.06	0.19	0.21	0.09	0.08	0.00	0.54
Zika	2	13.59	4.26	0.98	0.33	8.47	17.20	0.11	0.12	0.00	0.01	0.00	0.20
St. Louis	1	10.49	1.21	0.42	0.21	9.12	11.87	0.26	0.41	0.28	0.16	-0.05	1.10
St. Louis	2	9.34	0.95	0.43	0.11	8.33	10.61	0.16	0.10	0.05	0.02	0.06	0.30
St. Louis	3	13.13	1.49	0.48	0.36	11.20	14.74	4.29	6.21	2.35	0.62	-0.20	13.00
Dengue 1	1	11.44	1.44	0.50	0.21	9.87	13.29	0.61	0.76	0.47	0.22	0.04	2.05
Dengue 1	2	10.49	1.08	0.52	0.20	9.19	11.97	0.08	0.04	0.01	0.00	0.03	0.14
Dengue 1	3	8.19	0.97	0.32	0.15	7.15	9.33	0.10	0.05	0.02	0.00	0.04	0.16
Aroa	1	9.80	1.10	0.40	0.00	8.50	11.10	0.00	0.00	0.00	0.00	0.00	0.00
Iguape	1	8.40	0.70	0.30	0.00	7.40	9.20	0.00	0.00	0.00	0.00	0.00	0.00
Bussuquara	1	10.16	1.00	0.56	0.11	8.97	11.56	0.13	0.02	0.01	0.00	0.08	0.17
Naranjal	1	12.90	0.80	0.30	0.30	11.90	14.00	0.10	0.10	0.10	0.10	0.00	0.40
Kokbera	1	14.13	1.20	0.40	0.17	12.60	15.40	1.17	1.87	1.27	0.70	-0.20	5.00
Kokbera	2	11.41	2.88	0.64	0.31	8.04	14.22	1.39	0.95	0.52	0.17	0.19	2.50
Stratford	1	10.05	0.83	0.15	0.28	9.00	11.10	1.00	1.48	0.73	0.28	-0.03	3.45

Appendix 2 (cont). Observed values for each of the eco-climatic variables considered and each of the virus species. Key to codes: a0: mean; a1: amplitude of the annual cycle; a2: amplitude of the biannual cycle; a3: amplitude of the triannual cycle; mn: minimum; mx: maximum.

		Precipitation (mm/day)						Minimum Temperaure (oC)					
Species	**Cluster**	**a0**	**a1**	**a2**	**a3**	**mn**	**mx**	**a0**	**a1**	**a2**	**a3**	**mn**	**mx**
Alfuy	1	2.75	4.20	1.95	0.80	-0.20	9.45	21.30	4.75	1.10	0.15	15.50	25.10
JE	1	3.50	4.43	1.76	0.52	0.05	9.74	19.47	6.43	1.54	0.30	11.78	25.06
JE	2	3.86	3.31	1.08	0.59	0.92	8.39	9.06	10.68	0.90	0.49	-1.98	19.71
JE	3	6.48	3.64	1.44	0.55	2.81	11.40	20.98	1.64	0.39	0.11	19.03	22.48
MVE	1	1.30	1.00	0.37	0.22	0.40	2.76	14.33	6.88	0.64	0.14	7.14	20.96
MVE	2	5.44	3.69	1.29	0.52	2.11	10.34	21.57	2.43	0.67	0.11	18.48	23.46
Usutu	1	3.26	2.80	1.43	0.51	0.27	7.16	18.81	2.56	0.93	0.24	15.50	21.41
Koutango	1	2.91	2.77	2.01	0.61	-0.18	7.37	20.40	2.78	0.79	0.47	16.76	22.96
Kunjin	1	2.16	2.60	1.07	0.45	0.21	6.06	17.60	5.98	1.00	0.14	10.79	22.90
Kunjin	2	8.00	3.23	0.92	0.48	4.65	11.87	21.77	0.68	0.22	0.10	20.87	22.32
WN	1	1.48	1.57	0.69	0.31	0.08	3.76	14.06	6.63	1.01	0.26	7.09	20.48
WN	2	1.91	0.66	0.35	0.15	1.05	2.86	5.93	9.51	0.63	0.42	-3.68	15.51
WN	3	3.77	3.16	1.56	0.56	0.34	8.10	19.89	1.90	0.68	0.26	17.48	21.86
Yaounde	1	2.95	2.40	1.70	0.63	0.18	6.88	20.55	2.73	1.23	0.45	16.80	23.48
Cacipacore	1	8.00	6.60	2.00	0.30	2.30	16.20	22.30	0.10	0.20	0.10	21.90	22.50
Bagaza	1	1.53	2.42	1.27	0.50	-0.15	5.63	20.98	4.07	1.63	0.45	15.65	25.18
Bagaza	2	4.32	3.06	1.54	0.68	0.70	8.64	19.92	1.08	0.76	0.24	18.22	21.40
ITM	1	1.40	1.85	0.53	0.15	-0.03	3.73	12.33	6.90	0.70	0.15	5.23	19.03
Ntaya	1	5.85	3.24	1.66	0.68	1.71	10.01	21.02	1.28	0.50	0.21	19.21	22.29
Ntaya	2	2.20	1.56	0.86	0.22	0.34	4.32	16.82	2.70	0.70	0.26	13.68	19.60
Tembusu	1	6.03	2.73	1.28	0.98	2.78	10.05	21.30	1.78	0.73	0.18	18.88	22.80
Ilheus	1	5.04	2.28	0.86	0.44	2.41	8.18	16.25	2.65	0.55	0.09	13.29	18.88
Ilheus	2	7.59	4.83	1.90	0.69	1.92	13.39	21.33	0.48	0.31	0.13	20.60	22.02
Rocio	1	4.77	2.55	0.70	0.38	2.33	8.10	17.76	2.78	0.58	0.06	14.67	20.55
Kedougou	1	2.97	4.57	1.97	0.43	-0.20	9.80	21.40	3.37	1.47	0.50	16.87	25.40
Spondweni	1	3.13	1.98	1.80	0.68	0.16	6.80	18.09	2.51	0.85	0.26	14.80	20.50
Zika	1	4.59	2.76	1.70	0.56	0.80	8.59	19.54	0.81	0.54	0.22	18.31	20.61
Zika	2	1.83	2.88	1.45	0.53	-0.18	6.54	20.81	4.21	1.36	0.49	15.44	24.99
St. Louis	1	3.74	2.51	0.99	0.54	1.13	6.88	18.40	4.59	0.57	0.21	13.54	22.66
St. Louis	2	6.60	3.74	1.48	0.58	2.34	11.37	18.88	1.53	0.44	0.08	17.05	20.50
St. Louis	3	2.32	1.21	0.40	0.26	1.05	3.73	10.26	8.24	0.56	0.28	1.73	18.41
Dengue 1	1	3.38	2.70	0.61	0.44	0.75	6.64	16.22	4.03	0.80	0.15	11.50	19.77
Dengue 1	2	3.81	3.04	1.16	0.73	0.53	7.49	20.47	1.63	0.44	0.16	18.49	21.94
Dengue 1	3	6.30	3.48	1.05	0.68	2.55	10.32	21.87	1.24	0.42	0.12	20.26	22.99
Aroa	1	2.50	0.80	0.70	0.40	0.80	3.80	22.00	1.30	0.60	0.10	20.30	23.00
Iguape	1	5.20	2.30	0.70	0.40	3.20	8.60	17.20	3.60	0.60	0.00	13.60	21.10
Bussuquara	1	6.76	4.16	2.08	0.53	1.74	12.41	20.39	0.87	0.32	0.10	19.18	21.40
Naranjal	1	2.80	3.20	1.40	0.40	0.60	7.50	18.70	1.50	0.20	0.20	17.20	20.10
Kokbera	1	1.07	0.33	0.10	0.10	0.67	1.60	13.10	7.93	0.47	0.10	4.87	20.73
Kokbera	2	4.34	5.44	2.30	0.54	0.37	12.24	19.25	5.60	1.22	0.13	12.54	23.81
Stratford	1	3.68	2.33	0.78	0.38	1.63	7.05	13.45	5.25	0.50	0.18	7.85	18.48

Appendix 2 (cont). Observed values for each of the eco-climatic variables considered and each of the virus species. Key to codes: a0: mean; a1: amplitude of the annual cycle; a2: amplitude of the biannual cycle; a3: amplitude of the triannual cycle; mn: minimum; mx: maximum.

		Mean Temperature (oC)						Maximum Temperature (oC)					
Species	**Cluster**	**a0**	**a1**	**a2**	**a3**	**mn**	**mx**	**a0**	**a1**	**a2**	**a3**	**mn**	**mx**
Alfuy	1	27.45	3.80	1.25	0.20	22.65	30.90	33.70	3.05	1.35	0.25	29.75	37.30
JE	1	25.08	5.36	1.89	0.32	18.27	30.20	30.75	4.61	2.30	0.40	24.65	36.16
JE	2	14.05	10.00	0.99	0.41	3.34	23.75	19.07	9.45	1.31	0.41	8.62	28.32
JE	3	25.23	1.43	0.45	0.07	23.45	26.71	29.53	1.47	0.56	0.11	27.72	31.26
MVE	1	21.18	7.19	0.77	0.12	13.58	27.98	28.08	7.55	0.91	0.15	20.02	35.22
MVE	2	26.65	1.81	0.75	0.11	24.15	28.32	31.82	1.44	0.83	0.18	29.83	33.57
Usutu	1	24.71	2.23	0.86	0.27	22.13	27.31	30.66	2.60	0.97	0.41	27.51	33.81
Koutango	1	26.48	1.99	1.00	0.51	23.68	28.81	32.63	2.11	1.28	0.51	29.84	35.53
Kunjin	1	24.13	5.44	1.08	0.16	17.83	29.09	30.73	5.01	1.19	0.23	24.90	35.62
Kunjin	2	26.07	0.57	0.30	0.03	25.20	26.63	30.48	0.63	0.43	0.10	29.55	31.38
WN	1	20.26	6.56	1.23	0.33	13.21	26.88	26.51	6.64	1.52	0.41	19.31	33.53
WN	2	10.60	10.76	0.77	0.39	-0.28	21.24	15.33	12.03	0.98	0.40	3.10	27.01
WN	3	24.97	1.87	0.70	0.23	22.70	27.05	30.13	2.18	0.81	0.26	27.39	32.63
Yaounde	1	26.78	2.30	1.25	0.45	23.73	29.78	33.05	2.58	1.50	0.63	29.58	36.73
Cacipacore	1	26.60	0.50	0.30	0.10	25.90	27.30	31.10	1.10	0.40	0.10	29.70	32.20
Bagaza	1	28.05	2.83	1.80	0.58	24.08	31.82	35.15	2.65	2.07	0.70	31.35	39.18
Bagaza	2	25.04	1.22	0.58	0.20	23.78	26.80	30.26	2.24	0.50	0.30	27.80	32.86
ITM	1	18.35	6.78	0.78	0.23	11.03	24.58	24.40	6.68	1.05	0.30	16.78	30.23
Ntaya	1	25.35	1.38	0.53	0.18	23.57	26.85	29.77	1.68	0.65	0.25	27.66	31.75
Ntaya	2	22.64	2.74	0.74	0.20	19.50	25.60	28.54	3.06	0.88	0.32	25.26	32.14
Tembusu	1	26.05	1.23	0.70	0.08	24.28	27.48	30.83	1.38	0.70	0.10	29.45	32.68
Ilheus	1	20.95	2.61	0.54	0.10	18.12	23.61	25.70	2.64	0.64	0.13	22.87	28.50
Ilheus	2	25.85	0.77	0.39	0.10	24.93	26.85	30.44	1.29	0.55	0.15	28.95	32.07
Rocio	1	22.34	2.59	0.58	0.08	19.53	25.07	26.97	2.47	0.62	0.09	24.38	29.65
Kedougou	1	27.90	2.73	1.63	0.33	25.13	31.80	34.47	3.60	1.87	0.27	29.93	39.23
Spondweni	1	24.01	2.13	0.86	0.36	21.31	26.44	30.01	2.38	0.98	0.43	27.03	32.83
Zika	1	24.53	0.97	0.49	0.21	23.22	25.81	29.58	1.56	0.59	0.27	27.61	31.49
Zika	2	27.57	2.89	1.54	0.53	23.79	31.12	34.39	2.71	1.82	0.58	30.69	38.28
St. Louis	1	23.62	4.20	0.57	0.14	19.04	27.48	28.89	3.92	0.66	0.15	24.45	32.46
St. Louis	2	23.53	1.50	0.45	0.10	21.87	25.27	28.22	1.74	0.55	0.12	26.41	30.29
St. Louis	3	16.81	8.42	0.67	0.29	7.96	25.14	23.39	8.71	0.82	0.38	14.12	32.06
Dengue 1	1	21.92	3.46	0.80	0.15	17.78	25.16	27.67	3.07	0.88	0.20	23.94	30.89
Dengue 1	2	25.69	1.49	0.46	0.15	23.86	27.19	30.96	1.52	0.62	0.21	29.03	32.74
Dengue 1	3	25.94	1.18	0.41	0.09	24.48	27.19	30.06	1.35	0.47	0.14	28.55	31.61
Aroa	1	26.90	0.80	0.50	0.10	25.70	27.70	31.80	0.30	0.50	0.10	31.10	32.60
Iguape	1	21.40	3.40	0.60	0.00	18.10	25.30	25.60	3.40	0.70	0.00	22.70	29.60
Bussuquara	1	25.43	1.04	0.38	0.11	24.20	26.62	30.54	1.46	0.53	0.14	28.74	32.20
Naranjal	1	25.10	1.20	0.20	0.20	23.80	26.20	31.60	1.00	0.40	0.30	30.10	32.50
Kokbera	1	20.13	8.33	0.60	0.03	11.33	28.03	27.27	8.70	0.80	0.07	17.93	35.33
Kokbera	2	24.93	4.40	1.45	0.15	19.11	28.45	30.65	3.39	1.69	0.24	25.75	34.15
Stratford	1	18.45	4.98	0.55	0.10	13.00	23.15	23.50	4.75	0.60	0.20	18.25	27.98

Appendix 2 (cont). Observed values for each of the eco-climatic variables considered and each of the virus species. Key to codes: a0: mean; a1: amplitude of the annual cycle; a2: amplitude of the biannual cycle; a3: amplitude of the triannual cycle; mn: minimum; mx: maximum.

		Vapour Pressure (hPA)						Wet Days Frequency (days)					
Species	**Cluster**	**a0**	**a1**	**a2**	**a3**	**mn**	**mx**	**a0**	**a1**	**a2**	**a3**	**mn**	**mx**
Alfuy	1	20.85	6.35	0.60	0.25	14.60	27.60	5.75	8.00	2.55	0.60	-0.05	16.25
JE	1	21.58	8.09	1.43	0.61	14.27	29.97	4.93	5.23	1.95	0.54	0.43	11.95
JE	2	12.95	8.92	1.98	0.64	5.42	23.66	12.89	5.34	2.17	0.66	7.58	20.13
JE	3	26.16	2.53	0.73	0.21	23.06	28.71	15.27	4.99	2.08	0.84	9.59	21.77
MVE	1	12.36	3.70	0.62	0.31	8.99	16.81	5.68	2.92	0.80	0.36	2.89	9.30
MVE	2	25.12	3.76	0.65	0.17	20.97	28.65	12.53	4.97	1.73	0.56	7.99	18.80
Usutu	1	21.01	4.80	1.09	0.43	15.61	25.47	9.30	5.69	2.83	0.87	2.46	16.51
Koutango	1	21.68	6.36	1.37	0.68	14.86	27.82	8.39	6.64	3.07	0.46	0.83	16.58
Kunjin	1	17.06	5.79	0.83	0.33	11.40	23.26	5.88	5.22	1.54	0.55	1.60	12.53
Kunjin	2	27.78	1.22	0.48	0.10	26.17	28.98	17.60	3.12	1.32	0.55	13.63	21.55
WN	1	14.16	6.15	1.16	0.41	8.69	21.17	5.94	4.61	1.42	0.53	1.41	11.32
WN	2	9.90	5.36	0.69	0.18	5.16	15.86	10.90	2.42	1.55	0.41	7.36	13.92
WN	3	23.45	3.15	1.09	0.47	19.42	26.41	9.95	6.16	2.87	0.83	2.75	17.51
Yaounde	1	20.65	5.63	1.10	0.60	15.18	26.60	8.35	6.13	3.50	1.00	1.25	16.75
Cacipacore	1	29.70	1.00	0.20	0.30	28.60	31.00	23.20	7.60	0.50	1.40	15.90	30.00
Bagaza	1	18.02	9.83	1.82	0.85	8.42	27.78	5.90	6.97	2.08	0.52	0.38	14.88
Bagaza	2	24.38	2.38	1.24	0.24	20.78	26.50	11.24	6.30	3.32	0.96	2.56	19.04
ITM	1	13.85	5.63	0.85	0.15	8.73	20.15	6.05	7.08	1.03	0.50	0.13	13.73
Ntaya	1	26.33	2.14	0.88	0.24	23.38	28.37	15.13	5.48	3.03	0.99	7.54	21.95
Ntaya	2	18.72	3.54	0.78	0.44	15.00	22.24	8.92	4.40	2.10	1.00	3.34	14.82
Tembusu	1	27.40	2.53	1.30	0.38	23.83	29.75	15.93	7.75	1.75	1.65	8.35	24.90
Ilheus	1	19.97	3.07	0.69	0.18	16.71	23.31	16.49	4.70	1.21	0.57	10.95	21.77
Ilheus	2	27.56	1.04	0.66	0.21	25.94	28.79	17.47	7.53	2.44	0.66	8.27	24.58
Rocio	1	21.49	3.32	0.65	0.19	18.08	25.22	18.93	5.35	0.52	0.72	13.38	24.46
Kedougou	1	19.77	9.33	1.30	1.40	9.80	27.83	7.33	9.90	2.73	0.37	-0.13	19.83
Spondweni	1	20.29	4.41	1.14	0.36	15.25	24.51	8.91	4.89	3.34	0.99	1.94	16.04
Zika	1	24.45	1.57	1.05	0.29	21.82	26.05	12.64	5.12	3.54	1.01	4.61	19.75
Zika	2	19.29	9.46	1.63	0.93	9.69	28.36	6.08	7.39	1.99	0.48	0.27	15.30
St. Louis	1	22.59	5.87	0.70	0.30	17.05	28.54	11.00	5.54	1.97	0.77	5.38	17.67
St. Louis	2	23.68	1.99	0.67	0.21	21.37	26.06	18.12	6.21	1.57	0.64	10.94	24.40
St. Louis	3	13.40	5.96	0.86	0.26	8.01	20.03	7.24	2.41	0.77	0.52	4.80	10.07
Dengue 1	1	19.76	4.55	0.64	0.27	14.98	24.06	12.98	6.58	1.11	0.86	6.17	19.83
Dengue 1	2	25.14	2.65	0.67	0.30	22.04	27.61	9.79	5.63	1.86	0.93	3.63	16.19
Dengue 1	3	27.22	2.22	0.72	0.21	24.51	29.32	16.21	5.88	1.66	0.85	9.53	22.56
Aroa	1	26.90	2.00	1.00	0.10	23.90	28.50	9.50	3.80	1.90	0.70	4.30	13.90
Iguape	1	20.20	4.10	1.00	0.10	16.40	25.30	19.20	4.80	0.70	0.30	13.90	23.80
Bussuquara	1	26.09	1.33	0.53	0.19	24.26	27.62	16.58	6.24	1.69	0.52	9.30	22.44
Naranjal	1	23.20	1.80	0.10	0.00	21.60	25.00	10.20	7.50	3.10	0.60	5.00	21.10
Kokbera	1	11.37	3.50	0.50	0.33	8.07	15.60	4.23	0.43	0.47	0.13	3.30	4.90
Kokbera	2	22.02	7.37	1.01	0.47	15.30	29.55	7.24	6.89	2.45	0.42	1.83	16.40
Stratford	1	15.68	4.63	0.40	0.20	11.10	20.50	10.53	2.35	0.58	0.45	7.88	13.18

Appendix 3. Matrix of genetic distances (Holmes & Twiddy, 2003) between the viruses of the Japanese encephalitis serocomplex as based on NS-5 gene sequences (990 bp). Distances were generated with the software PAUP 4.0b10 for Macintosh (PPC/Altivee), and reproduced from material kindly provided by the authors.

	Alfuy	Aroa	Bagaza	Bussuq.	Cacipac.	Den1	Den2	Den3	Den4	Iguape	Ilheus	Is/Turk M.	Jap.En.	Kedougou
Alfuy	0													
Aroa	0.48051	0												
Bagaza	0.52168	0.51903	0											
Bussuquara	0.49429	0.42446	0.5084	0										
Cacipacore	0.43481	0.47781	0.50574	0.49252	0									
Den1	0.519	0.52153	0.56476	0.56613	0.5376	0								
Den2	0.53851	0.51903	0.53185	0.54696	0.60298	0.35898	0							
Den3	0.53231	0.54166	0.53058	0.45972	0.55933	0.34015	0.33846	0						
Den4	0.52665	0.50883	0.55946	0.54251	0.56084	0.40504	0.38312	0.36437	0					
Iguape	0.49572	0.27455	0.47317	0.4055	0.46091	0.52894	0.55445	0.50664	0.53138	0				
Ilheus	0.45871	0.52785	0.39497	0.45306	0.47917	0.48316	0.47554	0.51536	0.55056	0.49343	0			
Isrl/Turk Men.	0.53384	0.52374	0.04902	0.50628	0.52904	0.56625	0.5165	0.49441	0.53765	0.48724	0.41022	0		
Japanese En.	0.34532	0.50075	0.49277	0.50003	0.4098	0.53926	0.56432	0.54301	0.54424	0.52854	0.45168	0.51052	0	
Kedougou	0.60837	0.60168	0.51593	0.48399	0.66234	0.54541	0.52753	0.58629	0.65947	0.5716	0.56763	0.52346	0.57813	0
Kokobera	0.52313	0.48181	0.50684	0.47518	0.48447	0.53762	0.60966	0.59724	0.60728	0.48335	0.51194	0.49835	0.52282	0.60626
Koutango	0.35757	0.51197	0.53579	0.51825	0.42963	0.55041	0.57877	0.57372	0.57858	0.49938	0.47294	0.52393	0.35715	0.70885
Kunjin	0.37801	0.54946	0.50156	0.47352	0.48118	0.58346	0.59667	0.53074	0.58355	0.46009	0.42354	0.48102	0.39105	0.64209
MurrayValleyEn.	0.33441	0.47232	0.44959	0.46058	0.40663	0.55983	0.58306	0.52815	0.60127	0.46325	0.45779	0.4475	0.33196	0.59179
Naranjal	0.47631	0.43843	0.51955	0.2792	0.50023	0.51691	0.54763	0.51751	0.58038	0.45953	0.44403	0.52185	0.5069	0.50805
Ntaya	0.47607	0.518	0.31828	0.49775	0.54078	0.51414	0.5782	0.57566	0.60797	0.51243	0.41282	0.31457	0.52272	0.5766
Rocio	0.49852	0.55168	0.45865	0.45736	0.51431	0.50667	0.51183	0.44137	0.53191	0.50878	0.37214	0.44857	0.45998	0.54591
St. Louis En.	0.43823	0.46931	0.42399	0.4254	0.44696	0.46175	0.4773	0.45668	0.53757	0.39738	0.38643	0.42781	0.41569	0.51293
Spondweni	0.49875	0.48965	0.50928	0.46871	0.55266	0.5418	0.53511	0.53903	0.56846	0.51804	0.46052	0.49939	0.46713	0.51823
Stratford	0.55764	0.53184	0.58426	0.54864	0.50405	0.58697	0.66181	0.64082	0.66925	0.50607	0.57895	0.60239	0.5548	0.65351
Tembusu	0.53669	0.54665	0.34046	0.50768	0.55457	0.59572	0.59842	0.58934	0.63866	0.499	0.45143	0.3405	0.5096	0.61577
Usutu	0.35536	0.48259	0.50876	0.51961	0.40886	0.54155	0.53284	0.51309	0.59061	0.50341	0.47016	0.50892	0.35081	0.60191
West Nile	0.32777	0.50514	0.52339	0.52734	0.45188	0.52647	0.57049	0.55313	0.64759	0.48781	0.45938	0.51587	0.36438	0.63898
Yaounde	0.39371	0.48842	0.45697	0.47058	0.41356	0.53367	0.49283	0.49651	0.53146	0.4553	0.44541	0.46198	0.36543	0.63285
Zika	0.48541	0.54848	0.49504	0.50858	0.55363	0.47552	0.49398	0.50976	0.53033	0.50979	0.43132	0.48625	0.513	0.52367

Appendix 3 (cont.). Matrix of genetic distances (Holmes & Twiddy, 2003) between the viruses of the Japanese encephalitis serocomplex as based on NS-5 gene sequences (990 bp). Distances were generated with the software PAUP 4.0b10 for Macintosh (PPC/Altivee), and reproduced from material kindly provided by the authors.

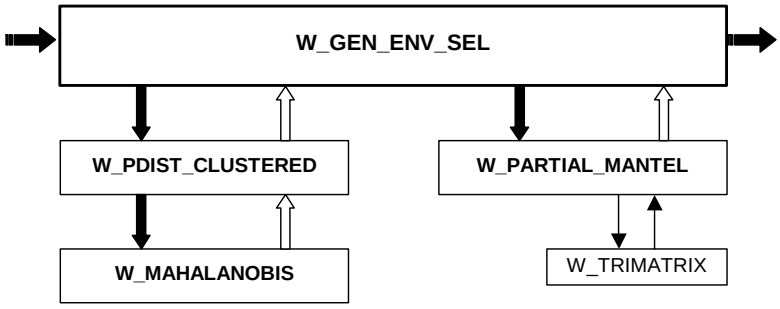
	Kokobera	Koutango	Kunjin	Murray V.	Naranjal	Ntaya	Rocio	St.Louis	Spondw.	Stratford	Tembusu	Usutu	W. Nile	Yaounde	Zika
Alfuy															
Aroa															
Bagaza															
Bussuquara															
Cacipacore															
Den1															
Den2															
Den3															
Den4															
Iguape															
Ilheus															
Isrl/Turk Men.															
Japanese En.															
Kedougou															
Kokobera	0														
Koutango	0.5512	0													
Kunjin	0.56346	0.32599	0												
Murr.ValleyEn.	0.53311	0.39419	0.33561	0											
Naranjal	0.53024	0.50205	0.52371	0.50226	0										
Ntaya	0.50028	0.48511	0.54387	0.50252	0.48775	0									
Rocio	0.54178	0.49082	0.4939	0.47245	0.46389	0.41755	0								
St. Louis En.	0.51712	0.41505	0.41122	0.41021	0.40692	0.40914	0.38939	0							
Spondweni	0.48899	0.48481	0.4722	0.51211	0.50526	0.44439	0.50501	0.4126	0						
Stratford	0.39052	0.53341	0.51651	0.50763	0.5794	0.57167	0.6	0.56093	0.58054	0					
Tembusu	0.50019	0.59087	0.55316	0.52233	0.50856	0.37569	0.482	0.44057	0.53926	0.60657	0				
Usutu	0.55204	0.3514	0.34177	0.31539	0.50418	0.50306	0.49415	0.44706	0.48704	0.58064	0.50359	0			
West Nile	0.50519	0.31457	0.28969	0.35636	0.56998	0.54356	0.49596	0.41805	0.49709	0.54147	0.54233	0.40859	0		
Yaounde	0.47684	0.34574	0.35748	0.34446	0.51284	0.48578	0.4713	0.40628	0.50787	0.46672	0.51204	0.41316	0.3246	0	
Zika	0.46895	0.48581	0.48873	0.48448	0.50032	0.49465	0.43863	0.40852	0.39278	0.54051	0.5706	0.49835	0.47767	0.50667	0

Appendix 4

The functions listed below and described in the following pages are necessary to select the environmental variables that maximize the correlation with the matrix of phylogenetic distances between the species, as implemented in Chapter 6 of this thesis.

w_gen_env_sel:	Stepwise selection of variables as based on phylogenetic distances and controlling for spatial autocorrelation (master program)
w_pdist_clustered:	Pairwise distance between species, to be used when they data have been partitioned into clusters
w_mahalanobis:	Mahalanobis distance among two matrices of objects x descriptors
w_partial_mantel:	Partial Mantel test for 3 symmetric distance matrices using the method of permutation of raw data
w_trimatrix:	Extracts in a vector the lower triangular submatrix of a square symmetric matrix

The diagram below shows how these functions are activated.



In the following pages the programming codes (written in Matlab, Mathworks Associates) associated to each function are described. For greater clarity, syntax highlighting colours follow Matlab's default, with comments appearing in green, commands of flow control in blue, closed strings in red and the remaining commands in black.

```
% W_GEN_ENV_SEL
%           Stepwise selection of variables as based on phylogeny and controlling for spatial autocorrelation
%
%           The stepwise process selects those variables which maximizes the
%           correlation with the matrix of genetic distances between the species.
%           It also offers the possibility of controling for differences in the geographic
%           distances between the species through partial correlative methods.
%
%           If the data for each species is partioned into clusters, the distances between each pair of
%           is obtained by calculating the minimum distance between their respective clusters
%
%
%
%  -----Usage:-----
%
%          [ top_variables, rAB_C, prAB_C, corr_var, vectors_pdist ] =
%                      w_gen_env_sel(gene,envi,spe,cl_n,geog,metric,ran,n_var,criter,excluvar)
%
%  -----Input:------
%
%          gene =  [p-by-p] genetic distances matrix
%          envi =  [m-by-p] environmental matrix
%          geog =  [p-by-p] geographic distances matrix
%          spe =  [m-by-1] vector of species code
%          cl_n =  m-by-1 vector of species cluster
%          metric =  can be any of the following character strings that identify ways to compute the distance:
%                      'Mahal' : Mahalanobis distance (default)
%                      'Euclid' : Euclidean distance - not recomended if variables were not standardized first!
%                      'SEuclid' : Euclidean distance on standardized values
%          ran =  number of randomizations (default = 10000)
%          n_var =  number of variables that the model should select (default = 10)
%          criter =  when species data were clustered, the distance to be considered between each species pair:
%                      mi -> (minimum) the shortest distance (closest clusters) between both species (default)
%                      me -> (mean) the mean distance among all possible distances between both species
%          excluvar = instruction allowing the same variable to be included in the model more than once
%                      'yes' -> do allow
%                      'no'  -> do NOT allow (default)
%  -------Output:-----
%
%          top_variables = [n_var-by-1] vector of selected variables
%          rAB_C = [n_var-by-1] vector of partial correlations
%          prAB_C = [n_var-by-1] vector of mantel test performed with rn randomizations
%          corr_var = [n_var-by-n_var] matrix of correlations among selected variables
%          vectors_pdist = [(((t-1)^2)/2)-by-n_var] matrix containing information about distances when
%                       each variable is added.
%
%  --- Requirement:---
%
%           w_mahalanobis
%           w_partial_mantel
%           w_pdist_clust
%           w_trimatrix
%
%           If the simple mantel test is to be used, the Strauss toolbox is needed
%           (free download at http://www.biol.ttu.edu/Strauss/Matlab/matlab.htm)
%
% ------------------------------------------------------------------------------
%           Wladimir J. Alonso
%           OTRG, Zoology Department, Oxford University (UK) & CAPES (Brazil)
%           2003
%           please report faults to : wladimir.alonso@zoo.ox.ac.uk
% ------------------------------------------------------------------------------
```

```
function [ top_variables, rAB_C, prAB_C, corr_var,vectors_pdist ] = ...
                                    w_gen_env_sel(gene, envi, spe, cl_n, geog, metric, ran, n_var,criter,excluvar)

if (nargin < 10) excluvar = 'No'; end
if (nargin < 9) criter = 'mi'; end
if (nargin < 8) n_var = 10; end
if (nargin < 7) ran = 10000; end
if (nargin < 6) metric = 'Mahal'; end
if (nargin < 5) geog = [ ]; end
if (nargin < 4) cl = ones(size(envi,1)); end

 [q,w,spe] = unique(spe);
 n_species_being_anal = max(spe);
 n_max_cl = max(cl_n); % maximum number of clusters
 [n_samples_ALL_spec n_variables] = size(envi);
 vectors_pdist = [ ];


if metric(1:3) == 'SEu' % 'SEuclid' - Standardization envi:
        mean_envi = mean(envi,1)';
        std_envi = std(envi,1,1)'; % stand. deviation producing the 2o moment of the sample about its mean.
        mean_envi_matrix = (mean_envi * (ones(1,n_samples_ALL_spec)) )';
        std_envi_matrix = (std_envi * (ones(1,n_samples_ALL_spec)) )';
        envi = (envi - mean_envi_matrix)./(std_envi_matrix);
end % if metric == 'SEuclid'

hhh = waitbar(0,'Running stepwise selection...');
top_variables = [ ];
for j = 1:n_var
   corr_list = [ ];
   for i_curr_var = 1:n_variables

      if  (excluvar(2)=='e') | (ismember(i_curr_var,top_variables)<1)
         this_var_dist = squareform(w_pdist_clust(envi(:,[top_variables;i_curr_var]),spe,cl_n,metric,criter));
          % important part - it generates the environmental distance matrix

         if geog ~= [ ]
         rAB_C_prov = w_partial_mantel(this_var_dist, gene, geog);
           % partial correlation
         else
         rAB_C_prov = corr(w_trimatrix(this_var_dist),w_trimatrix(gene));
          % simple correlation
         end


        corr_list = [corr_list; rAB_C_prov];
       else
        corr_list = [corr_list; 0 ]; % 0 is a dummy value, used only to export the right indexes of the variables
      end
   end
   i_max_corr = find(corr_list == max(corr_list));
   top_variables = [top_variables; i_max_corr];

         vector_pdist = w_pdist_clust(envi(:,top_variables),spe,cl_n,metric,criter);
         envi_dist = squareform(vector_pdist);
```

```
        if geog ~= []
          [rAB_C(j,1),prAB_C(j,1)] = w_partial_mantel(envi_dist,gene,geog,ran);
          % partial correlation and partial mantel test
        else
          [rAB_C(j,1),Z,prAB_C(j,1)] = mantel(envi_dist,gene,ran);
          % simple correlation and simple mantel test
        end

    vectors_pdist = [vectors_pdist,vector_pdist'];

    waitbar(j/n_var);
end

if nargout > 3
    corr_var = corr(envi(:,top_variables(1:n_var)));
end

close(hhh);
```

```
%  W_PDIST_CLUST
%           Pairwise distance between species, in cases where data was
%           previously partitioned into clusters.
%
%           Computes the Euclidean or Mahalanobis distance among each
%           pair of species in "env". The novelty of this function over
%           "pdist" (The Matworks Copyright) is that the distances can be
%           calculated even if they had been previously clustered (vector
%           "cl"). In this case, the distances between each two species are
%           going to be calculated from each cluster of the "1st species"
%           to each cluster of the "2nd species". The criterion for
%           arriving to a final distance for each pair of species can be the
%           minimum value (short distance) or mean value.
%
%
%           "The output Y is arranged in the order of ((1,2),(1,3),..., (1,M),
%           (2,3),...(2,M),.....(M-1,M)).  i.e. the upper right triangle of
%           the M by M square matrix. To get the distance between species
%           i and species j, either use the formula Y((i-1)*(M-i/2)+j-i)
%           or use the helper function Z = SQUAREFORM(Y), which will return a
%           M by M symmetric square matrix, with Z(i,j) equaling the distance
%           between species i and species j "
%                 (Explanation for pdist in MatLab, The Matworks Copyright).
%
%
%
%  -----Usage:-----
%
%       [Y] = w_pdist_clust(env,sp,cl,metric,criterion)
%
%  -----Input:------
%
%       env = [m-by-p] matrix of m samples with p variables
%       sp = [m-by-1] vector of species code
%       cl = [m-by-1] vector of species cluster
%       metric = computes the distance between objects in the data matrix  env,
%                   using  Mahalanobis as default, or Euclidean if there is something
%                   else than left in blank or written 'mahal'
%       criterion = when species had been clustered, the distance that is going to be
%                   considered between each pair of species:
%                   mi -> (minimum)the shortest distance (closest clusters)(default)
%                   me -> (mean) the mean distance among all possible distances
%
%  -------Output:-----
%
%        Y = [(((t-1)^2)/2)-by-1] vector containing the distance information.
%
%  --- Requirement:---
%
%        w_mahalanobis
%
%
%
% ------------------------------------------------------------------------------
%           Wladimir J. Alonso
%           OTRG, Zoology Department, Oxford University (UK) & CAPES (Brazil)
%           2003
%           please report faults to : wladimir.alonso@zoo.ox.ac.uk
% ------------------------------------------------------------------------------
```

```
function Y = w_pdist_clust(env,sp,cl,metric,criterion)
if (nargin < 4)  metric = 'Mahal'; end
if (nargin < 5)  criterion = 'mi'; end

[q,w,sp_] = unique(sp);
vt_n_cl = [ ];



 n_spp = max(sp_);
   for g=1:n_spp
      vt_n_cl = [vt_n_cl; max(cl(sp_==g))]; % vector with number of clusters for each species
   end
max_n_cl = max(vt_n_cl);



vt_dist_partial =[ ];
x2nd_obs =2;


for i_x1st_spp =1:(n_spp-1)
   % first species = for 1 to the number of species
   % (minus one, because the last "1st species" will
   % have already been properly compared with all the others
   % when its turn arrive (so we skip it)

 value_to_2nd_spp =  NaN * ones(  (n_spp-x2nd_obs)+1,max_n_cl);
   % Creates a provisory NaN matrix for
   % storing distances from "1st species"

   for cl_fst_spp=1:vt_n_cl(i_x1st_spp)
      % loop for including all the clusters of the first species

      counting = 0;
      MM1= env((sp_==i_x1st_spp)&(cl==cl_fst_spp),:);
      % sub-matrix refering to current cluster of the current "1st species"

      if metric(2) ~= 'a'
         mean_MM1 = mean(MM1,1);
         % means of the current cluster of the "1st species", needed for Euclidean calculation
      end

      for i_x2nd_spp= x2nd_obs:n_spp
         for cl_2nd_spp=1:vt_n_cl(i_x2nd_spp)
          MM2= env((sp_==i_x2nd_spp)&(cl==cl_2nd_spp),:);
          % sub-matrix refering to current cluster of the current "2nd species"

               if metric(2) ~= 'a'
                 mean_MM2 = mean(MM2,1);
                 % means of the current cluster of the "2nd species", for  Euclidean calculation

                 D12(cl_2nd_spp) = sqrt(sum((mean_MM1-mean_MM2).^2 , 2));
                 % Euclidean distance
               else
                 D12(cl_2nd_spp) = w_mahalanobis(MM1,MM2);
                 % Mahalanobis distance
               end
         end
       counting = counting + 1;
```

```
      if criterion(2) == 'e'
        value_to_2nd_spp(counting,cl_fst_spp) = mean(D12);
        % gets the mean distance of the current cluster of the "1st species"
        % to all clusters of the "2nd species"
      elseif criterion(2) == 'i'
        value_to_2nd_spp(counting,cl_fst_spp) = min(D12);
        % gets the minimum distance of the current cluster of the "1st species"
        % to all clusters of the "2nd species"
      end

      D12 = [];
    end
  end

  x2nd_obs = x2nd_obs + 1;

  vt_dist_partial = [ vt_dist_partial;value_to_2nd_spp ];
  % each row contains several distances for one pairwise computation,
  % from where one is going to be calculated (just below)

end

  if criterion(2) == 'e' % (e for 'mean')
    Y = mean(vt_dist_partial, 2);
    % gets the mean value of the mean for all clusters of the "1st species"
  elseif criterion(2) == 'i' % (i for 'minimum')
    Y = min(vt_dist_partial,[],2);
    % gets the minimum value of the minimums for all clusters of the "1st species"
  end


Y = Y';
```

```
%  W_PARTIAL_MANTEL
%           Partial Mantel test for 3 symmetric distance matrices using the method
%           of permutation of raw data
%
%          " The partial Mantel statistic (sensu Smouse et al., 1986), rM(AB.C), estimating
%            the correlation between matrices A and B while controlling for the effect of C,
%            is computed in the same way as a partial correlation coefficient " (Legendre, 2000)
%
%
%          " Results obtained by Anderson and Legendre (1999) in the multiple regression
%            context indicated that in most instances, permutation of the raw
%            data had correct type I error and good power. When the covariable contained
%            an extreme outlier, however, permutation of raw data resulted in
%            unstable (often inflated) type I error. The simulations reported in this paper
%            will examine the behavior of permutation method 1 in the context of
%            the partial Mantel test. " (Legendre, 2000)
%
%
%  -----Usage:-----
%
%  [ rAB_C, prAB_C ] = w_partial_mantel(A,B,C,n_per)
%
%  -----Input:-----
%
%       A,B,C = [p-by-p] square symmetric matrices with zeros on the diagonal.
%       n_per = number of permutations for the randomization test (default = 10000)
%
%  -----Output:-----
%
%       rAB_C = correlation between matrix cells
%      prAB_C = right-tailed significance level (p-value)
%
%  ----Disclaimer:----
%
%      Explanations of the steps taken along the function where extracted (literally or not) from "model 1"
%      described in Legendre (2000; method proposed by Smouse et al. ,1986). Of course, the authors
%      of this article are not responsible for any possible error that can arise from the
%      interpretation and transcription to their guidelines into a programming language.
%
%  --- Requirement:---
%
%      w_trimatrix
%
%  ----Bibliography:----
%
%     Legendre, P. 2000. Comparison of permutation methods for the partial correlation
%            and partial Mantel tests. Journal of Statistical Computation and Simulation
%            67: 37-73.
%            http://www.fas.umontreal.ca/BIOL/legendre/reprints/Partial_Mantel_paper.pdf
%
%     Legendre, P. & Legendre, L. 1998. Numerical ecology. Amsterdam, Oxford: Elsevier.
%
%     Smouse, P. E., Long, J. C. & Sokal, R. R. 1986. Multiple-Regression and Correlation
%            Extensions of the Mantel Test of Matrix Correspondence. Systematic Zoology,
%            35, 627-632.
%
% ------------------------------------------------------------------------------
%            Wladimir J. Alonso
%            OTRG, Zoology Department, Oxford University (UK) & CAPES (Brazil)
%            2003
%            please report faults to : wladimir.alonso@zoo.ox.ac.uk
% ------------------------------------------------------------------------------
```

```
function [rAB_C, prAB_C] = w_partial_mantel(A,B,C,n_per)

t = 1;
if (nargout < 2) t = 2; end;

%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
% 1. Compute the Mantel correlations rM(AB), rM(AC) and rM(BC).

v_A = w_trimatrix(A);
v_B = w_trimatrix(B);
v_C = w_trimatrix(C);

rAB = corr2(v_A,v_B);
rAC = corr2(v_A,v_C);
rBC = corr2(v_B,v_C);

% Calculate the reference value of the test statistic, rM(AB.C):
rAB_C = (rAB - (rAC * rBC)) / ( (sqrt(1 - (rAC^2))) * (sqrt(1 - (rBC^2))) ); % eq. 2

%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%

if t ~= 2;
shuffled_A = A;
r_shuffledA_B__C = [];
values_r_shuffledA_B__C = [];
[n_rows n_col] = size(A);

   if (nargin < 4) n_per = 10000; end;

for i=1:(n_per - 1)

% 2. Permute A at random using matrix permutation to obtain A*.

   h = randperm(n_rows);                    % get permuted indices
   k = A(h,:);                              % permute rows
   shuffled_A= k(:,h);                      % permute cols

%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
% 3. Compute rM(A*B) and rM(A*C).

   v_shuffled_A = w_trimatrix(shuffled_A);

   r_shuffledA_B = corr2(v_shuffled_A,v_B);
   r_shuffledA_C = corr2(v_shuffled_A,v_C);



% Using the value rM(BC) calculated in step 1, compute rM(A*B.C), using eq. 2,
% to obtain a value of the partial correlation statistic under permutation.

   r_shuffledA_B__C = (r_shuffledA_B - (r_shuffledA_C * rBC)) / ( (sqrt(1 - (r_shuffledA_C^2))) * (sqrt(1 - (rBC^2))) );
   values_r_shuffledA_B__C = [values_r_shuffledA_B__C, r_shuffledA_B__C];

end
% 4. Repeat steps 2 and 3 a large number of times to obtain the distribution
% of under permutation. Add the reference value rM(AB.C) to the
% distribution.

   values_r_shuffledA_B__C = [values_r_shuffledA_B__C, rAB_C];
   sorted_values_r_shuffledA_B__C = sort(values_r_shuffledA_B__C);

%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
```

```
% 5. For a one-tailed test involving the upper tail, calculate the probability
% as the proportion of values greater than or equal to rM. In the lower
% tail, the probability is the proportion of values smaller than or
% equal to rM.


   for  i=1:n_per
      if sorted_values_r_shuffledA_B__C(i) == rAB_C
         statistics_rAB_C = i;
      end
   end

   prAB_C = (1 - statistics_rAB_C/n_per);

end % if test ~= 2;
```

```
% W_MAHALANOBIS
%           Mahalanobis distance among two matrices of objects x descriptors
%
%           For two groups of sites W1 and W2, containing n1 and n2 sites, respectively,
%           and described by the same p variables, the square of the generalized distance
%           is given by the following matrix formula:
%           D^2 (W1, W2) = d12 * inv(V) * d'12
%
%           V = the pooled within-group dispersion matrix of the two groups of sites,
%           estimated from the matrices of sums of squares and cross products between
%           group-centred descriptors for each of the two groups, added up
%           term by term and divided by (n1 + n2 -2), as in discriminant
%           analysis and multivariate analysis of variance (Legendre & Legendre 1998:280)
%
%  -----Usage:-----
%
%      [D2] =   w_mahalanobis(W1,W2,method)
%
%  -----Input:-----
%
%      W1 = [m-by-p] matrix with the first group of observations
%      W2 = [m-by-p] matrix with the second group of observations
%      method = covariances matrices
%             - pooled
%             - separate (default)
%
%  -----Output:-----
%
%     D2 = squared of the generalized distance
%
%  ----Bibliography:----
%
%
%      Rogers, D. J. 2000. Satellites, space, time and the African trypanosomiases. In:
%           Remote Sensing and Geographical Information Systems in Epidemiology, pp.
%           129-171. London: Academic Press.
%
% ------------------------------------------------------------------------------
%           Wladimir J. Alonso (wladimir.alonso@zoo.ox.ac.uk)
%           Ben McCormick (benjamin.mccormick@zoo.ox.ac.uk)
%           OTRG, Zoology Department, Oxford University (UK) & CAPES (Brazil)
%           2003
% ------------------------------------------------------------------------------

function [D2] = w_mahalanobis(W1,W2,method)

if nargin<3
   method='separate';
end

     S1=cov(W1);                              % -> Dispersion matrix for group 1
     S2=cov(W2);                              % -> Dispersion matrix for group 2
     d12 = mean(W1,1) - mean(W2,1);                    % -> vector (lenght p) of the diferences between
                                          %    the means of p variables
```

```
switch lower(method(1:2)),
  case {'po'}


    [n1] = size(W1,1);                          % -> n1 = number of sites of group 1
    [n2] = size(W2,1);                          % -> n2 = number of sites of group 2
    V = (1/(n1 + (n2 - 2))) * ( ((n1 - 1) * S1) + ((n2 - 1) * S2));% -> pooled within-group dispersion matrix
    D2 = ((d12) * (inv(V)) * (d12'));               % -> Mahalanobis distance (actually it is D^2)

  case {'se'}


    D12 = ((d12) * (inv(S2)) * (d12'));             % -> Mahalanobis distance (actually it is D^2)
    D21 = ((d12) * (inv(S1)) * (d12'));             % -> Mahalanobis distance (actually it is D^2)

    D2 = min(D12,D21);

end
```

```
% W_TRIMATRIX
%       Extracts in a vector the lower triangular submatrix of a square symmetric matrix
%
%
%
%   -----Usage:-----
%
%   v = w_trimatrix(x)
%
%   -----Input:-----
%
%      x = [p-by-p] square symmetric matrix
%
%   -----Output:-----
%
%      v = [(((p-1)^2)/2)-by-1] vector containing the data of the lower triangular submatrix
%
%
% ------------------------------------------------------------------------------
%               Based on a suggestion by Ivan Araujo
% ------------------------------------------------------------------------------

function v = w_trimatrix(x)

 [n,p] = size(x);

 if (n~=p)
   error('Matrix must be square');
 end;

 v = [ ];

   for i = 1:n-1
            v = [v; x(i+1:n,i)];
   end
```

Appendix 5

Academic work published from 2001-2003

ALONSO, W. J., WYATT, T.D., KELLY, D.W. 2003. Are vectors able to learn about their hosts? A case study with *Aedes aegypti* mosquitoes. Memórias do Instituto Oswaldo Cruz, 98: 665 – 672.

ALONSO, W. J., SCHUCK-PAIM, C. 2002. Sex ratio conflicts, kin selection and the evolution of altruism. Proceedings of the National Academy of Sciences of the USA (PNAS), 90: 6843 – 6847.

ALONSO, W.J. 2002. Epidemiology and ethology in the study of the infectious agents: a desirable mutual contamination (in Portuguese). Proceedings of the XX Meeting of Ethology, 20: 221 – 229 (Brazil, Natal).

SCHUCK-PAIM, C., ALONSO, W.J. 2001. Behavioural mechanisms underlying the spatial distribution of *Nephilengys cruentata* females (Araneae, Tetragnathidae). Animal Behaviour 62: 1007 – 1012

Correspondence

FERNÁNDEZ-JURICIC, E., W.J. ALONSO & C. SCHUCK-PAIM. 2002. Peer-review to select academic job applicants. Nature, 420: 16.

ALONSO, W.J., FERNÁNDEZ-JURICIC, E. 2002. Regional network raises profile of local journals. Nature, 415: 471 - 472.

Previous work published in peer-reviewed journals

ALONSO, W.J., LUCENA, T., FRACASSO, C., VELTHUIS, H., IMPERATRIZ-FONSECA, V. L. 1998. Do *Melipona bicolor* (Apidae, Meliponinae) workers distinguish relatedness among different physogastric queens? Apidologie 29: 503 – 512.

ALONSO, W.J. 1998. The role of kin selection theory on the explanation of biological altruism: a critical review. Journal of Comparative Biology 3: 1 – 14.