Combining hazard and exposure to model the spatial distribution of two zoonoses, based on human case records

Caroline Zeimes

Thèse présentée en vue de l'obtention du grade de
Docteur en Sciences

Université Catholique de Louvain
Louvain-la-Neuve, April 2015

Composition du jury

Sophie O. Vanwambeke *Promoteur*

Professeur, Earth and Life Institute, Université Catholique de Louvain

Marie-Laurence De Keersmaecker *Président*

Professeur, Earth and Life Institute, Université Catholique de Louvain

Hugues Goosse

Professeur, Earth and Life Institute, Université Catholique de Louvain

Marius Gilbert

Docteur, Biological Control and Spatial Ecology, Université Libre de Bruxelles

Herwig Leirs

Professeur, Evolutionary Ecology Group, Universiteit Antwerpen

Acknowledgement

This thesis was funded by EU grant FP7-261504 EDENext and is catalogued by the EDENext Steering Committee as EDENext327 (http://www.edenext.eu).The contents of this publication are the sole responsibility of the authors and don't necessarily reflect the views of the European Commission.

Mes remerciements vont en tout premier lieu à ma promotrice, Sophie Vanwambeke. Pendant toute la durée de ma thèse, elle m'a guidée, soutenue, encouragée, cadrée et rassurée. Elle a toujours été très disponible. Ouverte à toutes mes suggestions, elle m'a proposé de nouveaux horizons. Elle m'a fait découvrir et aimer la recherche. Elle m'a permis de développer ma créativité et ma rigueur scientifique. En plus d'être une excellente scientifique, chercheuse et professeur, elle possède de belles qualités humaines qui en font un mentor exceptionnel.

I would like to thank the members of my jury, Prof. Marie-Laurence De Keersmaecker, Prof. Herwig Leirs, Dr. Marius Gilbert and Prof. Hugues Goosse.

Thanks to EDENext, I discovered a scientific community of passionate and exciting people. It is not possible to cite all of them here but my first thought goes to my friends: Dr. Nienke Hartemink and Daniela Cianci.

Many thanks also to all my collaborators which made me realize that Europe is a large family such as Gert Olsson, Marika Hjertkvist or Chantal Reusken.

A special thank goes to my colleagues at the Medical Geography lab: Eva, Sen, Valérie, Jasper, Mathilde and Pierre.

Finalement, j'aimerais remercier mon conjoint, ma famille, mes amis et tous mes collègues de l'institut qui m'ont donné l'envie d'aller au bureau tous les matins pour partager, rigoler, râler sur les problèmes techniques et tout simplement, vivre ensemble.

Contents

Composition du jury i

Acknowledgement iii

Contents v

List of abbreviations ix

Summary xi

Chapter 1 Introduction 1

1.1. Zoonoses 2

1.1.1. Definition 2

1.1.2. Why a geographical approach? 3

1.1.3. Scale issues 4

1.1.4. Cases studies 5

1.2. Where and why there? 6

1.2.1. Identification of relevant factors 7

1.2.2. A new framework for zoonoses 7

1.2.3. State of the current knowledge 8

1.3. Challenges of modelling zoonoses 18

1.3.1. Data sources 18

1.3.2. Modelling techniques 22

1.4. Thesis outline 26

Chapter 2: Comparison of methods to model human zoonoses data . 29

2.1. Introduction 30

2.2. Materials and methods 31
2.2.1. Materials 31
2.2.2. Methods 38
2.3. Results 46
2.3.1. Logistic regression 46
2.3.2. Boosted regression tree 48
2.3.3. Cross-validated logistic regression 48
2.3.4. MaxEnt model 51
2.3.5. Comparison between models 51
2.3.6. Partial analyses with variables related to bank voles, virus and humans 52
2.4. Discussion 53
2.4.1. Input data 53
2.4.2. Ease of use 55
2.4.3. Goodness-of-fit and predictive power 55
2.4.4. Interpretation 57
2.4.5. Further proposals for modelling 61
2.4.6. Human hantavirus infections in Sweden 61
2.5. Conclusion 62

Chapter 3 Shaping zoonoses risk: the case of tick-borne encephalitis in Sweden 63

3.1. Introduction 64
3.2. Materials and methods 65
3.2.1. Materials 65
3.2.2. Methods 74
3.3. Results 77

3.3.1. Principal component analyses 77
3.3.2. Boosted regression trees 78
3.3.3. Measures of the predictive power 82
3.3.4. Use of BRT as explanatory model 84
3.4. Discussion 85
3.4.1. Comparison of hazard and exposure 85
3.4.2. Variables influencing the distribution of TBE in Sweden 88
3.4.3. Use of BRT as explanatory model 90
3.5. Conclusion 92

Chapter 4 Untangle the effect of variables on hazard and exposure for hantavirus in Germany 93

4.1. Introduction 94
4.2. Materials and methods 95
4.2.1. Materials 95
4.2.2. Methods 102
4.3. Results 103
4.4. Discussion 112
4.4.1. Untangling the effect of hazard and exposure 113
4.4.2. The complexity of Germany 117
4.5. Conclusion 118

Chapter 5 Landscape and regional environmental analysis of distribution of hantavirus human cases in Europe 121

5.1. Introduction 122
5.2. Materials and methods 123
5.2.1. Materials 123

5.2.2. Methods 132
5.3. Results 136
5.3.1. At which levels do the variables operate and how do they influence NE distribution? 136
5.3.2. What are the effects of environmental variables across European ecoregions? 140
5.4. Discussion 144
5.4.1. Large-scale modelling 144
5.4.2. Model variables 147
5.4.3. Model predictions 148
5.4.4. Limitation of the models 148
5.5. Conclusion 150

Chapter 6 Conclusion 151
6.1. Contributions of the thesis 151
6.1.1. Hazard and exposure framework 151
6.1.2. Modelling tools 153
6.1.3. Spatial scale issue 154
6.1.4. Factors influencing the spatial distribution of zoonoses 155
6.2. Models limitations 158
6.3. Further perspectives 161

References 163
List of figures 179
List of tables 181
Data table 183

List of abbreviations

AIC	Akaike Information Criterion
AUC	Area Under the Curve
BIC	Bayesian Information Criterion
BRT	Boosted Regression Tree
CV	Cross-Validated
ICC	Intraclass Correlation Coefficient
NDVI	Normalized Difference Vegetation Index
NE	Nephropathia Epidemica
PC	Principal Component
PCA	Principal Component Analyses
PUUV	Puumala Virus
ROC	Receiver Operating Characteristics
SWI	Soil Water Index
TBE	Tick-Borne Encephalitis
VIF	Variance Inflation Factor

Summary

Zoonoses, diseases that usually circulate among animals and that are sometimes transmitted to humans, are complex systems that involve the pathogen, the host, potentially the vector and, humans. Spatial distribution models are often based on disease case records, as they frequently are the most readily available data, including through passive surveillance, if the disease cases are correctly diagnosed and reported. Conversely modelling the distribution of the host or vector requires a massive sampling effort while searching and modelling the pathogen itself necessitates vector, host or human blood samples.

Human disease cases may be seen as the tip of the iceberg, with the undetected zoonotic cycle hidden under the water surface. The question thus arises to know how much can be learned about the immersed part of the iceberg based on the emerging part. A set of tools and a new framework are here suggested to better address the issues raised by the use of human disease case data for modelling zoonoses. Hantavirus and tick-borne encephalitis are examined in diverse environments and at diverse scales to illustrate these concepts.

Our framework is based on the concept of risk assessment that is a combination of hazard (defined as pathogen circulation in the wild) and exposure (defined as people entering into infected landscape). Results suggest that the combination of hazard and exposure is needed to improve the predictive power of models calibrated using human

disease case data. The framework is also useful to investigate how factors are involved in the various parts of the disease transmission system.

The tools presented range from linear regression to niche models and from the landscape to the European scale. A multilevel approach is also advised in many circumstances. For large-scale models, three scenarios of variable response are identified, which bear diverse consequences for modelling and modelling results interpretation.

Chapter 1 Introduction

In the sixties and seventies, with the eradication of smallpox, the emergence of vaccination, antibiotics and better sanitation, it was widely accepted that infectious diseases would meet their end. It was said by T. Aidan Cockburn, physician and anthropologist: "*We can look forward with confidence to a considerable degree of freedom from infectious diseases at a time not too far in the future. Indeed… it seems reasonable to anticipate that within some measurable time… all the major infections will have disappeared"* or by the U.S. Surgeon General: *"It might be possible with interventions such as antimicrobials and vaccines to close the book on infectious diseases and shift public health resources to chronic diseases"* (Fauci 2005).

Today, we can see that they were overly optimistic.

A major reason of their false hope is that nature is an important source of infectious agents and that the global change affords more opportunities for pathogens to reach human (Patz et al. 2000). The majority of emerging diseases in humans originates from wildlife (Daszak et al. 2000, Kruse et al. 2004). Between 1940 and 2004, 60% of the events recorded as a disease emergence were of zoonotic origin (Jones et al. 2008). Zoonoses, diseases that usually circulate among animals and that are sometimes transmitted to humans, are often misunderstood and difficult to control. They can sometimes be dramatic as observed recently with Ebola or simply encountered on

the side of the road, where a tick is questing. As zoonotic diseases originate mostly from wildlife, studying them is challenging: preliminary hypotheses are difficult to establish because of the lack of knowledge, massive sampling efforts are required and statistical analyses cannot capture all the phenomenon complexity. Many questions are still pending and a geographic approach may bring some clues to understand the spatial distribution of the disease (or the pathogen) and the factors explaining this specific distribution. In this thesis, we offer a new conceptual framework and explore tools to better approach the issue of zoonotic diseases. This is illustrated here with the examples of hantavirus and tick-borne encephalitis in Europe.

1.1. Zoonoses

1.1.1. Definition

The definition of zoonoses raised much debates in the past, even if grouping these diseases was needed in order to join veterinary and medical effort (Nelson 1960). Zoonoses were eventually defined as *"diseases and infections naturally transmitted between vertebrate animals and humans"*(WHO 1959). Nowadays, zoonoses are recognized by national and international agencies as a major health and development issue (Marano et al. 2006).

The transmission system of zoonoses involves the pathogen, the host(s), humans and sometimes, the vector(s). Zoonotic pathogens may be parasites, microbes, virus or fungi (Soulsby 1974). The

diversity of hosts, from domesticated to wild animals, and of vectors, often insects or arachnids, increases the complexity of prevention and control of zoonoses (WHO 2004).

1.1.2. Why a geographical approach?

Zoonoses are at the crossroads of epidemiology, ecology and veterinary sciences, and geographers are well placed to have a multidisciplinary approach.

The relevance of geography in the field of zoonoses is also obvious as the probability of infection usually decreases with the distance to infected host and disease dynamics are influenced by the spatial distribution of the pathogen, host and vector (Ostfeld et al. 2005).

Pathogen, host, vector and human are located in space according to diverse elements (Figure 1). The environment can be characterized by climatic conditions, slopes and type of soil. According to this environment, different land uses and different types of vegetation are observed. Wildlife species have some preferences and limitations and are found in particular places of this environment. Human have also their preferences that depend among else on historical and cultural factors. The transmission of zoonoses happens at the intersection of all of these preferences and factors.

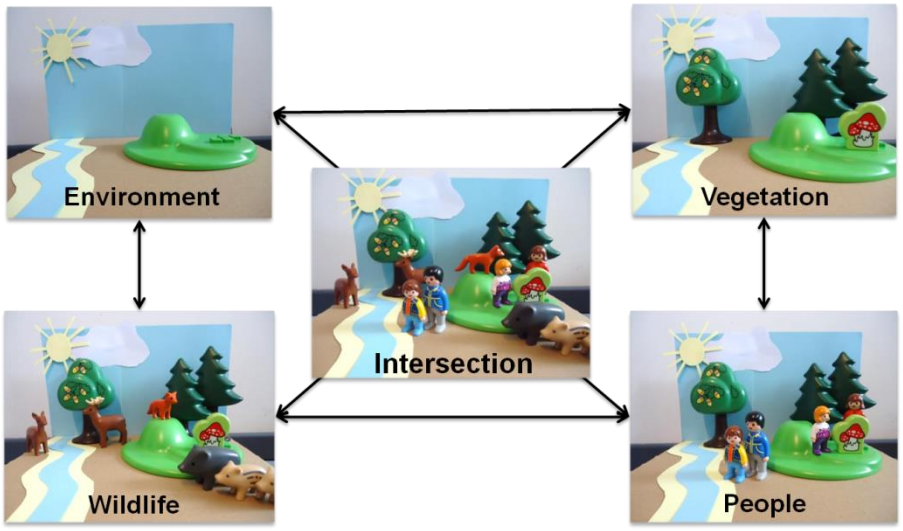


Figure 1: Places for zoonoses transmission

1.1.3. Scale issues

A geographical approach allows to consider issues related to spatial scales. The main issue is that a factor may follow different processes according to the study scale. One example is the distance to the sea, which is a proxy for the attractiveness at a local scale and a proxy for climatic buffering effect at a broader scale. The different scale levels range from the planet, continental, regional, landscape and local ecosystem level (Forman 1995). Transitions between these scales may not follow a gradual change, even if the structure is nested. Processes occurring differently according to the scale level, correlation at one level cannot be used to make assumptions for a smaller level, this is the ecological fallacy (Robinson 2009). This does not preclude that variables acting at a large-scale (e.g. climatic variables and regional

level) can be studied at a smaller scale (e.g. landscape scale). Indeed, the landscape scale can be indicated to model the complex interactions that happen between several components: climatic changes and ecosystem changes, land use and human behaviours and vector ecology and infected hosts (Lambin et al. 2010). Also, pathogen, host and vector abundance and movements are largely determined by the landscape (McCallum 2008).

A large study extent allows covering broad areas but implies the identification of the different scale levels covered in the study area, including the landscape scale. In this thesis, different scales are combined in order to better approach these challenges, which was poorly done in previous studies on zoonoses spatial distribution.

1.1.4. Cases studies

In this thesis, we focus on two viral zoonoses: tick-borne encephalitis in Sweden and hantavirus in various countries in Europe.

Tick-borne encephalitis (TBE) virus belongs to the family of Flaviviruses and the western subtype of TBE virus is usually transmitted by *Ixodes ricinus* ticks (Dumpis et al. 1999). Ticks pass through three active life stages (larvae, nymphs and adults) and need a blood meal to develop into the following respective stage or, for female adults, to lay eggs (Parola and Raoult 2001). Transmission of TBE virus among ticks mainly occurs during co-feeding, especially

between uninfected larva and infected nymph feeding on rodents (Randolph 2011, Randolph et al. 1996, Randolph et al. 1999).

Human hantavirus infections are a zoonotic disease of public health importance in Europe, and a major rodent-borne disease (Vaheri et al. 2012). Hantavirus belong the Bunyaviridae family (Bishop et al. 1980, Hart and Bennett 1999). Puumala virus (PUUV), the most prevalent hantavirus in western and northern Europe (Clement et al. 2003), is responsible for hemorrhagic fever with renal syndrome, often called nephropathia epidemica (NE) in humans (Olsson et al. 2010, Heyman et al. 2011). The main reservoir is the bank vole (*Myodes glareolus*) (Leduc 1989). Transmission between bank voles can be direct but is mainly indirect by inhalation of dust of urine and faeces of infected bank voles, just like transmission to humans may be direct, by biting, but is mainly indirect (Vanapeldoorn et al. 1992, Kozakiewicz et al. 2000, Kallio et al. 2006). At room temperature (and colder) and away from UV light, the virus remains infective for at least two weeks (Kallio et al. 2006). Transmission to human may therefore happen while cleaning closed and un-aired buildings or handling firewood where the bank voles have previously defecated (Dearing and Dizney 2010).

1.2. Where and why there?

The main objective of this thesis is to help answering two fundamental questions on zoonoses: “Where is the disease located?” and “Why is

the disease there?". This can be done with the help of explanatory models investigating the role of the environment as per the spatial determinants outline in section 1.1.2. We also suggest a new framework based on the concepts of risk assessment to untangle the different sources of risk.

1.2.1. Identification of relevant factors

We already saw in section 1.1.2. that zoonoses infections happen at the intersection of multiple factors. Such a frame allows outlining preliminary assumptions on important factors, and models can be used to assess their(s) impact(s). In this context, the elaboration of the "best model" (e.g. with the best predictive power) is not the primary aim. Potential factors are chosen based on preliminary studies or assumptions and, according to the data availability of data, factors or their proxies are used in the model.

1.2.2. A new framework for zoonoses

The global change that currently affects our world increases the uncertainty on emergence and distribution of zoonoses. However, the precise mechanisms leading to a greater risk of zoonoses for humans are still poorly understood. In this context, there is a need to frame and to go further in the understanding of processes. This had led us to think about a new framework to approach the risk of zoonotic infections based on the concepts of risk assessment.

Risk assessment includes the identification of hazard and the characterization of exposure (FAO 1999). The hazard is any potential source of damage (e.g. radioactive radiation), while the exposure is the chance that populations will be in contact with the hazard (e.g. work in a nuclear power plant).

In the context of zoonotic diseases, we define hazard as the number of infected hosts or vectors in the environment. This is determined by ecological conditions allowing the hosts, vectors and pathogens to complete their life cycles and to overlap. Exposure concerns the degree to which humans get in contact with infected hosts/vectors. This largely relates to land use, including the ability and the attraction to access places where infected hosts/vectors are found.

Many disease ecology studies focus on what is here defined as hazard. Exposure is more commonly addressed by the field of public health which often does not include landscape-related variables. However, the distribution of disease cases potentially results from the combination of both hazard and exposure, and therefore cannot be approached solely from one angle.

1.2.3. State of the current knowledge

The creation of risk maps and/or the identification of relevant factors are of primary importance to improve people's health. Previous studies have therefore already investigated the spatial distribution of

zoonotic diseases and identified factors affecting their distribution, mostly from the perspective of hazard.

Factors can be identified through four different types of source: laboratory analysis, field study (including public health survey), modelling and, expert and common knowledge coming from experience. These studies can of course combine sources: a preliminary field data collection followed by the elaboration of a mathematical model.

Such studies bring their contribution to the building of the knowledge necessary for public health surveillance, prevention, diagnosis and treatment. This thesis likewise contributes to the building of this knowledge but also offers a more holistic approach. Indeed, studies generally focus either on hazard or exposure and it appears from a review of the literature that these concepts are rarely approached together. The scarce studies that apply a similar framework or perspective (e.g. (Linard et al. 2007a) or (Vanwambeke et al. 2010)) do it implicitly or incompletely, e.g. by omitting to account for both aspects in the interpretation. With our unifying framework, interesting factors that are often studied under one point of view can be highlighted in a global context. The framework also affords further clues to go beyond of the current disease understanding.

Some examples of typical studies of each category are here reviewed for tick-borne encephalitis and hantavirus. Chosen studies identify

relevant factors that will be used in the following chapters of this thesis.

1.2.3.1. Tick-borne encephalitis

Diverse factors influence the spatial distribution of TBE (Table 1). Hazard is related to the tick *Ixodes ricinus*, the hosts (from small to large animals such as rodent or red deer) and the virus. Ticks have been extensively studied and several factors impacting their mortality and behaviours were already identified. Exposure is investigated through risk activities, socio-economic status and places accessibility.

a) Hazard

Climatic and environmental factors determine ticks survival and the ability of ticks to find and to feed on hosts. Climatic conditions, especially moisture and temperature, affect the survival of ticks but also their behaviour (Gray 2002) such as, for example, a higher temperature favours the tick development and birth rate (Bormane et al. 2004). Several models highlighted the importance of moisture and temperature. A spatial model of probability of ticks presence in the United States found a significant relation to temperature and vapour pressure (Brownstein et al. 2003); a spatial model on different ticks species presence in California related to temperature and wetness (Eisen et al. 2006); an empirical model linking ticks abundance with temperature and humidity in Slovakia (Pangrácová et al. 2013); a

model identifying precipitations as a significant predictor of tick presence (Jore et al. 2014) or; a model linking ticks and low minimum temperature in winter and, an extended spring and autumn in Sweden (Lindgren et al. 2000). There are also a lot of laboratory experiments concerning the effect of humidity on ticks. They showed that dry conditions are limiting for the ticks and increase their mortality (Rodgers et al. 2007). It also affects the tick behaviour as the tick will search moisture rather than host (Crooks and Randolph 2006) and it was also observed that the ticks go down to the moist lower vegetation and will feed more often on hosts found there: rodents (Randolph and Storey 1999).

Environments affecting the hazard are identified through field studies, models or expert knowledge. In the United States, the tick presence was modelled in relation to landscape composition, soil and bedrock composition (Guerra et al. 2002). Forests appeared as an important factor to predict TBE in Czech Republic (Daniel et al. 1998) as well as land cover classes, temperatures and vegetation index in the Baltic States (Šumilo et al. 2006). In Sweden, a field study revealed that forested areas are preferred by the ticks over open areas and, more precisely, beech forests are preferred over pine forest (Lindström and Jaenson 2003). Another field study in Spain showed that there are no ticks in young coniferous forest and open areas, few in coniferous forests and several in old forests with secondary plant regrowth (Estrada-Peña 2001). Also, in Belgium, the abundance of ticks is more important in oak forests rather than pine forests and increases with the presence of shrub cover (Tack et al. 2012). Another model showed

that transition from coppice to high stand forest is also a risk factor in Italy (Rizzoli et al. 2009). In addition, this study highlighted the importance of tick hosts and, particularly, roe deer. The roe deer is also identified as a major host for the ticks according to expert knowledge (Jaenson et al. 2012a, Jaenson et al. 2012b) or knowledge coming from field trips, where deer faeces were notified in TBE areas in England (Robertson et al. 2000). Other important hosts for the persistence and transmission of TBE are rodents such as field voles (*Apodemus flavicollis and A. sylvaticus*) or bank voles (*Myodes glareolus*) (Labuda et al. 1997). Forest ecotones, the area of transition between the forest and another land use, are important for both ticks and tick hosts (Despommier et al. 2006, Lindström and Jaenson 2003, Tack et al. 2012, Allan et al. 2003). Finally, the forest patch size and isolation are significantly related to tick density around Lyme in the United States, highlighting that the risk is lower in fragmented forested landscape (Brownstein et al. 2005).

b) Exposure

Any activity leading to contact with ticks (and so, any activity in contact with vegetation) is a risk activity. Surveys conducted among people presenting antibodies identified risk activities, such as orienteering, in Sweden (Gustafson et al. 1993), or picking berries and mushroom, or owning a summer house in the area of Stockholm (Haglund et al. 1996). TBE was linked with poverty and household expenditure on food in a model on the particular context of central and Eastern Europe (Šumilo et al. 2008). In the same context, another

model highlighted that land ownership of forested areas is important to understand the risk of TBE at the landscape level (Vanwambeke et al. 2010). Finally, it is part of the common knowledge that TBE is caught during leisure time as it is noticed for Germany in (Süss et al. 2004).

c) Hazard and exposure

We only found few studies containing factors related to both hazard and exposure (Vanwambeke et al. 2010, Linard et al. 2007a, Cortinas et al. 2002), without being explicit on these terms. In a first study on TBE, Šumilo identified factors affecting the hazard (Šumilo et al. 2006) and emphasised the importance of considering other variables linked to socio-economic condition and human behaviour. These were considered in a second study (Šumilo et al. 2008) but without linking the two models.

Table 1: Factors influencing hazard and exposure to tick-borne diseases

	Hazard	Exposure
Humidity	(Randolph and Storey 1999) (Gray 2002) (Brownstein et al. 2003) (Crooks and Randolph 2006) (Eisen et al. 2006) (Rodgers et al. 2007) (Pangrácová et al. 2013) (Jore et al. 2014)	
Temperature	(Gray 2002) (Bormane et al. 2004) (Brownstein et al. 2003) (Eisen et al. 2006)	

	(Pangrácová et al. 2013) (Lindgren et al. 2000)	
Land use	(Guerra et al. 2002) (Šumilo et al. 2006)	(Haglund et al. 1996)
Forest	(Daniel et al. 1998) (Estrada-Peña 2001) (Lindström and Jaenson 2003) (Brownstein et al. 2005) (Rizzoli et al. 2009) (Tack et al. 2012)	
Tick hosts	(Labuda et al. 1997) (Robertson et al. 2000) (Rizzoli et al. 2009) (Jaenson et al. 2012a) (Jaenson et al. 2012b)	
Forest ecotone	(Allan et al. 2003) (Lindström and Jaenson 2003) (Despommier et al. 2006) (Tack et al. 2012)	
Vegetation index	(Šumilo et al. 2006)	
Risk activities		(Gustafson et al. 1993) (Haglund et al. 1996) (Süss et al. 2004)
Socio-economic conditions		(Šumilo et al. 2008)
Land ownership		(Vanwambeke et al. 2010)

1.2.3.1. Hantavirus

A broad range of conditions and diverse environments are encountered across Europe and various factors identified from the literature may be more relevant to explain the NE distribution in some or other areas. Several interesting reviews on hantavirus in Europe (Heyman et al. 2012, Olsson et al. 2010, Heyman et al. 2011, Vaheri

et al. 2013) tried to map the uneven spatial distribution of hantavirus disease(s) and made hypotheses on the underlying factors. To answer the question “Where is the disease located?”, these studies collected disease data at various scale but never very precisely and the results were inaccurate maps. To answer the question “Why there?”, they made assumptions mainly based on expert knowledge but also on local studies. In this thesis, we offer to test these assumptions thanks to our framework and accurate multilevel spatial modelling. A summary of factors is presented in Table 2.

a) Hazard

Hazard is determined by the bank vole presence and abundance as well as the ex-vivo virus survival, important for the indirect transmission. Factors related to hazard may be driven by the climate or the environment.

Kallio et al. (2006) established in a laboratory experiment that higher humidity and cold temperatures favour ex-vivo virus survival (Kallio et al. 2006). Snow, allowing cold and wet conditions for the ex-vivo virus survival, is an important but complex variable with a nonlinear effect. Field studies showed that in areas where the snow cover is too abundant (northern Fennoscandia), there is lower prey diversity and more pressure on bank voles (Hansson and Henttonen 1985, Hanski et al. 1991). This was also reflected in a prey-predator model (Hanski et al. 2001). Another field study in Fennoscandia indicated that snow, in the absence of an ice sheet, preserves food for the bank voles

(Korslund and Steen 2006). In areas where snow is often abundant (southern Fennoscandia), the prevailing expert opinion is that higher winter temperatures may force the bank voles to migrate near dwellings because of the lack of protective snow cover (Evander and Ahlm 2009). Conversely, in areas where the snow cover is generally low (e.g. Belgium, France, Germany, the Netherlands), expert opinion is that extreme cold winter with thick snow cover may force bank voles to migrate near human dwellings (Faber et al. 2010). It is also hypothesized that mild winter allows a better rodent survival during the winter and earlier food resources for the next spring (Schwarz et al. 2009).

Field analysis in Belgium identified that bank voles are more found in wet environment over drier environment (Verhagen et al. 1986) and two models reflected that soils with thinner particles, retaining more water, allow better virus persistence (Linard et al. 2007b, Sauvage et al. 2007). An empirical study in Belgium identified forest and especially broadleaved forest as the main habitat of bank vole (Linard et al. 2007b). Coniferous mature and moist forest with well-developed undergrowth can however be favourable as identified in a model for hantavirus in Sweden (Olsson et al. 2005). The combination of field studies with modelling techniques demonstrated that contiguous, well-connected forests may favour virus circulation in rodents (Guivier et al. 2011, Kozakiewicz et al. 2000) and that habitat fragmentation in the Netherlands decreases the population of bank vole (Vanapeldoorn et al. 1992).

Another factor identified from empirical-statistical studies is the vegetation growing period that increases the population of bank vole by enhancing the carrying capacity and the number of breeding opportunities for bank voles in Belgium (Barrios et al. 2010).

b) Exposure

In France and Belgium, the proportion of built-up areas around forest patches and buildings within 50m of forest were modelled as risk factors (Linard et al. 2007b, Crowcroft et al. 1999). In Fennoscandia, Belgium and the Netherlands, different surveys among infected people were conducted to identify risk activities (Groen et al. 1995, Vapalahti et al. 1999) as well as one model (Linard et al. 2007a). These activities are being a trapper, a farmer or a forestry worker. As indicated in (Piechotowski et al. 2008), it is part of common knowledge that higher winter temperature may relate to more human outdoors activities and a higher exposure to bank vole excreta.

c) Hazard and exposure

As for TBE, hazard and exposure are rarely studied together (Linard et al. 2007b, Linard et al. 2007a). It is then done implicitly and the interpretation often focused on each aspect separately.

Table 2: Factors influencing hazard and exposure of hantavirus infections

	Hazard	Exposure
Humidity	(Kallio et al. 2006)	
Temperature	(Evander and Ahlm 2009) (Faber et al. 2010) (Schwarz et al. 2009) (Kallio et al. 2006)	(Piechotowski et al. 2008)
Snow	(Korslund and Steen 2006) (Kallio et al. 2006)	
Prey diversity	(Hanski et al. 1991) (Hanski et al. 2001) (Hansson and Henttonen 1985)	
Wet environment	(Linard et al. 2007b) (Sauvage et al. 2003) (Verhagen et al. 1986)	
Forest	(Linard et al. 2007b) (Olsson et al. 2005)	
Built-up	(Linard et al. 2007b)	(Linard et al. 2007b) (Crowcroft et al. 1999)
Contiguity of forest	(Guivier et al. 2011) (Kozakiewicz et al. 2000) (Vanapeldoorn et al. 1992, Guivier et al. 2011)	
Vegetation index	(Barrios et al. 2010)	
Risk activities		(Linard et al. 2007a) (Groen et al. 1995) (Vapalahti et al. 1999)

1.3. Challenges of modelling zoonoses

1.3.1. Data sources

Investigating the spatial distribution through statistical models of the host or vector usually requires a massive sampling effort and tracing the pathogen circulation necessitates extensive analyses of vector, host or human blood samples. Often, human records of zoonotic disease are the easiest component to record and spatial modelling of zoonoses

are therefore frequently based on human disease case data. These records reflect at the same time hazard and exposure, their accuracy is often restrictive and the absences found in the database, which indicate absence of records, do not necessarily reflect the absence of pathogen circulation.

1.3.1.1. Dependent variable

a) The issues of human records

Recording a presence may be interpreted as a probabilistic function that depends on the abundance of the species/disease and on its detectability (Brotons et al. 2004). Several issues arise from the use of human data.

The main issue is that human records reflect both hazard and exposure. They can be seen as the tip of the iceberg which hides an undetected enzootic cycle below the surface (Randolph 2008). These records only cover a part of the total number of infections and of the disease extent. It is difficult, based on human cases, to quantify the disease in wildlife or in domesticated animals.

Another issue of human records is that their spatial precision and accuracy are often limiting. The most precise and accurate records are geographical coordinates of the infection place (Table 3). Data are often provided by place of residence or place of records (doctors, hospitals, laboratory centres ...) and by administrative units.

Administrative units are often more available than coordinates due to privacy rules. Data on the place of infections are quite rare because it requires surveys and/or is difficult to identify. Precise and accurate human data are thus uncommon.

Table 3: Precision and accuracy of human zoonose records

<table>
<tr><td>Place of record</td><td>Place of residence</td><td>Place of infection</td></tr>
<tr><td colspan="3">----------ACCURACY AND PRECISION---------></td></tr>
<tr><td colspan="2">Administrative unit</td><td rowspan="2">Geographical coordinates</td></tr>
<tr><td>Larger</td><td>Smaller</td></tr>
</table>

b) The issue of absences

Real absences are also challenging to record. Absences, i.e. places where it is undoubted that the organism/pathogen is not present, may be recorded but are often a set of points randomly chosen through the study area. Absences of a zoonotic disease imply the absence of at least one of five elements: animal donor, vector, animal recipient, pathogen and an external environment allowing pathogen circulation (Lambin et al. 2010). Absences may be interpreted in three ways (Lobo et al. 2010):

- Environmental absences, related to unfavourable environmental and climatic conditions (not in potential or realized distribution),
- Contingent absences, located in favourable areas (within the potential but not in the realized distribution) and,
- Methodological absences, caused by a bias in the data collection

Hopefully, if the ability to detect a species is constant across the study area (and differs from zero), then absences are reliable or associated to habitats where prevalence is low (Brotons et al. 2004).

1.3.1.2. Explanatory variables

Variables can be limiting according to their availability and their resolution. Data at high resolution often does not cover large study extent, which can be a limitation for large-scale model.

To represent the landscape scale, the explanatory variables are computed in a certain radius around the disease record. The size of this buffer depends of the record accuracy. When data are provided by geographical coordinates of infection place, the buffer size should be related to the biology of the host-vector. When records are given by geographical coordinates of place of residence, the buffer should reflect the potential movement of the dweller. Finally, when the records are provided at coarser resolutions such as municipalities, the buffer size should reflect at least the municipality size. In this thesis,

different relevant buffer sizes were tested according to the nature of the dependent variable. The buffer size with the most number of variables significant in bivariate logistic regressions was chosen.

1.3.2. Modelling techniques

Models are useful to address the questions of the location of zoonoses and of why are the diseases there and not there. In this thesis, the relation between a set of factors and the disease are investigated through collections of data: the disease records and the measurements of factors. This approach is useful to quantify the phenomenon and to detect the complex reality in its wholeness. Indeed, measures of factors are directly taken from the environment (with e.g. remote sensing), without control.

Spatial epidemiology can be seen as an (re-)emerging discipline (Ostfeld et al. 2005), even if factors are still often identified through laboratory or field studies. In this thesis, different tools and methods from various sources are presented and it can be hoped that more people, first attracted by disease knowledge, may be then influenced by spatial epidemiology.

1.3.2.1. Presence and absence models

Many ecological and epidemiological spatial records relate to location-specific records of discrete units such as organisms or

reported disease cases. These records are then represented as presence point records. A number of models allow investigating and predicting the presence of organisms and pathogens based on a set of independent variables. An extensive review of models predicting species habitat distribution is presented in (Guisan and Zimmermann 2000) and diverse methods are compared in (Beale et al. 2010, Anderson et al. 2006). Some of these methods address in various ways the issue of confronting (or not) presences with absences. Models can be classified into three categories with regard to input dependent variable: presence and absence, presence and available, or presence-only.

Some examples of existing techniques are investigated in this thesis and are explained in Chapter 2. First, boosted regression trees (BRT), based on presence and absence data, are decision trees where predictive performance is improved by boosting (Elith et al. 2008). Second, logistic regressions, based on presence and absence data, fit a logistic curve between dependent variable and explanatory variables (Hosmer and Lemeshow 1989). Third, cross-validated logistic regression (CV method), compares presences to available sites, having a different intensity of use rather than being strict absences. Finally, MaxEnt fits model by maximizing the entropy according to the presence distribution and background data (Elith et al. 2011).

A presence-absence approach is used for each study as data were often accurate and provided by coordinates. In Chapter 4, the number of NE was provided by municipalities. BRT is however used as the

population at risk per municipalities is not known and as the results are compared with another presence-absence model on bank vole. In Chapter 5, a presence-absence approach is used in order to standardize the dataset.

1.3.2.2. Data modelling vs. algorithmic modelling culture

The function linking a set of factors and an independent variable can be seen as a black box. On one hand, the data modelling culture fills this black box with a fixed shape (e.g. linear or logistic curve) and, subsequently, fits the model by estimating the parameters of the predictors (Breiman 2001). On the other hand, the algorithmic modelling culture, also called "machine learning", fits the model by filling the black box with a step by step set of operations.

BRT and MaxEnt are machine learning models while logistic regression and the CV method are part of the data modelling culture. These two last methods are much used compared to machine learning models which are rare in the field of spatial epidemiology. As BRTs allow modelling nonlinear processes, which are often encountered in wildlife, this technique is applied for all our studies, filling the gap of machine learning underutilization in the fields of spatial epidemiology and medical geography.

1.3.2.3. Explanatory vs. forecasting models

Models that establish a relation between an independent variable and a set of predictors can be used for three different purposes (Shmueli 2010). First, a model can be descriptive and aiming to summarize and represent a simplified reality. Second, a model can be explanatory and tries to test causal explanation. Third, a model can be predictive and used to predict new or future observations, including forecasting. In epidemiology, models are often used in a mixed way balancing between explanatory and predictive aims (Shmueli 2010).

In this thesis, a focus on explanatory model is used as knowledge on many zoonoses is not yet enough to understand the full dynamics. For this, factors are selected based on our own suspicions or on the existing literature. Explanatory models can be compared based on their predictive power, allowing a "reality check" (Shmueli 2010). Even if the first aim is to explain, the predictive power and the model predictions are here used to compare our models.

Boosted regression trees are often used in the purpose of predictions but, as they allow modelling nonlinearity, they can be of great value as explanatory models. This is investigated in Chapter 3.

1.4. Thesis outline

In this introduction, the challenges raised by the spatial modelling of zoonoses are highlighted: human presences represent both hazard and exposure, human case records often lack of precision and accuracy, the absences records may result from contingent or methodological absences, modelling methods may be more advantageous or disadvantageous when working with human zoonosis data, BRT are mainly used for prediction and not for explanation and, one factor may follow different processes at diverse scales.

We attempt to address these challenges in the following chapters. The methodological issues are studied in Chapter 2 as well as the use of BRT for explanation in the Chapter 3. Our framework may help to better approach disease systems and go a step further in the understanding of zoonoses. The concept of hazard and exposure is applied all along this thesis. In Chapter 3 and 4, evidence for the usefulness of the framework is proposed. Finally, the spatial scale is addressed in Chapter 5.

Chapter 2

The different modelling techniques used in the following chapters are here explained in detail. They concern boosted regression trees, logistic regression, cross-validated logistic regression, MaxEnt model and the validation and evaluation methods. They are applied and compared for hantavirus human infection in Sweden. Here, the data are provided by geographic coordinates of infection places, a high

level of precision and accuracy. General advices on good practices are provided to novices that want to model zoonoses based on human records. Also, variables that affect the hantavirus distribution in the North of Sweden are approached at the landscape scale and nonlinear processes and thresholds are identified.

Chapter 3

In this chapter, the concepts of hazard, related to the landscape ecology for wildlife, and exposure, related to the landscape attractiveness for people, are applied on tick-borne encephalitis in an endemic area of Sweden. Three models are built: a model with variables related to the hazard, a model with variables related to the exposure and a global model containing all the variables. These models are compared in order to highlight the importance of joining hazard and exposure approaches. Furthermore, this chapter investigate the distribution of wildlife, the landscape components and the landscape structure that play a role in a region where the disease is well established.

Chapter 4

After applying our framework for TBE in Sweden, the effect of environmental variables on the hazard and exposure are untangled for a set of factors influencing the spatial distribution of hantavirus in Germany. To better approach the fact that human cases are only the tip of the iceberg, a model based on the bank vole presence is computed in addition to a model based on the human cases. The hantavirus spatial distribution in Germany is for the first time

investigated under the geographic angle and the impacts of variables on the different parts of the disease transmission system are examined.

Chapter 5

This chapter brings clues in the understanding of the uneven and sometimes unexpected spatial distribution of hantavirus in human in Europe. A global model is for the first time built, allowing to go further in the understanding of where the disease is located and why there. Multiscale issues are approached and recommendations for modelling diseases for large study areas are given. The changing effects of each variable are investigated at different scales and across different ecoregions.

Chapter 2: Comparison of methods to model human zoonoses data[1]

In this chapter, we compared several methods for the three approaches that can be used to model spatial distributions: based on presence/absence, presence/available or presence data. General advices on the good practices that should be used by beginners were given. Human cases of hantavirus infection reported by place of infection in Sweden were used as a case study. Logistic binomial regression and BRT were used to model presence and absence data. Presence and available sites (where the disease may occur) were modelled using cross-validated logistic regression. Finally, a MaxEnt model, based on presence-only data, was used.

Logistic regression, when having reliable absences, including variables reflecting the surrounding environment and using the step procedure, had the best predictive power, followed by boosted regression trees, MaxEnt and cross-validated logistic regression. The cross-validated method partly avoids the issue of absence data but requires fastidious calculations. Boosted regression trees and MaxEnt account for nonlinear responses. The advantages and disadvantages of each method are reviewed.

[1] Adapted from: Zeimes, C. B., G. E. Olsson, C. Ahlm & S. O. Vanwambeke (2012) Modelling zoonotic diseases in humans: comparison of methods for hantavirus in Sweden. *International Journal of Health Geographics,* 11**,** 39.

Chapter 2

2.1. Introduction

Modelling techniques are numerous but some may perform better than others to model zoonoses based on human records. They present diverse (dis)advantages linked to the input data or the interpretation of the results. Here, we compared modelling techniques and practices in order to present these advantages and disadvantages. In this way, people that are unfamiliar with spatial modelling can have a first approach and apply good practices in a straightforward fashion. For this purpose, models were based on data at fine resolution, the geographic coordinates of places of infection of hantavirus in Sweden, avoiding issues linked to inaccuracy of the data.

In Sweden, Puumala hantavirus (PUUV) is the most prevalent hantavirus and the only pathogenic one (Olsson et al. 2003). In Scandinavia, the peak of NE occurs from November to December but cases may be recorded year-round (Brummer-Korvenkontio et al. 1999). In Sweden, 90% of all NE cases notified are reported from the four northernmost counties (Olsson et al. 2003) and the study area was per consequent limited to these counties. We focused on environmental factors related to bank vole habitat, ex-vivo virus survival and human presence.

Models can be classified into three categories with regard to input data: presence and absence, presence and available, or presence-only. Here, representative methods of each approach were investigated and compared. For the presence and absence, logistic regression and

boosted regression trees (BRT) were tested. For presence and available, cross-validated (CV) logistic regression was computed and, for presence-only, MaxEnt software was used. The probabilities of the four models were mapped and the outputs of each model were compared. The predictive power and the goodness-of-fit of presence-absence approach and presence-only approach have already been compared (Brotons et al. 2004, Elith et al. 2006, Jimenez-Valverde et al. 2008, Fielding and Bell 1997, Gastón and García-Viñas 2011) but here, in addition to comparing the predictive power or the goodness-of-fit, advantages and disadvantages were reviewed and good modelling practices that can be easily implemented by beginners were identified.

2.2. Materials and methods

2.2.1. Materials

The study area covers the distribution range of hantavirus in Sweden (Figure 2). NE has been a notifiable disease in Sweden since 1989. In the present study, cases of NE recorded between 1991 and 1998 were used. Detailed locations of alleged sites of human PUUV exposure were acquired by mail and telephone survey.

During this period and in the region, a total of 1 724 cases of NE were notified, and 1 305 persons (76%) responded to the survey. Of these, 862 were confident about the time and location of PUUV exposure but only 217 could provide information detailed enough to link them to

such an exact location as a property. Data were reported by centroid of the land holding where the infection was acquired. Of the 217 cases recorded, some occurred in the same location. Only one record was kept for each location, leaving 212 presence points. 300 isolated dwellings were selected at random from the Lantmäteriet database (Swedish mapping, cadastral and land registration authority). They were used as absence points in the logistic regression and BRT and as available points in the CV model (Figure 2). As the isolated dwellings reflect the population distribution and density, these absences were considered as reliable.

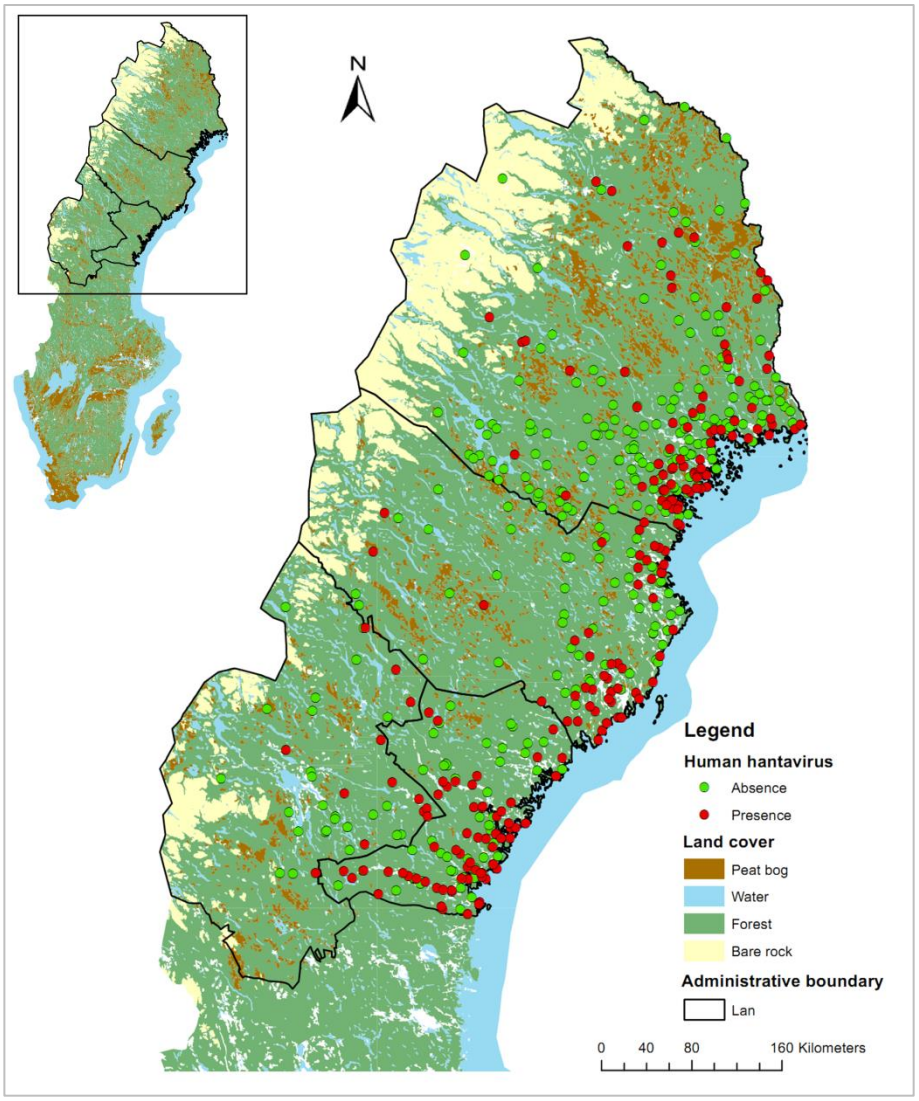


Figure 2: Human PUUV infections in Sweden

Three groups of environmental influences on the distribution of NE were explored (Table 4), relating to bank vole habitat, ex-vivo virus survival and human presence. In northern Sweden, the primary habitat of bank voles is mature and moist coniferous forest. Spruce forests are preferred over pine forests as they provide more food and shelter

(Olsson et al. 2005). Forest data were extracted from the SLU Skogskarta (Sveriges lantbruksuniversitet: http://skogskarta.slu.se). The area of forests, the mean volume of spruce and the mean volume of pine were calculated in a radius of three kilometers around the infection place. Due to their extensive coverage in the area, forests are all connected but locally, forest coverage and configuration vary. Landscape structure indices in a radius of three kilometers around the infection place were computed (number of forest patches, average shape index of forest, distance of the furthest cell from forests (using shortest path), mean contiguity index of forest and mean Euclidian nearest-neighbour distance between patches of forests). As bank vole habitat is often related to peat bogs in North of Sweden (Olsson et al. 2003), the area of peat bogs in a three-kilometers radius was also calculated, based on the land cover data from Lantmäteriet.

Soil grain size was used as a proxy for soil humidity. Data on soil grain size were extracted from the Geological Survey of Sweden (SGU) and classified into coarse, medium and fine particles. Snow depth and average snow duration (when only present for at least 10 days), affecting both the virus ex-vivo survival and the bank vole abundance, were computed from interpolated weather station data of the Swedish Meteorological and Hydrological Institute.

Human density presence follows a double gradient: from South to North and from East to West. Population density (Gridded Population of the World from Center for International Earth Science Information Network (CIESIN)) was used to reflect the spatial distribution of

humans. Distance to the sea and elevation follow also this gradient. Elevation data were extracted from the Aster GDEM elevation data (Global Digital Elevation Model, Earth Remote Sensing Data Analysis Center (ERSDAC)). These two variables may act as proxies for climate, soil composition and attractiveness of the landscape. Other variables were also chosen to reflect the exposure (by attractiveness and accessibility of the landscape): distance to the nearest holiday house, extracted from the Central Statistical Bureau data (Statistiska Centralbyran) and, length of water ways and roads in a three kilometer radius (Lantmäteriet).

Independent variables with non-normal distribution were log-transformed (volume of spruce, distance to forests, area of peat bogs, elevation, distance to sea coast, population density and distance to holiday homes). For logistic regression, BRT and CV model, some variables were expressed as a value in a radius around the infection place, allowing consideration of the landscape encountered around the place of infection. MaxEnt however only allows spatially continuous variables and cannot integrate these variables in a straightforward fashion. Data layers included in MaxEnt were: forests, volume of spruce, volume of pine, peat bogs, soil grain size, snow depth, snow period, elevation, distance to the sea, population density, roads, holiday houses and water ways.

Table 4: Variables and hypothesized relationships with the abundance of bank voles, ex-vivo virus survival and human presence (* variable log-transformed)

Variable in logistic model, BRT and CV model	Variable in MaxEnt model	Abundance of bank voles	Ex-vivo virus survival	Human presence	Source
Area of forests in a 3-km radius around the dwelling (m^2)	Forests	x			SLU Skogskarta
*Mean volume of spruce per hectare in a 3-km radius around the dwelling (m^3/ha)	Volume of spruce	x			SLU Skogskarta
Mean volume of pines per hectare in a 3-km radius around the dwelling (m^3/ha)	Volume of pine	x			SLU Skogskarta
*Maximum distance to forests in a 3-km radius around the dwelling (m)		x			SLU Skogskarta
Number of patches of forests 3-km radius		x			SLU Skogskarta
Mean shape index of forests 3-km radius		x			SLU Skogskarta
Mean contiguity index of forests in a 3-km radius		x			SLU Skogskarta
Mean Euclidian nearest-neighbor distance between patches of forests in a 3-km radius (m)		x			SLU Skogskarta

*Area of peat bogs in a 3-km radius around the dwelling (m^2)	Peat bogs	x			SVK
Mean snow depth between 1991 and 1998 (cm)	Snow depth	x	x		SMHI
Average duration of the snow when it is present for at least 10 days (days)	Snow period	x	x		SMHI
Majority of grain size of the soil (1 = coarse, 2 = medium, 3 = fine) in a 3-km radius	Soil grain size		x		SGU
*Elevation (m)	Elevation	x	x	x	Aster GDEM
*Distance to the sea coast (m)	Distance to the sea	x	x	x	SVK
*Population density (inhabitant/km^2)	Population density			x	Gridded population of the world
Total length of public roads in a 3-km radius (m)	Roads			x	SVK
*Distance to holiday homes (m)	Holiday homes			x	Statistiska Centralbyran
Total length of the water ways in a 3-km radius (m)	Water ways			x	Swedish Places

2.2.2. Methods

2.2.2.1. Presence vs. absence: logistic regression

The logistic regressions fit a logistic curve between dependent variable and explanatory variables for which the estimators are calculated by maximum likelihood to maximize the probability of obtaining the observed sample (McCullagh and Nelder 1989, Hosmer and Lemeshow 1989). The intercept determines the position of the logistic curve on the dependent variable (Rogers 2006). The coefficient of independent variable is the rate of change of the logit function per unit of change of the variable (Hosmer and Lemeshow 1989). This estimator shows how fast the curve will increase or decrease.

First, analyses with one explanatory variable were carried out for each independent variable. After, variables were selected for the multiple model using a backward stepwise procedure in R (“stats” package). The AIC was used to select the best model and the VIF (“car” package) was checked to avoid multicollinearity issues. Multicollinearity happens when predictors are linearly related, leading to an overestimation of the estimators (Dormann et al. 2013). The square root of VIF indicates the number of time the standard error is higher than if the predictors were not correlated. Interactions between variables were tested but none was significant at the level of five percent. As quadratic terms decreased the goodness-of-fit of our model, they were not included.

2.2.2.2. Presence vs. absence: boosted regression trees

BRT combines decision trees with boosting to improve the performance ("gbm" package in R) (Elith et al. 2008). In a regression tree, a branch leads to several internal nodes or to a terminal node. The path chosen at each internal node depends on the value of the explanatory variables. At a terminal node, a decision is made on presence or absence. Decision trees are built by recursive binary split: initial trees are enlarged by new binary splits made on the previous trees. Boosting allows improving the optimization by adding new trees that reduce the most the loss in predictive performance. The procedure is forward and stage wise: after one step, a new tree is fitted on the residual of the previous tree and the new model, with new residuals, contains the previous and the new trees.

To improve BRT performance, three parameters may be at play (Ridgeway 2007). First, a low learning rate, which is the contribution of each tree to the model, is advised. Second, the bag fraction, the percentage of data randomly selected at each step to build the tree, introduces stochasticity in the model. It allows the model to run faster, improves the accuracy and avoids overfitting (Elith et al. 2008). Overfitting occurs when the model starts to fit the noise in addition to the signal, leading to poorer interpolation abilities. The default bag fraction used in this thesis is 0.5. Third, the tree complexity represents the number of nodes in a tree. For a given number of trees, a higher tree complexity implies thus a lower learning rate.

In this thesis, cross-validation (based on 10 folds) was used to select the number of trees for which predictive deviance was the lowest. Models were run on nine subsamples and the predictive deviance was calculated on a tenth sub-sample. This step is repeated 10 times, using each time a different excluded sub-sample. The final predictive deviance is the mean of the predictive deviance calculated on the 10 evaluation sub-samples. Then, the final model was subsequently fitted based on this optimal number of trees (function "gbm.step" from (Elith and Leathwick 2014)).

The contribution of each variable can be calculated and the effect of the variables graphed. The relative importance of each variable represents the number of times a variable is used in successive trees, weighted by the mean of the squared improvement provided by this variable to each tree (Elith et al. 2008). Response curves are graphs representing the evolution of the fitted probability function according to the variation of the variable.

Interactions are automatically modelled because the response of one variable depends of the previous responses of the other variables higher in the tree (Elith et al. 2008). The relative strengths of interactions can be quantified and plotted.

BRT was used to model hantavirus infections. The learning rate and tree complexity were chosen based on visual analyses of graphs and on the predictive power.

2.2.2.3. Presence vs. availability: cross-validated logistic regression

The cross-validated method (CV method) (Boyce et al. 2002, Johnson et al. 2006, Wiens et al. 2008) is based on the principle that the presence of an organism depends on the presence of resources it uses. Each point has a different resource availability and hence of intensity of use (and not just the presence or absence of resources). The probability to find an organism in one place depends of its intensity of potential use. This method uses a classic logistic regression method, but the evaluation and use of the results focus on the computed probabilities and a cross-validation of the predicted probabilities.

In the cross-validation, data are divided into different subsamples. Logistic regressions are then calculated using each time a different combination of sub-samples. The other sub-samples, not used for calibrating the model, are put together and used for validation. Estimators of the different regressions are averaged to produce the final model, which is applied to the validation sample to predict probabilities. The predicted probabilities calculated for the validation sample are clustered in groups of probabilities using quantiles. Validation requires calculating the utilization of resources for each cluster $U(x_i)$:

$$U(x_i) = \frac{w(x_i)A(x_i)}{\Sigma_j w(x_i)A(x_i)}$$

Where $w(x_i)$ is the mid-point probability of the cluster i and $A(x_i)$ is the area of cluster i (here, the number of observations in cluster i).

New predicted presences are calculated by multiplying $U(x_i)$ by the total number of observations for each class. To validate the model, these predicted presences can be compared with observed presences for each class:

1) Spearman coefficient (and $\chi 2$ test of goodness-of-fit) can compare predicted and observed values. A high positive correlation is desired.
2) A linear regression of predicted cases (x) on observed cases (y) can be modelled:
    a. R^2 is used to assess the predictive power.
    b. The intercept is expected to be zero and the estimated regression coefficient is expected to be around one.

For our case study, the cross-validation was based on five sub-samples and the predicted probabilities were clustered in 10 classes.

2.2.2.4. Presence-only: MaxEnt model

MaxEnt model is a presence-only approach using background data. MaxEnt is a model that maximizes entropy satisfying any constrain on the unknown distribution (Phillips et al. 2006), and that minimizes the relative entropy between two probability densities defined in the covariate space, one estimated from the presence data and one from the landscape (Elith et al. 2011). The entropy is here defined as the

inverse of the sum of the probability distribution for each pixels multiplied by the natural logarithm of this distribution (Phillips et al. 2006). To maximize the entropy, the constraints on the probability distribution must be limited.

MaxEnt builds an occurrence model starting from a uniform distribution of probability for each cell of the raster (Phillips 2012). Then, it improves the model iteratively until the gain becomes saturated. The gain is a likelihood statistic which maximizes the probability of presence according to the background data.

All dependent variables are used simultaneously. Collinear variables are usually not considered as a problem because if a variable has a significant impact on probabilities, variables correlated to it will have little impact (Elith et al. 2011).

MaxEnt allows to account for sampling bias by including an additional layer representing the relative survey effort across the landscape. MaxEnt also provides response curves showing the influence of a variable on the probability of presence. Jackknife analyses can be used in order to evaluate the contribution of each variable to the model. Cross-validation can also be asked.

Unlike the two previous models, dependent variables are continuous spatial data layers with identical extent and resolution. Here, the extent raster input was the four northern Swedish counties where NE is recorded and the resolution was one kilometer. The population

density layer was tested as sampling bias and Jackknife analyses were asked. Also, five-fold cross-validation was used.

2.2.2.5. Comparison between models

Goodness-of-fit of several regressions can be compared with the Akaike Information Criterion (AIC), and the predictive power of models can be measured with the area under the curve (AUC) and Cohen's kappa statistic.

The AIC gives a relative measure of the goodness-of-fit of the model according to the maximized value of the likelihood function and the number of parameters (Akaike 1987). Between different models, the one with the lowest AIC should be preferred.

The AUC comes from receiver operating characteristics (ROC) analysis (Pearce and Ferrier 2000, Fawcett 2006). The rate of true positives is plotted against the rate of false positives at all thresholds of classification into presence and absence. An AUC equally to 0.5 is a random distribution of predictions, an AUC higher than 0.7 indicates a good model and, an AUC equally to one, a perfect prediction.

Cohen's kappa statistic is an index of agreement for positive and negative observations (Kirkwood and Sterne 2003). A kappa over 0.75 indicates an excellent agreement; between 0.4 and 0.75, a fair to good agreement and under 0.4, a poor agreement.

Due to the limited size of the database, all data points were used for model training. No external validation was carried out and internal indices were used to compare models.

As the logistic and CV models were built with the same variables and dataset, the Akaike's Information Criteria (AIC) was used to choose the model with the best goodness-of-fit.

To evaluate the predictive power of the four models, the AUC and the Cohen's kappa statistic were calculated ("PresenceAbsence" package in R).

Probability maps were created for each model. For logistic regression, BRT and CV, predicted points probabilities were interpolated by kriging to obtain continuous maps. MaxEnt provides a continuous probability map.

False positives and false negatives were also mapped. Even if the CV model usually does not classify probabilities into presences and absences, a map was made for comparison. The probability threshold was chosen at the level where sensitivity (number of true positives divided by the sum of true positives and false negatives) equals specificity (number of true negatives divided by the sum of true negatives and false positives).

The AUC provided by the MaxEnt software is calculated over the entire study area (presences and background), while the others model

AUC only considered the set of points. Therefore, AUC and kappas were computed on an identical set of points and variables for the MaxEnt model, in order to make an accurate comparison.

2.2.2.6. Partial analyses with variables related to bank voles, virus and humans

Partial logistic regressions, BRT and MaxEnt models were fitted using variables related to each element. In this manner, the relative importance of bank voles, virus and humans distributions on human infections distribution may be speculated.

2.3. Results

2.3.1. Logistic regression

Several variables were significant in the bivariate analyses ($p < 0.05$):

- with a positive sign: logarithm of mean volume of spruce, mean volume of pine, logarithm of distance to forests, logarithm of population density and,
- with a negative sign: logarithm of area of bogs, snow depth, snow period, logarithm of elevation, logarithm of distance to the sea, logarithm of distance to holiday homes.

The multiple logistic model included six explanatory variables (Table 5). Forest contiguity and snow depth were retained in the stepwise

procedure but were not significant ($p > 0.05$). The probability of presence significantly increased with the area of forests, logarithm of distance to forests, contiguity and population density. It decreased with snow depth and logarithm of distance to the sea.

With an AUC of 0.97 and a kappa index of 0.76, the logistic regression had a good predictive power and an excellent agreement. The probability of presence decreased from South to North and from East to West (Figure 3). False positives were found mostly in the East, where the model overestimated presence. False negatives were sparser.

Table 5: Models obtained by logistic regression and cross-validated logistic regression method

	Estimate of logistic regression	**Mean-estimate of CV method**
Intercept	−5.371**	−5.447
Area of forests	$8.048*10^{-8}$***	8.133^{-8}
Log (distance to forests)	1.665***	1.689
Contiguity	1.198	0.226
Snow depth	−0.016	−0.016
Log (distance to sea)	−0.470**	−0.471
Log (population density)	0.544*	0.109
AIC	629.74	792.77
AUC	0.972	0.721

2.3.2. Boosted regression tree

A learning rate of 0.01 and a tree complexity of five were chosen, giving an optimal number of trees of 350.

Variables with the most important contributions were: area of forests (11.45%), distance to holiday homes (11.01%), distance to the sea (9.82%), elevation (8.81%) and mean volume of spruce (8.53%).

Interaction effects were the most important between the sum of roads and area of forests, the snow period and snow thickness, and the elevation and area of forests.

The AUC of 0.92 and kappa of 0.65 indicated a good model and a good agreement. Predicted probabilities generally increased from West to East, with an area of minimum probability in the center (Figure 3). The highest probabilities were found along the coast. No false absences were found and only 16 observations were false positives.

2.3.3. Cross-validated logistic regression

The estimated regression coefficients are found in Table 5. As these are averages, the significance degree was not known but no coefficient was close to zero. The Spearman correlation between observed and predicted values was significant (0.92; p-value = 0.0013). The linear regression between observed and predicted values had an adjusted-R^2

of 0.84. Predicted values were slightly lower than observed values and not around one as expected (0.303; p-value < 0.001).

The AUC (0.72) indicated that the predictive power was satisfactory but the kappa (0.32) indicated a poor agreement. Probability of presence of disease decreased from South to North and from East to West. Most false positives were in the East (Figure 3).

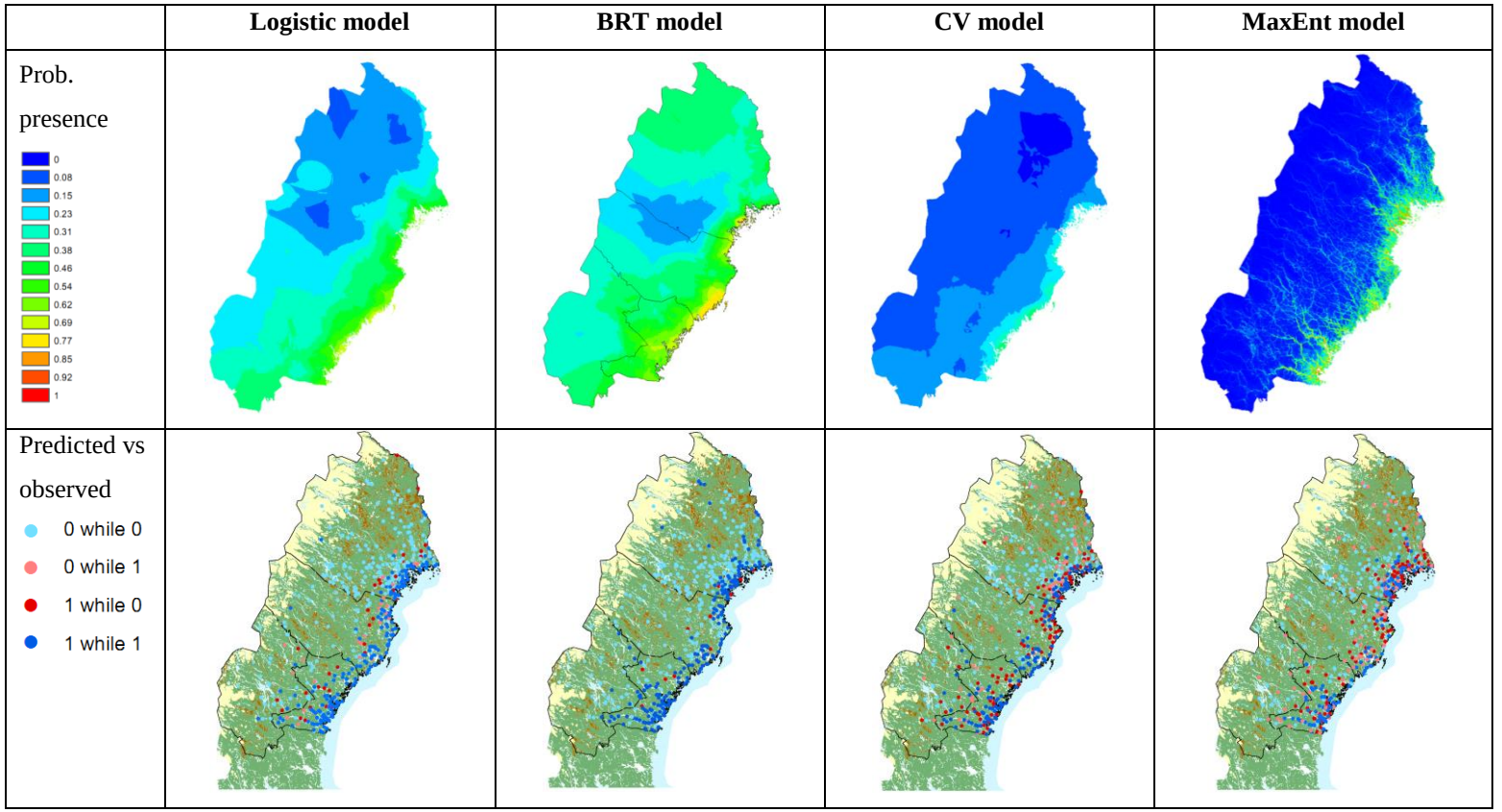


Figure 3: Comparison between logistic regression, boosted regression tree, cross-validated logistic regression and MaxEnt model

2.3.4. MaxEnt model

Results with and without the population density layer as sampling bias were similar. Jackknife analyses showed that elevation brought the highest gain when used in isolation from other variables. Roads decreased the most the gain when omitted from the model.

The probability of presence increased with the volume of spruce and decreased with the volume of pines. Approximately after 60 m^3/ha of pine, the probability of presence decreased, indicating that habitat is less favourable to bank voles. The two variables are probably complementary. When there are little pines, there are more spruces and vice-versa.

The AUC of MaxEnt model was good (0.91) but, when calculated on the same points than in logistic and CV model, it decreased to 0.66. Kappa was only calculated on the selected points and indicated poor agreement (0.19). The highest predicted probabilities were found near the coast, roads and water ways (Figure 3). There were many false positives in the East.

2.3.5. Comparison between models

When accounting for the confidence interval, the estimators of logistic regression and CV method were similar, except for contiguity and

population density (Table 5). AUC and AIC were the best for logistic regression.

The linear pattern which appeared in MaxEnt was related to including spatially detailed data on roads and water ways (Figure 3). As the other models were based on points and then interpolated, such a linear pattern cannot be represented, but may appear if probabilities were computed per pixel.

Based on the AUC, the logistic regression produced the best model. If a logistic regression is built with the same variables as MaxEnt (as calculated at the point and not in the surrounding buffer) and if the AUC of MaxEnt is calculated only on the original data points, both AUC were equal to 0.66 and ROC curve were similar, indicating similar predictive power.

The thresholds identified for classifying probabilities into presences and absences were 0.44 for logistic model, 0.42 for BRT, 0.17 for CV model and 0.49 for MaxEnt model. Except for BRT, all methods overestimated presence. Many false positives were near the sea coast.

2.3.6. Partial analyses with variables related to bank voles, virus and humans

For logistic regressions and BRT, AUC were the best for models with variables related to bank vole habitat, followed by models related to the virus and finally models related to humans (Table 6). Inversely,

for the MaxEnt models, the best model was built on variables related to humans.

Table 6: AUC of partial models based on variables related to each element

	Rodents	Virus	Humans
Logistic regression	0.732	0.695	0.684
Boosted regression trees	0.886	0.8244	0.801
MaxEnt	0.891	0.893	0.922

2.4. Discussion

The input data, the ease of use, the goodness-of-fit, the predictive power and the interpretation are here investigated for each method to help the modeller to make a choice of a particular modelling technique according to its purposes. Modeller should also pay attention to the different practices when using a particular method, e.g. selection of variables with a step procedure or inclusion of variables reflecting the surrounding environment. The advantages and disadvantages of each method, as applied with the accessible methodologies proposed here, are summarized in Table 7.

2.4.1. Input data

The good predictive powers of all modelling techniques indicate that human data are appropriate to model the disease in human. Moreover, in the case of a host with a wide ecological niche, like the bank vole,

models are less accurate (Tsoar et al. 2007) and it is better to use human case data rather than host for modelling. Human data are conversely not useful to model the disease in the host population as shown in a MaxEnt study on Juquitiba hantavirus in Brazil (Donalisio and Peterson 2011).

A major advantage of logistic regression, CV method and BRT is that they are working on point data, allowing the implementation of variables which reflect the surrounding environment such as the composition and configuration of the landscape. Zoonotic transmission indeed relates to factors extending beyond the place of record.

Logistic regression and BRT required absence data. Here, absences were identified from accurate data on dwellings but, these absence points may be unidentified cases or just an absence of human hantavirus transmission over the study period. Random locations would have been less appropriate as they would not consider human distribution. The CV method considered availability rather than absence, therefore avoiding the issue of unreported cases or absences related to the stochasticity of zoonotic disease transmission to humans. The issue of absences was also avoided for MaxEnt which only required presence records.

In MaxEnt, heavy constraints lied on the dependent variables (continuous raster maps of same resolution and geographical extent). Here, several variables at the landscape scale concerned landscape

structure. This could not be operationalized as continuous variables in a comprehensive and straightforward fashion. On one hand, measures concerning the landscape surrounding infection sites could no longer be used. Continuous rasters could be constructed to represent the landscape variables, but loss of information is inevitable. On the other hand, the spatial pattern of the input variables, which played an important role for the final model was preserved. In partial analyses, the best MaxEnt fit used variables related to humans. The human population in northern Sweden is highly structured along the sea coast and inland along roads and rivers. As the final model is calculated on each pixel, the linear pattern was more evident than in the other models.

2.4.2. Ease of use

Logistic regression is widely available in statistical softwares and is easily implemented. The R package “gbm” or “dismo” used for BRT includes a complete tutorial. The CV method required fastidious calculations. MaxEnt, a free software with a graphical interface, is user-friendly.

2.4.3. Goodness-of-fit and predictive power

In our study, the logistic regression gave the best results. It had the best AUC and so, the best predictive power, followed by BRT and

MaxEnt and, finally by CV model. It should be noted that BRT results were quite heavily influenced by the bag fraction.

When the same independent variables were used in logistic and MaxEnt models, AUC and Kappa were comparable. So, the stepwise procedure and the input variables based on the surrounding environment allowed a better fit and predictive power. These methodologies should always be considered.

The predicted probabilities of Figure 3 may first appear similar. All models predicted a higher risk near the coast even if a variation between regions can be observed according to the different modelling techniques. Depending of the study purposes, the differences between regions or localities may be of great interest and then, the choice of modelling techniques may involve more consequences.

False positives for a zoonotic disease can be interpreted as a poor prediction, a non-reported case, or the presence of the pathogen in the wild but its absence in humans. NE is generally under diagnosed, and many PUUV infected humans may go undetected. Indeed, up to seven in eight PUUV infected humans may go unrecognized with subclinical symptoms or symptoms mistaken (Ahlm et al. 1998). In maps of predicted versus observed (Figure 3), CV and MaxEnt had more incorrect predictions, indicating a poorer prediction comparing to logistic regression and BRT. Models only based on presence were most likely to overestimate presence. However, false positives may give indications on the potential distribution, while the others

approximate the realized distribution. The use of different sets of explanatory variables may also contribute to this, but tests using identical sets of predictors confirmed the results. All models overestimated presence near the coast, but BRT did the least.

Previous comparisons between presence/absence and presence-only methods have highlighted that logistic regression is more appropriate in some cases and MaxEnt are more appropriate in others. If absence data are available, logistic regression is better than MaxEnt to discriminate sites with high disease risk (La Manna et al. 2011). Also, penalized logistic regression, which avoids performance problems caused by overfitting, performs similarly to MaxEnt and has been found better than standard logistic regression (Gastón and García-Viñas 2011). Another study shows that MaxEnt is slightly better within the known distribution but logistic regression predicts better outside the data distribution (Cleve et al. 2011).

2.4.4. Interpretation

Logistic regression, BRT and CV models had higher flexibility for the inclusion of diverse variables. Variables with more straightforward biological interpretations and/or closer proxies could be added in the model. It could be argued that it is an attractive feature for explanatory models. In our case, landscape structure variables (e.g. relating to forest structure and arrangement with respect to human habitat) could be included. MaxEnt found powerful associations with altitude, a

variable of little biological significance that proxies several other biologically relevant variables such as temperature, snow cover or population density. The use of MaxEnt may thus be less recommended to build explanatory models.

A major advantage of MaxEnt was the production of a spatially continuous result, allowing finer detail and more visually pleasing output and avoiding the necessity to interpolate results spatially. However, this may come at the price of many false positive pixels. It may still be useful for identifying further study sites. As the interpolated surfaces of the other methods are also uncertain, major risk areas could be first outlined by MaxEnt, then, in these areas, other models could be fitted to gain more detail.

MaxEnt and BRT facilitated the identification and interpretation of nonlinear responses. Contributions and Jackknife analyses showed immediately interesting variables. Response curves showed the variation of probabilities in relation to the dependent variables. These curves brought a lot of information but could be quite complex to understand and a good understanding of the system was required. Moreover, in BRT, variables could be strongly correlated, meaning that curves were not reliable and that only the general trend should be considered. Also, results may be difficult to generalize: in MaxEnt, nonlinearity generated complex estimators and for the BRT, it was not possible to visualize the final tree.

In partial analyses, the importance of bank vole distribution was highlighted in logistic regression and BRT. These models allowed including landscape structure variables that describe the rodent habitat in more detail. In our study, MaxEnt model indicated the importance of human distribution because of its spatial pattern. Modelling the spatial distribution of human hantavirus infections requires thus both hazard and exposure. Similar modelling approaches have been used for other hantaviruses. A study in Argentine used reservoir host data and logistic regression to estimate risk areas for humans (Carbajo and Pardiñas 2007). Another study on Andes hantavirus in Argentina, comparing MaxEnt and logistic regression using rodent data and human infection data, found good predictive power for both methods in predicting rodent distribution, while MaxEnt performed less well on human data (Andreo et al. 2011).

Our four models were based on environmental conditions and tried to define the intersection of the spatial distributions of bank voles, humans and virus. Care is therefore needed when interpreting results, particularly differences between potential and realized distributions (Jimenez-Valverde et al. 2008) and, even if all favourable conditions are present, the disease/species is not necessarily found.

Table 7: General advantages and disadvantages of logistic regression, boosted regression trees, cross-validated logistic regression and Maxent model based on methodologies accessible to beginners

	Advantages	Disadvantages
Logistic regression	- Good goodness-of-fit and predictive power when using step procedure and including variables reflecting the surrounding environment - Inclusion of variables reflecting the surrounding environment	-Need of real absence points
BRT	- Account for nonlinearity of biological processes - Inclusion of variables reflecting the surrounding environment - Modelling of interactions	- Need of real absence points - Overfitting and difficulty to extrapolate
CV method	- Available sites instead of absence sites - Inclusion of variables reflecting the surrounding environment	- Fastidious calculations
Maxent model	- Ease of use - Spatially continuous results - Accounts for nonlinearity of biological processes	- Complex estimators, difficulty to extrapolate - Need of spatially continuous data - Limited by the coarsest resolution and the smallest extent of variables

2.4.5. Further proposals for modelling

An option could be to use first BRT or MaxEnt, in order to delineate areas of high probabilities. Variables could then be sliced according to their response curve into several variables or transformed into categorical variable. This way, nonlinear processes could be considered. If absences are available, they can be added in logistic regressions or, if not, in the CV method. Non-continuous landscape variables can then be added. The final purpose of the model, explicative or predictive, would also direct the choice, as would data availability and specificity of the system at hand.

2.4.6. Human hantavirus infections in Sweden

Disease cases were found at the intersection of the distribution of bank voles, humans and virus. Unlike Germany (see Chapter 4), the spatial pattern of hantavirus infections reflects a double gradient (from the coast to inner land and from South to North) that is followed by different explanatory variables and for which, the distance to the sea and the elevation were proxies. Distance to the sea was indeed significant in the logistic model and elevation was the variable that brought the highest gain in the MaxEnt model. Going from the mountainous areas in the North-West to the coast in the South-East, the climate becomes milder, pines are replaced by spruces, the soil contains more moisture and human density grows. Even if the correlations between our variables were not always strong, all

variables were interconnected. All of these variables are consequently important in the understanding of the disease in Sweden.

Other variables that appeared essential were the snow cover and the connectivity of the forest, represented by the distance to forests and the contiguity. Caution must nevertheless be taken as the connectivity index used here does not necessarily represent functional connectivity (Ewers and Didham 2006, Chetkiewicz et al. 2006).

2.5. Conclusion

Diverse methods were implemented based on human records, tested and performed (well or less well). The models differed in what they permit and offer, each of which may be more important depending on the study objectives. Every method has advantages and disadvantages. A solution to keep the most pros could be to combine the different methods as it will be done in Chapter 5 by using jointly BRTs and logistic regressions.

Chapter 3 Shaping zoonoses risk: the case of tick-borne encephalitis in Sweden [2]

In this chapter, the hazard and exposure concepts from risk assessment are applied in an innovative approach to understand zoonotic disease risk. Hazard is here related to the landscape ecology determining where the hosts, vectors and pathogens are and, exposure is defined as the attractiveness and accessibility to hazardous areas. Tick-borne encephalitis in Sweden is used as a case study. The use of BRT as an explanatory model is also investigated.

Three BRT are compared: a hazard, an exposure and a global model which combines the two approaches. The global model offers the best predictive power and the most accurate modelling. The highest probabilities are found in easy-to-reach places with high landscape diversity, holiday houses, waterbodies and, well-connected forests of oak, birch or pine, with open-area in their ecotones, a complex shape, numerous clear-cuts and, a variation in trees height.

While conditions for access and use of hazardous areas are quite specific to Scandinavia, this study offers promising perspectives to improve our understanding of the distribution of zoonotic and vector-borne diseases in diverse contexts.

---

[2] Adapted from: Zeimes, C. B., G. E. Olsson, M. Hjertqvist & S. O. Vanwambeke (2014) Shaping zoonosis risk: landscape ecology vs. landscape attractiveness for people, the case of tick-borne encephalitis in Sweden. *Parasites & Vectors,* 7**,** 1-10.

Chapter 3

3.1. Introduction

In this study, we attempt to distinguish hazard (pathogen circulation in the wild) and exposure (people entering infected landscape) by comparing the predictive power of three models that focus on different aspects of the landscape: a hazard model containing a set of variables found in the ecological literature, an exposure model containing a set of variables found in the touristic and public health literature and, a global model containing both sets of variables. TBE in Sweden is used as a case study. TBE has already been well studied under the hazard angle (e.g. (Randolph 2008)) but less under the angle of exposure (which is emphasized in (Stefanoff et al. 2012)).

TBE is of concern in Sweden, as the tick population has spread and the incidence of the disease has been increasing sharply over the past few years (Lundkvist et al. 2011). Two phenomena are currently observed in Sweden. On one hand, the range of human cases of TBE is expanding westward within the known tick range, and on the other hand, the expansion of ticks northward along the coast. While some common factors may be at play, the mechanisms behind each phenomenon have not been fully clarified. The expansion may result from climate changes, host populations dynamics and human behaviour changes (Jaenson et al. 2012b, Medlock et al. 2013). However, even in the well-established TBE endemic areas around Stockholm, the effect of these variables on spatial distribution of the disease is unclear. They will be reviewed and classified into hazard and exposure in the following section.

BRT are most often used as predictive or forecast model. As explained in Chapter 1, the predictive power is here used to compare different models together and to assess the relative quality of the model, assuming that a model with a good predictive power should have a good explanatory power (Shmueli 2010). In this Chapter, the use of BRT as a useful tool for explanation is also investigated.

3.2. Materials and methods

3.2.1. Materials

The study focused on Stockholm and the five neighbouring counties (Gävleborgs län, Dalarnas län, Uppsala län, Västmanlands län and Södermanlands län) (Figure 4). TBE is well established in that region of Sweden, where the disease has been recorded for the past century. Records of cases by nearest settlement of infection (SMI, Swedish Institute for Communicable Disease Control) were included for a ten-year study period (January 1998 to December 2007). Presence at any time during the study period was translated into a presence record, totalling 125 presence records. The other settlements extracted from the Lantmäteriet database (Swedish mapping, cadastral and land registration authority), constituted the 4297 absence records.

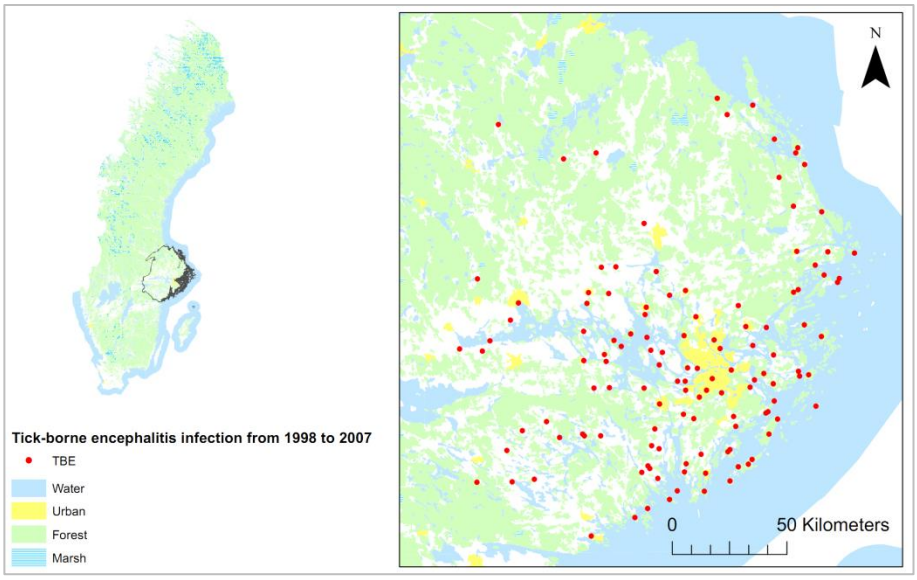


Figure 4: Human disease cases of tick-borne encephalitis in Sweden

Hypotheses were made based on literature, and candidate explanatory variables were allocated to hazard or exposure. However, some variables could not be allocated to a single group in an unequivocal way and were included in both. Variables either represented the surrounding environment, as calculated in a radius of two km around the point location (a two km buffer), or were calculated at the exact record location.

3.2.1.1. Variables describing hazard

Variables describing the hazard, i.e. places where infected hosts or vectors are found, were identified in the literature (Table 8). They are linked to the ecology of the hosts and vector. Some commonly used landscape metrics were also included.

a) Host species

In Sweden, roe deer are a major blood meal host for reproducing adult female ticks (Jaenson et al. 2012a, Jaenson et al. 2012b). However, other larger game species, e.g. red deer, fallow deer and wild boar, are available in large numbers and are also likely to be important hosts for ticks. Bag records of these game species were included (Dr Jonas Kindberg, Wildlife Monitoring Unit, Swedish Association for Hunting and Wildlife Management, personal communication). Data, available by center of hunting districts, were interpolated by Thiessen polygons, leading to a lower resolution compared to the other variables. The number of animals per hectare found at the location is used as a proxy for blood meal availability.

In Sweden, at the end of each hunting season, the bag records must be reported. Hunter often cluster in teams, hunting in local areas for which specific quotas are defined (Carlsson et al. 2010). The quotas are evaluated according to the hunting data from the previous year(s). Consequently, the data provided do not necessarily reflect the abundance of the species but can nonetheless be interpreted as proxies for variation in space of the game species.

b) Forest

Deciduous forests are a highly suitable habitat for ticks, as well as for some host mammals (Jaenson et al. 2009, Tack et al. 2012). The total proportion of forest, broadleaved forest, conifer and mixed forest

inside the two km buffer were included (100 m resolution, CORINE Land Cover, EEA). The average shape and proximity index of forest patches in the buffer were calculated. A patch with the most compact shape (i.e. the smallest edge to area ratio), in the case of raster data, a square, has a shape index of one. Increasing values indicate a more complex shape, and more contact between the patch and its surroundings. The proximity index of forest patches relates to the amount of forest within a specified radius around a patch, and indicates whether a patch is isolated or fragmented.

Also, as various tree species may impact tick habitat suitability differently, the mean volume of spruce, oak, birch and pine per hectare in the buffer were added (30 meter resolution, SLU Skogskarta, Swedish University of Agricultural Sciences).

Forest areas where tree height was lower than 50 centimeters were used as a proxy to represent clear-cuts (SLU Skogskarta). Intensive clear-cutting usually is non-valuable for wildlife but, in the study area, clear-cuts were small (mean area of 1682.68 m^2 and mean cross-section of 146.54 m) and offered a greater diversity of resources. The area of clear-cuts was divided by the area of forest in the buffer to represent the proportion of clear-cuts inside the forest. While clear-cuts may provide food for various host species, it does not provide as good shelter as forests.

c) Land cover

Forest ecotones, particularly where they connect to open areas, can be very suitable for ticks and hosts as these habitats offer a high diversity of resources (Lindström and Jaenson 2003). The main roe deer habitat is also deciduous or mixed forest with open areas (Jaenson et al. 2012a, Tack et al. 2012). The proportion of open areas (agricultural and transitional areas from CORINE) in ecotones of 150 meters around forests in the buffer was added to the hazard model.

The Shannon diversity index, representing the richness of the landscape in the buffer, was included (Shannon and Weaver 1949).

Since high humidity favours tick questing, moist areas are more suitable for ticks (Lindström and Jaenson 2003). The proportion of waterbodies in the buffer (CORINE) and the distance to the nearest water course (Lantmäteriet) were included as proxies for moister areas.

3.2.1.2. Variables describing exposure

Variables describing the exposure, the degree to which people enter infected landscape, were identified through accessibility and the scientific literature studying landscape attractiveness for touristic activities (Table 8).

a) Accessibility

A study of tourist preferences indicated that accessibility increases the touristic value of forest (Bostedt and Mattsson 1995). Indeed, in Sweden, there is a traditional right of public access to private land, e.g. to enter forests and to harvest resources such as mushrooms and berries (Bostedt and Mattsson 1995, Lindhjem 2007). Assuming that roads increase access, and that forests with roads are more likely to be entered by visitors, we included the length of roads in the buffer (Lantmäteriet) and the length of roads in forests in the buffer to describe accessibility.

Assuming that places with holiday cabins would relate to outdoor activities, the area occupied by holiday houses in the buffer was included (Statistics Sweden (Statistiska Centralbyrån)). In Sweden, 50% of holiday houses are within a radius of 32 kilometres from permanent homes (Müller 2006). The distance to Stockholm, from which many holiday cabin users originate, was included in the model (Lantmäteriet database), assuming that areas closer to Stockholm would be more frequently used for outdoor recreation. Population density was also included (2.5 arc-minutes resolution, Gridded Population of the World from Center for International Earth Science Information Network (CIESIN)). Occupational data were not available, preventing the identification of the population at highest risk (e.g. the forestry workers). This population is nevertheless probably vaccinated (vaccination data were neither available).

b) Scenic beauty

Landscape features documented to increase the perceived scenic beauty include water features (Nielsen et al. 2012) and broadleaved forests (Bostedt and Mattsson 1995). The distance to the nearest water course (Lantmäteriet database), proportion of waterbodies in the buffer (CORINE), and the proportion of forest and of broadleaved forest (CORINE) were used. In Finland, a preference for forest stands with a higher mean tree height and a skewed distribution of height has been highlighted (Silvennoinen et al. 2001). The mean tree height in the buffer was calculated and standard deviation of tree height was used as a proxy for the skewness (SLU Skogskarta). In Sweden, the touristic value of a forest increases with the number of clear-cuts and decreases with the size of the clear-cuts within a given area (Bostedt and Mattsson 1995). The proportion of clear-cuts in the forest (SLU Skogskarta) in the buffer was thus added to the exposure dataset.

Table 8: Variables selected in the hazard or exposure model

	Hazard	Exposure	Resolution	Units	Sources
Roe deer / red deer / fallow deer / wild boar	X		Low (interpolation based on hunting districts centers)	Number of animals per hectar	Dr. Jonas Kindberg
Proportion of forest in the buffer	X	X	100m	Percentage	CORINE
Proportion of broadleaved forest in the buffer	X	X	100m	Percentage	CORINE
Proportion of coniferous forest in the buffer	X		100m	Percentage	CORINE
Proportion of mixed forest in the buffer	X		100m	Percentage	CORINE
Shape index of forest in the buffer	X		100m	None	CORINE
Mean proximity index for forest patches in the buffer	X		100m	None	CORINE
Mean volume of spruce / pine / birch / oak in the buffer	X		30m	m^3/ha	SLU skogskarta
Proportion of clear-cuts (tree height < 50cm) in the forest in the buffer	X	X	30m	Percentage	SLU skogskarta
Proportion of waterbodies in the buffer	X	X	100m	Percentage	CORINE

Distance to the nearest water course	X	X	High (shapefile)	m	Lantmäteriet
Proportion of open areas in ecotone of 150m around forest in the buffer	X		100m	Percentage	CORINE
Shannon diversity index in the buffer	X		100m	None	CORINE
Length of roads in the buffer		X	High (shapefile)	m	Lantmäteriet
Length of roads in forest in the buffer		X	High (shapefile)	m	Lantmäteriet
Distance to Stockholm		X	High (shapefile)	m	Lantmäteriet
Proportion of area occupied by holiday houses in the buffer		X	High (shapefile)	Percentage	Statistiska Centralbyrån
Mean population density		X	2.5 arc-minutes	Person/km²	Gridded Population of the World
Distance to the sea		X	High (shapefile)	m	Lantmäteriet
Standard deviation of tree height in the buffer		X	30m	m	SLU skogskarta
Mean height tree		X	30m	m	SLU skogskarta

3.2.2. Methods

3.2.2.1. Principal component analyses

The potential explanatory variables outlined above are numerous, mostly proxies, and sometimes redundant. Therefore, principal component analyses (PCA) were used to identify sub-groups of similar variables ("Rcmdr" package and plugin "FactoMineR" in R 2.12.0). The factorial coordinates were used as new variables. Different groups of variables were tested: accessibility, forest beauty and area beauty related variables for the exposure and, land-use, forest structure, tree species, soils and wildlife species related variables for the hazard. Two groups of variables allowed reducing the number of variable to two mains principal components: one summarizing the variables on wildlife species (wild boar, red deer, fallow deer and roe deer) and another, accessibility variables (population density, length of roads, distance to Stockholm and length of roads in forest).

3.2.2.2. Boosted regression trees

The multivariate models were built using BRT ("gbm" package in R) (Elith et al. 2008). Fitted probabilities of response curve graphs were interpreted here as a relative probability of disease presence at various levels of the predictor variable.

Three models were built containing respectively the hazard variables, the exposure variables, and both hazard and exposure in a global

model. Some variables were included in both hazard and exposure models as they may relate to either aspect (Table 8).

3.2.2.3. Spatial autocorrelation

Spatial autocorrelation may result from endogenous or exogenous factors (Dormann 2007, Bonada et al. 2012). Endogenous factors are linked to the biology of the species and are often related to contagion processes. For instance, cherries fallen on the ground around a cherry tree may lead to a cluster of cherry trees. Exogenous factors mainly result from environmental conditions. For instance, a bird that ate cherries may disseminate the seeds, also leading to a cluster of cherry trees. In the case of exogenous factors, the inclusion of the factor (e.g. the bird) remove or decrease the bias due to spatial autocorrelation (Dormann 2007). To correct spatial autocorrelation resulting from endogenous factors, autoregressive terms may be added.

TBE is still spreading, indicating that endogenous process may be at play. The proportion of infected places within a radius of 20 km was thus added to each model to account for potential spatial structure in the distribution of TBE cases.

3.2.2.4. Measures of the predictive power

a) Internal validation of the predictive power

As BRT builds the trees on random subsamples, each model (hazard, exposure and global) was run 25 times. AUC ("PresenceAbsence" package in R) were compared using a Student t-test. False presences and absences, using the sensitivity equals the specificity as probability threshold, were mapped.

b) External validation of the predictive power

External AUC for each model were calculated from cross-validation on 10 subsets. This was run 25 times and compared with a Student t-test.

The three models were computed on the southern half of the data and run on the northern half for which the AUC were calculated. This was done to evaluate the predictive power of BRT on small areas and to assess the overfitting.

Moreover, TBE records from 2011 were also used to assess the predictive power of our models. Continuous predicted probability maps were created for each model by kriging, on which TBE presences records in 2011 were overlaid.

The presences in 2011 were completed with absences (settlements with no presence records between 1998-2007 and in 2011). Then, predicted probabilities were calculated for this new dataset. The means of predicted probabilities located at presences points were compared to absences points by a Welch test.

3.2.2.5. Use of BRT as explanatory model

The global model was run five times and, in the list of the top five variables with the highest relative importance, the ones that appeared in each of the five models were removed. This was repeated several times in order to remove the 10 most important variables in common. The variables relative importances of the reduced model were compared to the relative importances of the global model.

3.3. Results

3.3.1. Principal component analyses

Only two groups of variables, the wildlife species and the accessibility related variables, had the two first principal components explaining more than 60%. The two first components of the PCA on the data on wildlife species explained 71.76% of the variance. The first component (variable PC1: Wildlife) was positively correlated with wild boars (correlation of 0.85), red deer (0.84) and fallow deer (0.64).

The second component was only positively correlated with roe deer (0.99), which was subsequently kept as an individual variable.

The first two components of the PCA on accessibility variables explained 66.36% of the variance. The first component (variable PC1: Accessibility) was positively correlated with the human population density (correlation of 0.76), the length of roads (0.65) and negatively correlated with the distance to Stockholm (-0.75). The second dimension was positively correlated with the length of roads in the forest (0.92) and was also kept as an individual variable.

3.3.2. Boosted regression trees

The variables with the highest relative importance (variables forming the first 50% of summed relative importance) in the hazard model were the number of TBE cases within 20 km (relative importance of 23.23%), volume of spruce (10.71%), distance to a water course (7.30%), total proportion of forest (5.40%) and proportion of coniferous forest (5.33%) (Table 9).

In the exposure model, the variables with the highest relative importance were the number of TBE cases within 20 km (relative importance of 25.42%), length of roads in forests (17.67%) and distance to a water course (7.79%) (Table 9).

Table 9: Relative importance of variables introduced in the hazard and in the exposure boosted regression trees

Hazard model		Exposure model	
Variable	**Relative importance (%)**	**Variable**	**Relative importance (%)**
Infections in 20km	23.23	Infection in 20km	25.42
Volume of spruce	10.71	Length of roads in forest	17.67
Distance to water course	7.30	Distance to water course	7.79
Volume of oak	6.33	Proportion of forest	7.75
Proportion of forest	5.40	Mean height of trees	7.09
Proportion of coniferous	5.33	PC1: Accessibility	6.78
Proportion of clear-cuts	5.28	Proportion of holiday houses	6.18
Volume of birch	5.24	Distance to the sea	5.49
Forest shape index	5.04	Standard deviation of height of trees	5.19
Shannon diversity index	4.83	Proportion of clear-cuts	5.19
Volume of pine	4.47	Proportion of broadleaved forest	3.74
Proportion of mixed forest	4.26	Proportion of waterbodies	1.72
Forest proximity index	3.60		
Open areas in ecotones	3.05		
Proportion of broadleaved forest	2.20		
PC1: Wildlife	1.31		

Proportion of waterbodies	1.31	
Roe deer	1.05	

The trends in the response curves of probability of finding TBE were similar in the three models. Response curves graphs of the global model are represented in Figure 5. Variables which showed a global positive trend are: infections in 20km (relative importance of 17.94%); roads in forests (11.56%), holiday houses (4.94%), PC1: Accessibility (4.16%), oak (3.76%), birch (3.50%), Shannon index (3.19%), forest shape index (3.17%), mixed forest (3.00%), clear-cuts (2.88%), standard deviation of trees height (2.71%), forest proximity index (2.07%), broadleaved forest (1.52%) and, waterbodies (0.97%). Variables which showed a global negative trend are: spruce (8.10%), coniferous (3.84%), mean trees height (2.78%) and roe deer (1.35%). Variables which showed an important decrease followed by an increase are: distance to water course (4.55%), distance to the sea (4.15%), forest (3.70%), open area in ecotone (2.31%) and, PC1: Wildlife (0.74%). Pine (3.12%) showed a first peak around 15m^3/ha followed by an increase around 70m^3/ha.

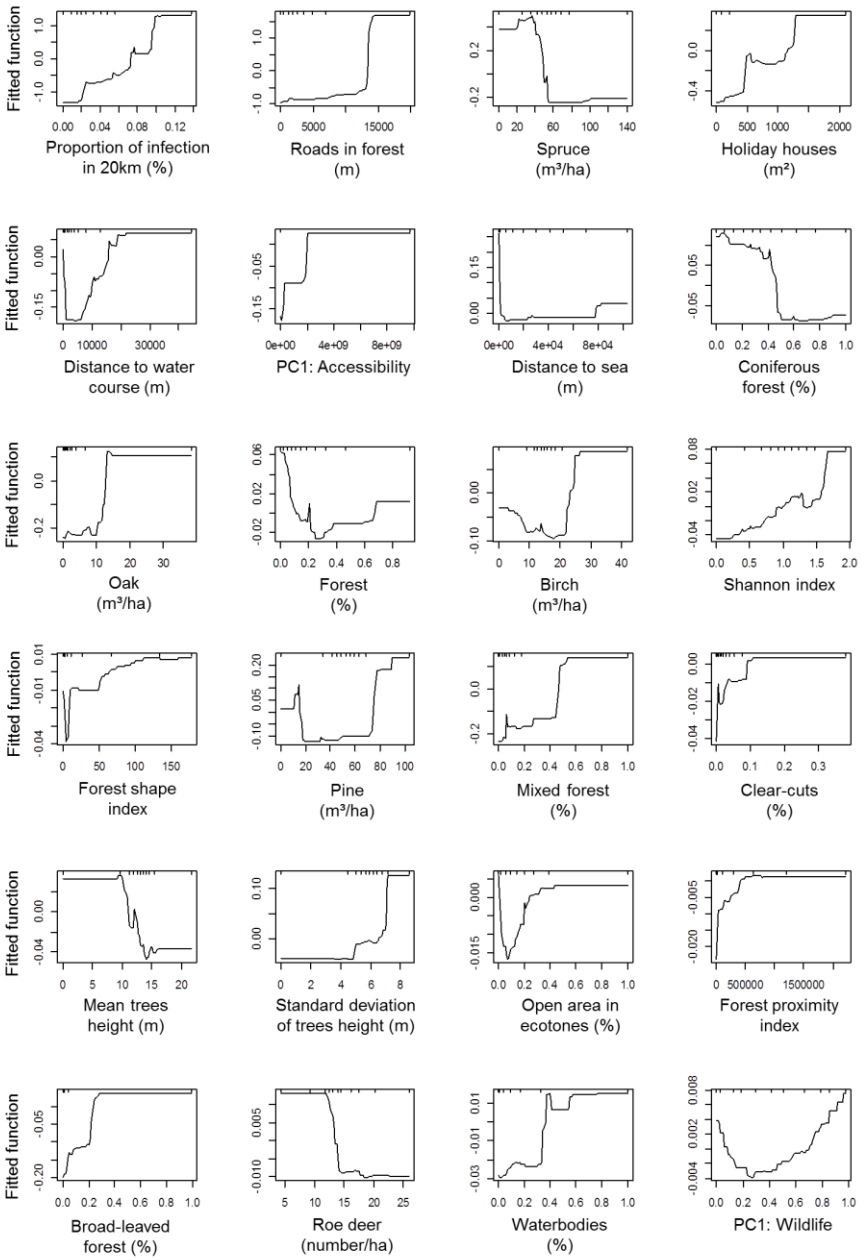


Figure 5: Graphs of each variable according to the fitted function of the global model

3.3.3. Measures of the predictive power

The mean AUC were 0.92 for the hazard and exposure models and 0.93 for the global model. AUC were lower for the cross-validated data with 0.74 for the hazard and exposure models and 0.75 for the global model. The mean AUC of global models was significantly higher (p-value <0.001) than the mean AUC of hazard models and exposure models.

For the global model, AUC resulting from a model computed on the southern half of the data (reduced sample size) was 0.92. The model applied on the northern half of the data had an AUC of 0.73, indicating still a good predictive power and low overfitting.

There were few false absences: 18 for the hazard model, 30 for the exposure model and 18 for the global model (Figure 6). There were more false presences: 681 for the hazard model, 482 for the exposure model and 593 for the global model (Figure 6).

The false presences of the hazard model were in areas within the TBE focus while false presences of exposure model were distributed in areas where disease cases have not yet been recorded (Figure 6). The global model had the best visual match between areas with high interpolated probabilities and high frequency of presence in 2011. The hazard and exposure models both appeared to contribute to the distribution of high probability areas observed in the global model map (Figure 6).

The mean probability predicted by the global model on presences in 2011 ($44.05*10^{-3}$) was significantly different (p-value<0.001) from the mean probability of absence ($27.51*10^{-3}$). Similar results were observed for probabilities extracted from the hazard and exposure models.

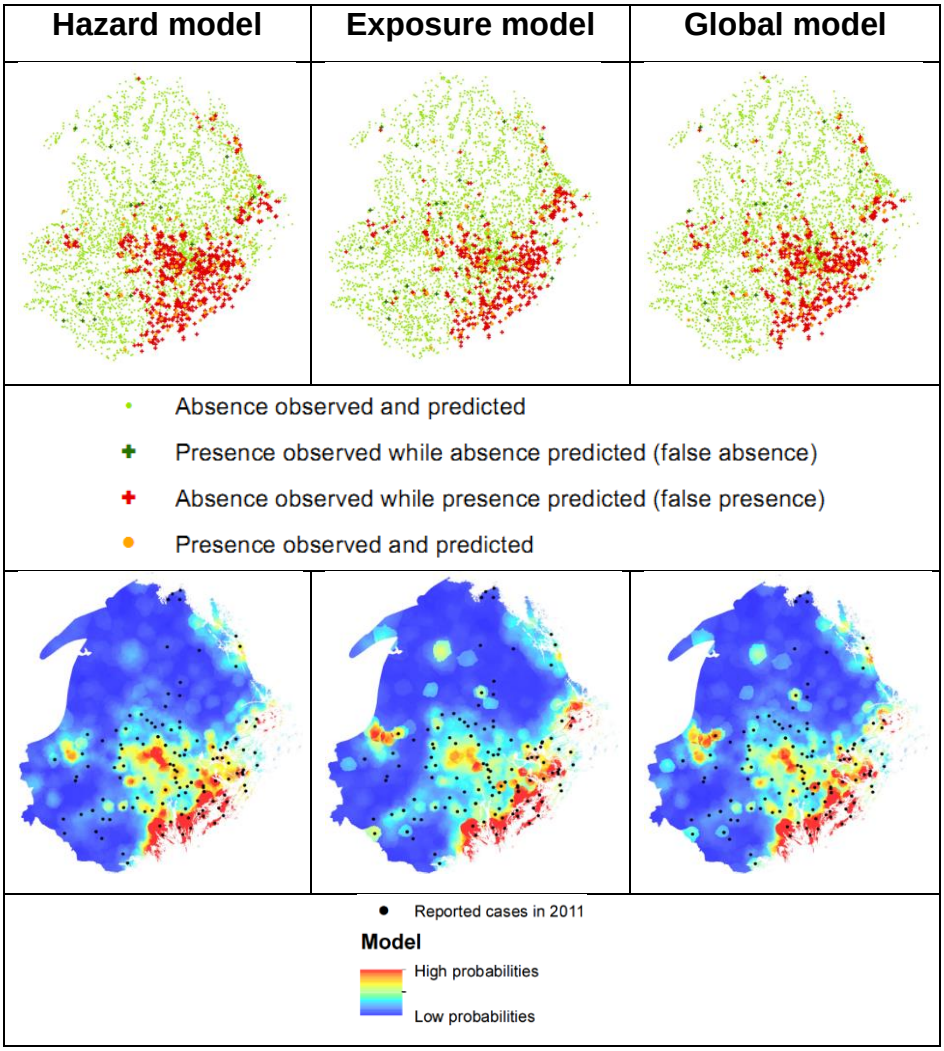


Figure 6: Resulting maps of the hazard, exposure and global models (based on 1998-2007 TBE records) and TBE records in 2011

3.3.4. Use of BRT as explanatory model

The roads in forests, spruce, holiday houses, distance to water course, PC1: Accessibility, coniferous forest, distance to the sea, birch, mean trees height and, standard deviation of trees height were removed. The reduced model had an AUC of 0.88 and a cross-validated AUC of 0.70, indicating a predictive power that is still good enough.

In the reduced model, the five variables with the most importance were proportion of infection, oak, clear-cuts, forest, pine and forest shape index (Table 10). The relative importances increased compared to the relative importance of the global model.

Table 10: Relative importances of variables included in the reduced and global model

Variable	Reduced model: relative importance (%)	Global model: relative importance (%)
Proportion of infection in 20km	34.94	17.94
Oak	8.57	3.76
Clear-cuts	8.06	2.88
Forest	7.85	3.70
Pine	6.91	3.12
Forest shape index	6.83	3.17
Shannon index	6.59	3.19
Forest proximity index	5.70	2.07
Mixed forest	4.41	3.00
Open area in ecotones	3.18	2.31
Broadleaved forest	2.76	1.52
PC1: Wildlife	1.51	0.74
Waterbodies	1.36	0.97
Roe deer	1.33	1.35

Graphs of each variable according to the fitted function of the reduced model were similar to the graphs produced for the global model (Figure 7).

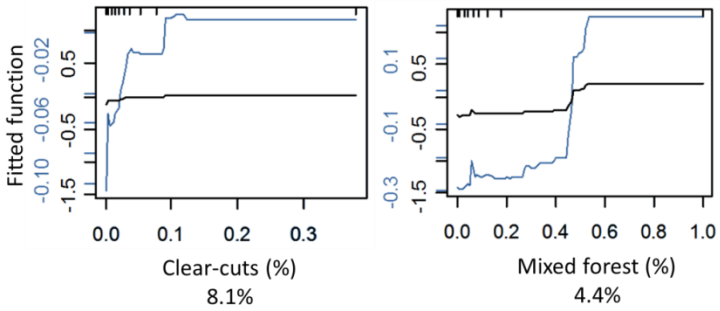


Figure 7: Graphs of clear-cuts and mixed forest according to the fitted function of the reduced model. Blue scale: adjusted scale. Dark scale: common scale at all variables graphs.

3.4. Discussion

3.4.1. Comparison of hazard and exposure

All three models, focusing on hazard, exposure and all factors, respectively, reached a good fit and a reasonable ability to predict hot spots for an independent year, i.e. 2011. However, the global model was clearly the most exhaustive in indicating areas of higher probabilities. Both the aspects of hazard and exposure therefore deserve consideration when examining risk and its spatial distribution. The two variables with the highest relative importance in the global model, other than the variable describing spatial structure, were the roads in forests, positively related to the exposure, and the volume of spruce, negatively related to the hazard, underlining the importance of

accounting for both aspects of risk. This makes perfect sense when approaching the question of the spatial distribution of a zoonotic disease using human case records: human land use is spatially heterogeneous. Beyond the need to account for all factors explaining the spatial distribution, factors related to exposure may offer important keys for the understanding of zoonotic disease emergence and human risk. In this study, the spatial distribution of TBE cases in Sweden looked beyond hazard-related factors and classic epidemiologic factors such as occupation to include variables depicting specifically where people at risk are most likely to enter infested areas.

False presences identified for the various models did not follow the same spatial pattern (Figure 6). These false presences could be: model errors; locations suitable for transmission but where the pathogen or susceptible humans are absent; areas where the pathogen is circulating but not transmitted to humans and; areas where the disease is found but not recorded (non-identified, non-reported or asymptomatic cases). Considering false presences as points were TBE may appear in the near future, the exposure model seems suitable for predicting the disease in new areas. Inversely, looking at the probability maps, the hazard model seems to show a better prediction in areas of spatially concentrated TBE infectious areas and so a better prediction of intensification of the disease. There are still few points outside the high probability areas, indicating that some variables may be missing in the models.

Challenges related to this hazard/exposure approach relate essentially to the interpretation of variables as influencing hazard or exposure. Some variables may be related to hazard or exposure in an unambiguous way, but several may be proxy for both vector (or host) habitat and landscape attractiveness for human. For example, while tree species would presumably relate to hazard, a study conducted in Finland indicates that touristic preferences increase with the volume of pines and birches (Silvennoinen et al. 2001). Scale may influence the interpretation of variables as related to hazard or exposure. For example, the distance to the sea, here used in the exposure model, may also influence, at a broader scale, the length of the vegetation season and the suitability for vectors and hosts through its buffering effect on temperature (Jaenson et al. 2009). Careful consideration of the interpretation and scale of variables included in risk, hazard or exposure models is therefore necessary.

Both hazard and exposure variables are needed for a better understanding of the spatial distribution of vector-borne diseases, but exposure variables may be specific to regions, just as some epidemiological risk factors can be culturally driven. Sweden, for instance, has a particular public right of freedom to roam ("Allemansrätten"). Still, accessibility is not evenly developed everywhere, and other factors may influence accessibility or attractiveness, such as land ownership (Vanwambeke et al. 2010).

3.4.2. Variables influencing the distribution of TBE in Sweden

This study highlights the main spatial variables influencing the distribution of TBE in a highly endemic region of Sweden. Forests with the highest probabilities of presence were well-connected oak, birch and pine forests, with a complex shape, numerous clear-cuts and a tree height variation of at least five meters. Landscapes with the highest probabilities were easy to reach, with high landscape diversity (Shannon index), holiday houses, waterbodies and broadleaf or mixed forest with open area in their ecotones. Fitted function curves mostly follow our preliminary assumptions based on the literature or field experience. Therefore, spruce forests are less favourable than pine forests, probably because they have less undergrowth. Also, the high probability in very close proximity to the sea is related to the presence of houses near the shore. The results for the distance to water course raise new questions as it is a less prominent feature of the landscape than waterbodies.

Few interactions were identified. The most important interaction was between area of holiday houses, a proxy for attractiveness, and places with more than 10 km of roads in forests, a proxy for accessibility. Furthermore, correlation between variables makes some probability distributions difficult to interpret. For instance, tree height, which is positively correlated to the proportion of coniferous, is negatively correlated to broadleaved forest. Decreasing probabilities with the tree height may thus relate to the decreasing probabilities with coniferous.

Also, low pine volume may imply the presence of larger volumes of deciduous trees and explain the peak around 17 m^3/ha.

The PCA showed that wild boars, red deer and fallow deer are not found in the same places as roe deer. Deer and wild boars (and, maybe, more specifically, young wild boars) most likely constitute important blood meal sources for adult female ticks before egg laying. A negative response of TBE with roe deer was previously highlighted in Sweden (Jaenson et al. 2012a) and, at a local scale, in Italy and Slovakia, with the increase of co-feeding ticks on rodent when deer density is decreasing (Cagnacci et al. 2012). Both studies hypothesized a dilution effect due to a high density of deer (incompetent hosts) diverting the questing ticks from rodents (competent hosts). However, a mathematical model estimating the threshold for tick-borne disease persistence reveals that, in the case of non-viraemic transmission, the dilution effect is less relevant (Rosà et al. 2003). Because of our PCA results, we argue that the negative effect between roe deer and TBE may result not from roe deer specifically, but from unsuitability for any aspect of the transmission cycle. Indeed, a study in the Czech Republic revealed a positive and significant association between TBE and wild boar but no significant association with roe deer (Kriz et al. 2014). Further investigation on the role of wildlife in feeding ticks and hosting the TBE virus would be useful. These results highlight the need for a better understanding of the TBE transmission system and the mechanisms underlying statistical relationship. Only in this way could such results be meaningful for risk prediction and public health.

3.4.3. Use of BRT as explanatory model

We used BRTs as explanatory models. Response curves allowed us to interpret phenomena and to answer preliminary hypothesis on how the variables behave. Nonetheless, the interpretation of these response curves can be risky for two reasons. First, the explanatory power of the model is not evaluated and second, the relative importance of some variables and their effects represented at a common scale may seem weak.

For BRT, the explanatory power that usually provides information on the strength of the relationship between a set of predictors and a response cannot be evaluated. Instead of the explanatory power, the predictive power is used in order to see if the model corresponds to the reality. The BRTs computed here had high predictive power, even if only 125 presences were recorded. When calculating on a smaller study areas, the southern half, the predictive power remained good. Also, the known issue of overfitting in BRT did not seem to have here a great impact. Indeed, the AUC from cross-validation and from the application of the model as calculated on the southern half to the northern half were still good. These very good predictive powers allow thinking that the models reflect well the complex reality. The strength of the relationship not being evaluated, the response curve cannot be used to truly quantify the level of probabilities (y-axis). Only the variables trends should be considered and marked thresholds may be identified (x-axis). The disadvantage that the strength of the relationship cannot be assessed is counterbalanced by the high

predictive power of BRT, indicating that the reality is approached in its completeness.

The nonlinear response curves provided by BRT models have great explanatory potential. Nonetheless, it is hard to trust their effect when represented at a commons scale (Figure 7) and when looking at their relative importance. It may not matter in our case, as the focus is to interpret how the variables behave and not how important they are to make predictions. The relative importance of variables varies greatly depending of the amount of other variables included in the model. When removing the 10 most important variables, the relative importance of the other variables has been substantially raised (Table 10). Between two correlated variables, one may have a much higher relative importance, decreasing the relative importance of the other variable. By removing the first variable, the relative importance of the second will probably increase more than what it would have risen so that the sum of relative importances reaches 100%. Also, when looking at the response curves, a common scale does not necessarily reflect the relative importance. Indeed, the curve of the mixed forests (relative importance of 4.41%) varies more importantly than the curve of the clear-cuts (relative importance of 8.06%).

In conclusion, BRTs are useful to explain the general trends of variables but not to quantify their effects. The response curves may also bring information about potential thresholds from which a change of trend can be expected but not on how these changes would be important.

3.5. Conclusion

Our study of the distribution of human cases of TBE in Sweden indicates that separating and accounting specifically for hazard and exposure in distribution models holds great potential for the understanding and the mapping of zoonotic disease spatial pattern and emergence. Exposure variables were extracted from standard GIS data bases following a similar strategy as is classically done for studies focusing on hazard.

TBE is emerging in different places in Europe and understanding this pattern is essential to help public health. Randolph compares human cases to “the tip of the iceberg” that emerges from the undetected enzootic cycles below the surface (Randolph 2008). As ecological processes driving the distribution of TBE are not yet completely described, it is of great value to be able to track the sources of human TBE back to infection sites, trying to unravel the role of local wildlife on the persistence and circulation of TBE. Accounting for exposure may also contribute to this by allowing more specific interpretation of any variable in the model.

In conclusion, linking ecology and public health is highly recommended. While the conditions for access and use of hazardous areas highlighted in this study may be specific to Scandinavia, this unified method offers promising perspectives to further understand the distribution of various zoonotic and vector-borne diseases in diverse contexts by the explicit inclusion of exposure-related variables.

Chapter 4 Untangle the effect of variables on hazard and exposure for hantavirus in Germany[3]

A step forward is done with the framework explained in Chapter 3. The effects of environmental factors on hazard and exposure for hantavirus in Germany are unravelled using two models: one on human cases and the other on bank voles. The spatial distribution of human disease cases and of the host, the bank vole, are for the first time jointly studied at the country scale with a common set of environmental variables related to climate, soil, land use, altitude and human density and activities.

Hazard was positively impacted by mild summers and winters with less snow cover, precipitations, wind speed, soil humidity, mixed and broadleaved forest, forest contiguity, and negatively impacted by air humidity, numerous warm days, built-up areas and built-up areas in forest ecotones. Exposure, the existence of human contact with an infectious environment, was positively related to mild winters, warm summers, less snow cover, extended growing season, mixed and broadleaved forest, population density and working activities, and negatively related to wind speed.

[3] In prep.: Zeimes CB, Reil D, Jacob J and, Vanwambeke SO. "The effect of environmental variables on zoonotic disease hazard and exposure: spatial modelling of hantavirus human infections and bank vole presence".

Chapter 4

4.1. Introduction

Untangling the effect of hazard- or exposure- related factors based solely on human disease case records can be challenging, especially when this effect is examined with proxy variables, as often done in environmental studies. The effect of environmental factors related to the climate, soil, land use, altitude and human societies on hazard and exposure are here differentiated using both human case and host data. This was applied to PUUV in Germany for which human PUUV cases were reported by county and bank voles were sampled across the territory. Bank voles should be related to specific environments described as hazardous areas while human cases should be related to the exposure by risk activities in these hazardous environments.

In Germany, since NE became notifiable in 2001, many outbreaks have been reported in different areas (Heyman et al. 2012). No study has so far investigated spatially the environmental factors of the distribution of hantavirus at the country scale. Local studies identified seed plants and especially beech forests and the production of beechnuts as promoters for increasing human PUUV infections in Baden-Württemberg and western Thuringia (Piechotowski et al. 2008, Schwarz et al. 2009, Boone et al. 2012, Faber et al. 2013). Rural dwellers are more at risk than urban dwellers (Faber et al. 2010) but humans have been infected also during recreational activities in a forested city park in Cologne (Essbauer et al. 2007). Climatic conditions associated with high human PUUV incidences are mild winters and springs (Schwarz et al. 2009). The bank vole density is

indirectly linked to the weather conditions two years before, reflecting the importance of tree mast prior to a population increase (Imholt et al. 2014).

In a previous study, investigating the effect of environmental factors on host abundance, PUUV prevalence among host population and human NE cases in Belgium (Linard et al. 2007b), it was concluded that NE cases are better predicted by variables promoting the indirect transmission path (virus survival outside the host) than by variables related to the host abundance. In our study, we move a step forward by looking at the effect of each factor on hazard and exposure. Human cases reflect both hazard and exposure but, when the variable effect is opposite between the bank vole and the human PUUV case model, the variable is related to exposure. This helps to better approach the disease transmission system.

4.2. Materials and methods

4.2.1. Materials

4.2.1.1. Human hantavirus cases

Clinical PUUV infections in humans were reported per counties from 2001 to 2013 (https://survstat.rki.de/, data status: 16.10.2013). If there was at least one case recorded, the county was considered PUUV positive (presence) (Figure 8). If there was no record, the county was considered PUUV negative (absence). In Germany, there are 107

small “urban” counties covering 16 030km² and 295 large “rural” counties covering 341 555km². At first, all counties were considered, and then rural counties were studied separately.

4.2.1.2. Bank voles abundance

Information on bank vole abundance in 2001-2013 was available from several locations in eight Federal States of Germany: Lower Saxony for 2001-2011, Hesse for 2006-2010, Saxony-Anhalt for 2001-2006 (Northwest German Forest Research Station and the Lower Saxony State Office for Consumer Protection and Food Safety); Brandenburg for 2001-2011 (Brandenburg Forestry State Agency); Mecklenburg-Western Pomerania for 2001-2006 and 2011 (Mecklenburg-Western Pomerania National Forest); Thuringia for 2001-2010 (ThüringenForst), and trappings done in North Rhine-Westphalia for 2010-2013 (Julius Kühn-Institute, Institute for Plant Protection in Horticulture and Forests, Vertebrate Research). Abundance data were collected as the number of individuals per 100 trap nights estimated from a standard snap trapping method of forestry authorities in Germany. As some trapping locations were in close proximity to each other, coordinate points were aggregated on a grid of 6km² resolution. If the bank vole abundance in a cell was > 0, the center of the cell was considered as presence for bank vole. If no bank vole was trapped, it was considered as an absence point (Figure 8).

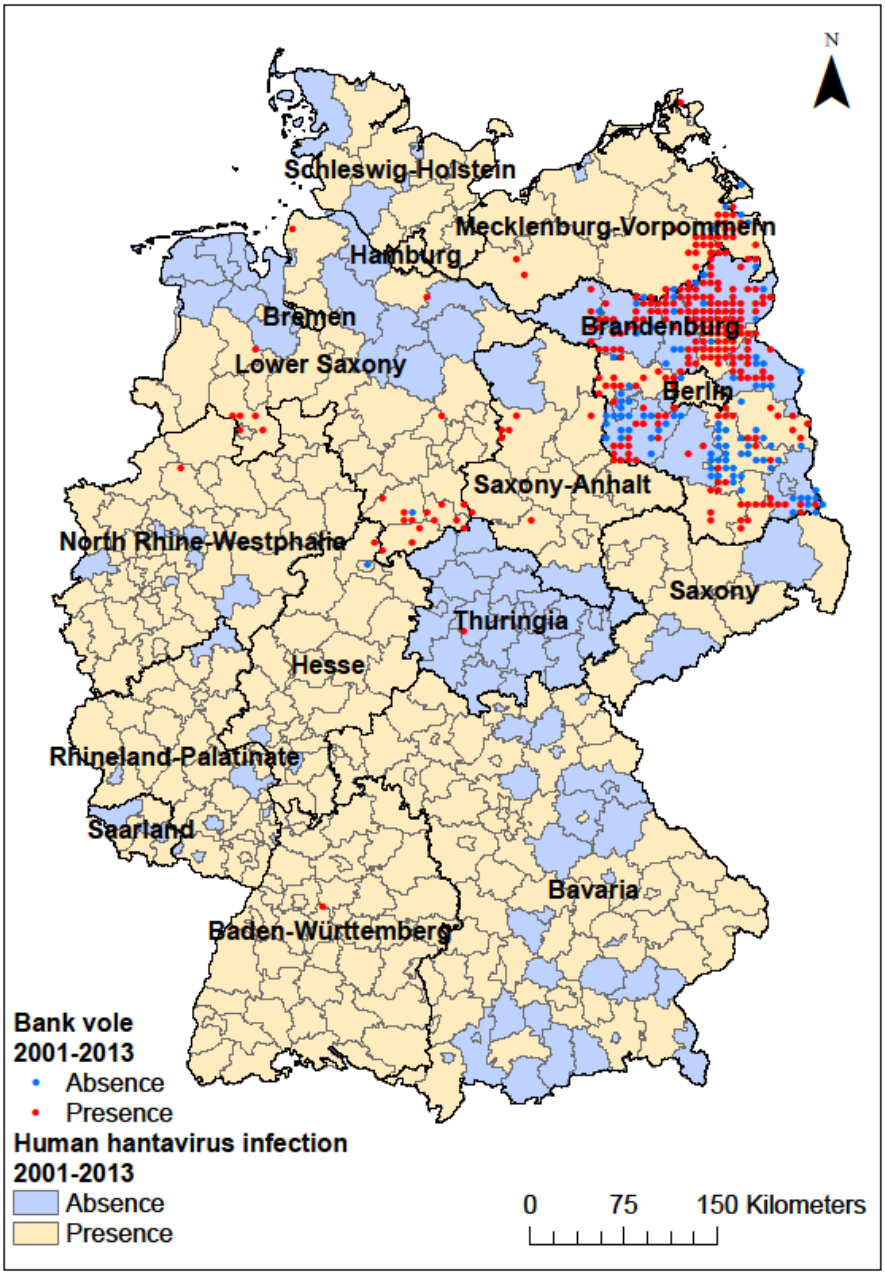


Figure 8: Presence and absence of human PUUV cases and bank voles between 2001-2013 in Germany

4.2.1.3. Explanatory variables

Variables were calculated and averaged at the county level for the human PUUV cases model and averaged in a circle of 3km radius for the bank vole model.

Climatic variables were extrapolated by kriging from meteorological stations (ECA&D, European Climate Assessment & Dataset). The variables were calculated by year and averaged over 2001 and 2013. They included the sum of precipitation, the daily mean of relative humidity, the maximum number of consecutive dry days (when the sum of precipitations were <1 mm), the number of maximum consecutive warm days (with a temperature >25°C), the daily maximum temperature in summer (June, July, and August), the daily minimum temperature in winter (December, January, and February), the average of daily snow depth, the average daily wind speed, and the growing season length.

The average monthly soil water index (TU-WIEN) and soil texture from coarse to thin (ESDB, European Soil Database distribution version 2.0, European Commission and the European Soil Bureau Network, EUR 19945 EN, 2004, (Panagos et al. 2012)) represented soil humidity. Soils with increasing texture and finer particles can retain more water, which facilitates the ex-vivo survival of the virus.

The percentage of CORINE land use classes were calculated in the counties (human PUUV cases model) and in the 3km buffers (bank

vole model) for coniferous forest, deciduous forest, mixed forest, agricultural land, built-up areas, and waterbody surfaces. The proportion of built-up areas in an ecotone of 250m around forests was also calculated. Forest contiguity and the minimum Euclidian distance between two forest patches were included to represent forest connectivity (FRAGSTATS, version 4) as bank voles are found in connected forest patches (Guivier et al. 2011). The mean shape index of forest patches was added to represent the complexity of the shape of the forest.

Further parameters included were the elevation (EEA, European Environment Agency) and mean slope ("Slope" from toolbox "Surface", ArcGIS).

Finally, the human population in 2012 (Statistisches Bundesamt) and the number of agriculture, forestry and fishery workers per 1000 inhabitants (Statistisches Bundesamt) were added to the human PUUV cases model.

Table 11: List of variables included in the models with their units, sources and temporal and spatial resolution

	Variables	Units	Sources	Temporal resolution	Spatial resolution
Climate	Precipitation	mm	ECA&D	Mean on 2001-2013	Interpolation by kriging based on weather stations (raster of 1km)
	Relative humidity	%			
	Number of dry days	Days			
	Number of warm days	Days			
	Maximum temperature in summer	° Celsius			
	Minimum temperature in winter	° Celsius			
	Snow depth	cm			
	Wind speed	m/s			
	Growing season length	Days			
Soil	Soil Water Index	Relative units	TU-WIEN	Mean on 2007-2010	25km
	Soil texture	Coarse to fine	ESDB	2004	1km

Land use	Proportion of coniferous, mixed and broadleaved forest Proportion of built-up areas Proportion of agricultural areas Proportion of waterbodies Proportion of built-up areas in 250m ecotone around forests	%	Corine	2006	100m
	Contiguity of forest	Relative unit			
	Minimum distance between forest patches	M			
	Forest shape index	Relative unit			
Altitude	Elevation	Meters	EEA	2004	1km
	Slope	Degree	EEA	2004	1km
Human	Population	Person	Statistisches Bundesamt	2012	Per municipality
	Workers in agriculture, forestry and fishery	Person per 1000		2005	

4.2.2. Methods

Boosted regression trees (BRT) were chosen because it allows to model nonlinear response of variables (Elith et al. 2008) ("gbm" package in R, R version 3.1.0). The predictive power of the model was evaluated with the area under the curve (AUC) (Fawcett 2006). Because of spatial autocorrelation and possible over-fitting (Bahn and McGill 2013), AUC based on 10 folds cross-validation were provided. A model is overfitted when, by increasing the complexity, the model based on the calibration data starts to fit the noise in addition to the signal. This improves the AUC of the calibration data based model but reduces the AUC of the validation data based model.

Predicted probabilities and predicted presences and absences, using sensitivity equals specificity as threshold, were mapped. The other outputs interpreted were a table of the relative importance of each variable in the model, response curves, the relative strength of interactions and, interaction graphs.

Two main presence-absence models were built: a human PUUV cases model and a bank vole model. To bring further details, a human PUUV cases model was also computed for rural counties only.

4.3. Results

The predictive power of human PUUV cases model was very good with an AUC of 0.985 ± 0.005 and a cross-validated AUC of 0.806 ± 0.030. The bank vole model had a good predictive power with an AUC of 0.951 ± 0.010 and a cross-validated AUC of 0.755 ± 0.022.

The false presences/absences and the predicted probabilities for presence of human PUUV cases and bank voles were mapped (Figure 9). In the human model, probabilities were highest along the axis Stuttgart – Cologne, in Mecklenburg Western-Pomerania, in the center of Saxony-Anhalt, Saxony and Bavaria. The lowest probabilities followed the direction South-South-East to North and the direction North-West to North-East. The classification threshold between presence and absence was 0.68. False positives were generally located in urban counties, within or along the general presence axis. False absences were more scattered. In the bank vole model, the highest predicted probabilities for bank vole presence were located in the North, center and in one location in South of Germany and the lowest predicted probabilities were in the South of Brandenburg. The classification threshold equalled 0.74. False presences were located in a cluster of high probabilities and true presences, while false absences were located in a cluster of low probabilities and true absences.

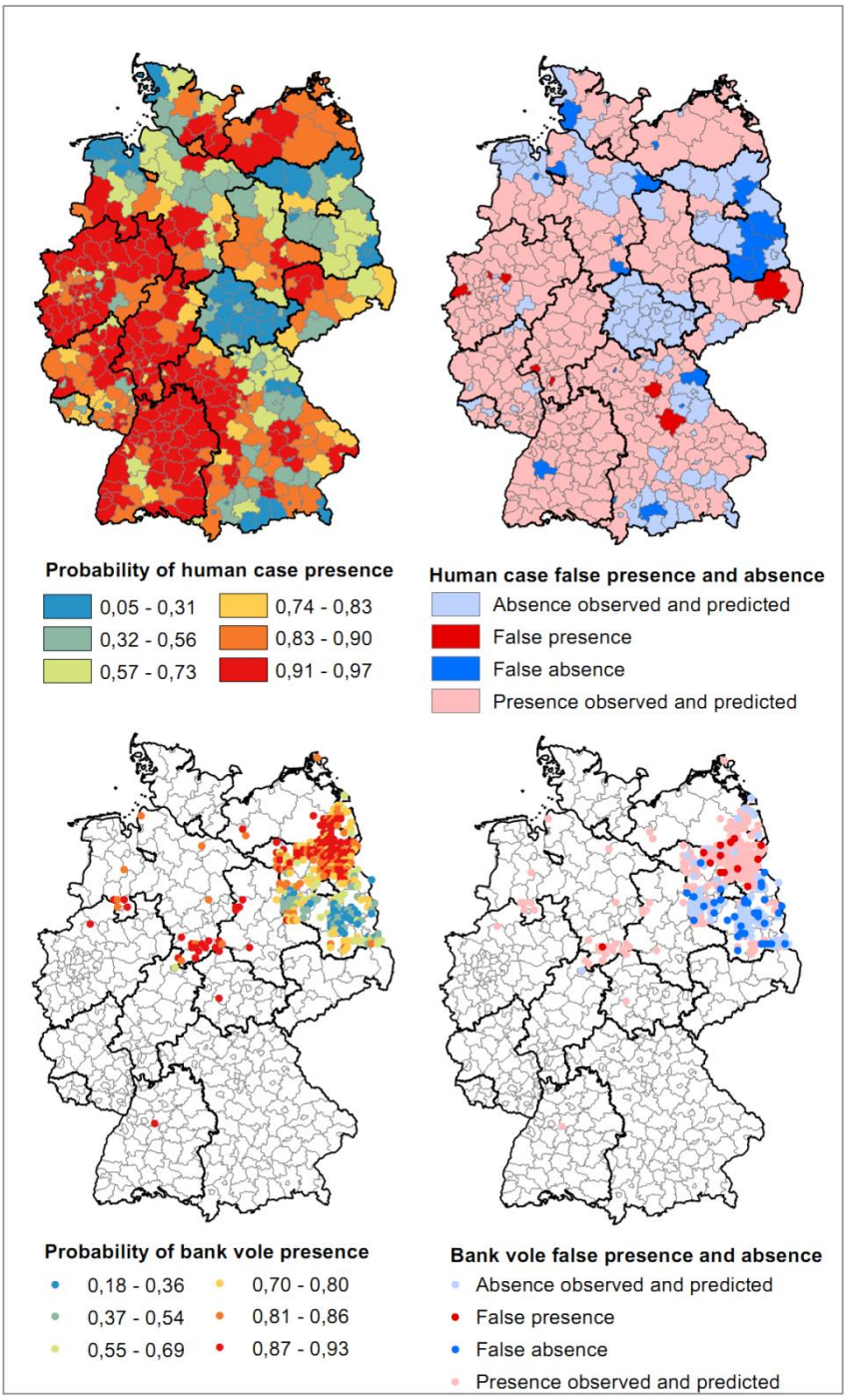


Figure 9: Probability and false presence/absence maps from boosted regression trees on human Puumala virus infection cases and bank vole presence

Table 12: Relative importance of the variable in the human hantavirus case model and in the bank vole model, colours gradient represent the magnitude of the relative importance

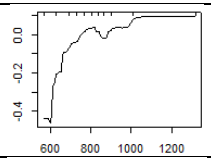
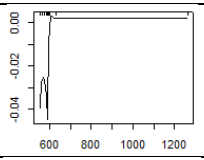
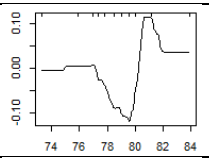
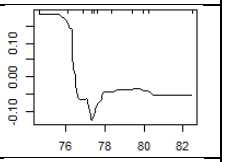
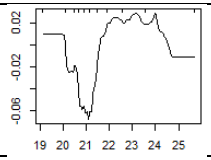
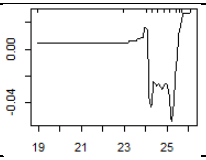
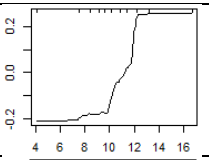
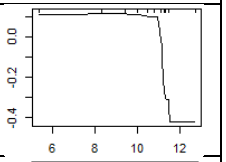
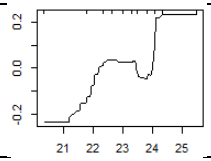
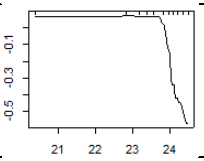
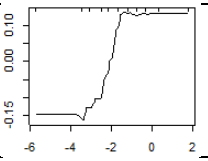
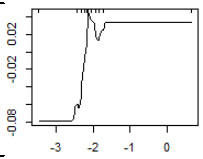
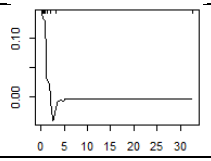
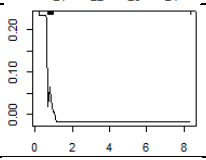
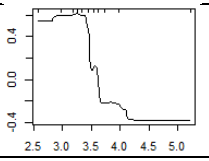
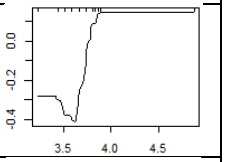
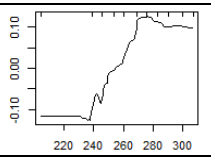
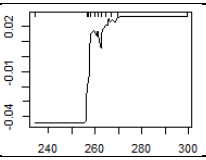
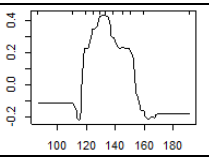
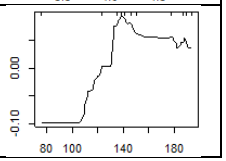
Variable	Relative importance (%)	
	Human case model	Bank vole model
Population	17.426	
Wind speed	8.697	9.329
Soil water index	7.504	2.920
Workers in agriculture, forestry and fishery	5.792	
Broadleaved forest	4.576	3.500
Number of warm days	4.543	7.179
Maximum temperature in summer	4.135	9.409
Waterbody	3.573	6.481
Precipitation	3.569	1.825
Mixed forest	3.476	7.323
Forest shape index	3.361	3.252
Built-up areas in ecotone	3.142	4.662
Coniferous forest	3.080	1.864
Minimum distance between forest patches	2.930	1.209
Minimum temperature in winter	2.812	2.198
Relative humidity	2.805	6.899
Agriculture	2.609	3.354
Forest contiguity	2.577	2.573
Built-up areas	2.438	4.005
Number of dry days	2.273	2.621
Slope	2.257	3.465
Growing season length	2.174	1.413
Elevation	2.151	7.794
Snow depth	1.946	4.375
Soil texture	0.153	2.349

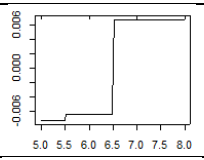
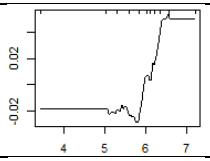
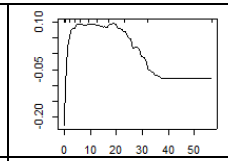
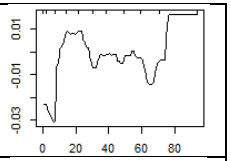
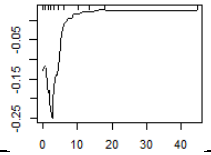
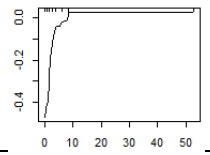
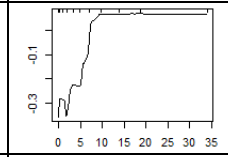
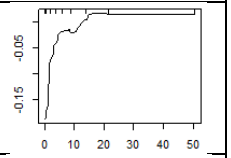
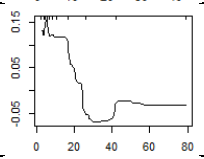
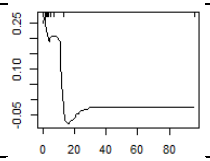
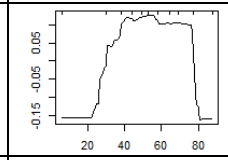
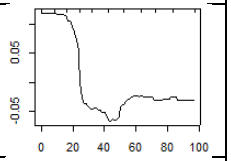
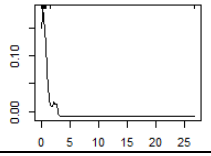
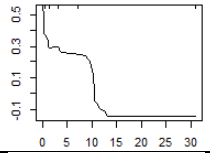
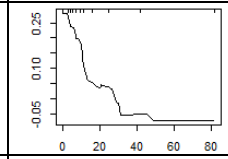
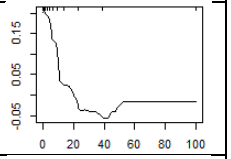
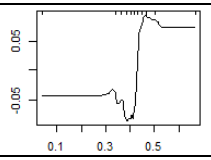
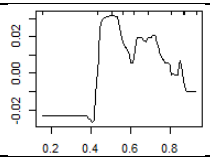
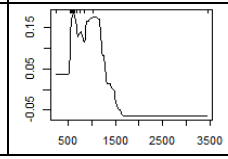
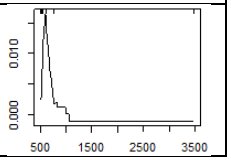
The five variables with the highest relative importance in the human PUUV case model were human population size, wind speed, soil water index, fraction of workers in agriculture, forestry and fishery, and the proportion of broadleaved forest, and in the bank vole model, maximum temperature in summer, wind speed, elevation, proportion of mixed forest, and number of warm days (Table 12). The relative importance of environmental variables varied greatly between the human PUUV case model and the bank vole model.

The human PUUV case model based on rural counties resulted in similar results than the human PUUV case model based on all counties. Only three variables showed different response curves: soil texture, proportion of built-up areas, and proportion of agricultural areas. Variables with a global positive trend on the presence of human PUUV cases were (sorted by relative importance): human population size, proportion of broadleaved forest, number of warm days, maximum temperature in summer, precipitation, proportion of mixed forest, minimum temperature in winter, forest contiguity, slope, growing season length, elevation, soil texture (and the proportion of built-up areas in rural counties model) (Figure 10). Wind speed, proportion of workers in agriculture, forestry and fishery, proportion of waterbodies, proportion of built-up areas in ecotone, proportion of coniferous forest, minimum distance between forest patches, proportion of built-up areas, snow depth (and proportion of agricultural areas in the rural counties model) had a negative effect. The last variables, soil water index, relative humidity, proportion of

agricultural areas, number of dry days (and the soil texture in the rural counties model) had mixed responses.

For the bank vole model, variables with a positive effect on the presence of the rodent were wind speed, elevation, proportion of mixed forest, proportion of broadleaved forest, slope, forest shape index, soil texture, minimum temperature in winter, precipitation and growing season length (Figure 10). A negative effect was observed for the following variables: maximum temperature in summer, relative number of warm days, humidity, proportion of waterbodies, proportion of built-up areas in ecotone, snow depth, proportion of built-up areas, proportion of agricultural areas, and minimum distance between forest patches. Finally, the soil water index, number of dry days, forest contiguity and proportion of coniferous forest had a mixed effect.

Variable	Human case	Bank vole	Variable	Human case	Bank vole
Precipitation (mm)			Relative humidity (%)		
Number of dry days (day)			Number of warm days (day)		
Maximum temperature in summer (°C)			Minimum temperature in winter (°C)		
Snow depth (cm)			Wind speed (m/s)		
Growing season length (day)			Soil Water Index (relative unit)		

Soil texture (coarse to fine)			Coniferous forest (%)		
Mixed forest (%)			Broadleaved forest (%)		
Built-up areas			Agricultural areas (%)		
Waterbody (%)			Built-up areas in forest ecotone (%)		
Forest contiguity (relative unit)			Minimum distance between forests (m)		

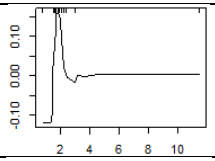
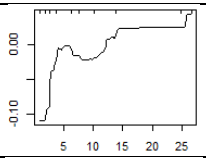
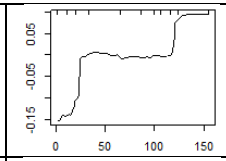
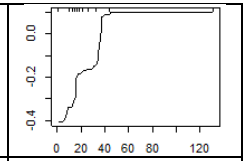
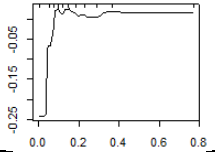
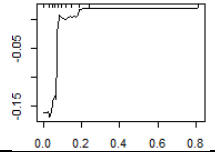
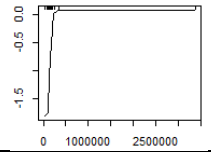
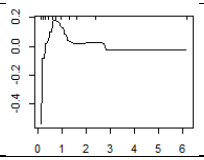
Forest shape index (relative unit)			Elevation (m)		
Slope (degree)			Population (person)		/
Workers in agriculture, forestry and fishery (person per 1000 inhabitants)		/			

Figure 10: Response curves from the boosted regression trees on the probability of human hantavirus and bank vole presence

In the human PUUV case model, the pairs of variables with a relative interaction size >1 were minimum temperature in winter and wind speed, the number of warm days and human population size, the soil water index and relative humidity, the proportion of built-up areas in ecotone and wind speed, the proportion of broadleaved forest and wind speed, wind speed and human population size, and the proportion of broadleaved forest and human population size.

In the bank vole model, the pairs of variables with a relative interaction size >1 were number of warm days and relative humidity, the maximum temperature in summer and relative humidity, wind speed and the proportion of mixed forest, and elevation and maximum temperature in summer.

The soil water index and the relative humidity in the human PUUV case model both showed a mixed response curve. In the interaction graph, there was a considerable increase of infection probability at the range 115-150 unit of soil water index and 81-83% of relative humidity (Figure 11). Wind speed and the proportion of mixed forest had a positive effect on the probability of bank vole presence (Figure 11). The probability of bank vole presence was highest at a proportion of mixed forest >15% and wind speed >3.5 m/s

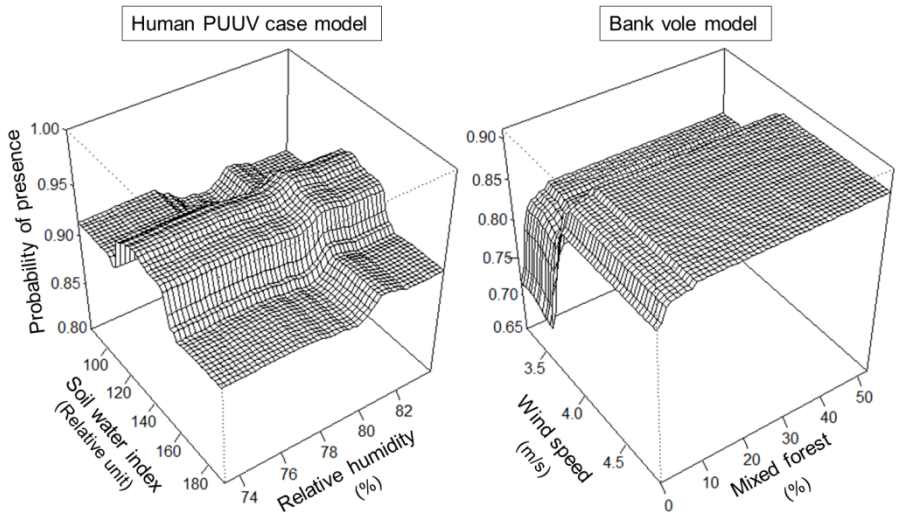


Figure 11: Interaction graphs for the probability of presence of human PUUV hantavirus cases based on soil water index and the relative humidity and, the probability of bank vole presence based on wind speed and the proportion of mixed forest

The distribution of the precipitations, dry days, forest contiguity, and forest shape index were very different between the two models. The third quartile of precipitations in the human PUUV case model was below the second quartile of precipitations in the bank vole model. The third quartile of dry days, forest contiguity and forest shape index in the bank vole model were below the first quartile of these variables in the human PUUV case model.

4.4. Discussion

Both models had very good predictive powers. Also, the human PUUV case probability map was consistent with the map of human PUUV cases per inhabitants (Heyman et al. 2012). The human PUUV

cases and the bank vole presence were well predicted by environmental factors. Bank voles are found in mixed/broadleaved contiguous forest with little built-up and milder winter with less snow cover was validated. PUUV human cases were also found in these environments but with some differences linked to the outdoors exposure. Indeed, warm days, high summer temperatures and lower wind speed, unfavourable for bank voles, were positively associated with PUUV human cases as well as population density and number of workers in agriculture, forestry and fishery.

4.4.1. Untangling the effect of hazard and exposure

Examining first the relative importance and second the response curves in the two models sheds light on the hazard and exposure aspects of risk.

First, looking at the relative importance, elevation was the variable with the third highest relative importance in the bank vole model while it was the third least important variable in the human model. Indeed, the relative importance of variables varied greatly between the human PUUV case model and the bank vole model. It indicates that the population of bank voles is not sufficient to understand the risk of human infection. The proportion of broadleaved forest, which was in the top five of relative importance in the human PUUV case model, seems to have a substantial impact on exposure while it appeared to be the mixed deciduous forest that had the greater impact on bank vole presence and hazard. Indeed, people prefer to spend time in

broadleaved forest than in coniferous forest (Edwards et al. 2012). People regard coniferous forest as artificial, man-made and dark, while broadleaved forest are characterized as native, natural, and with diverse colours (Schriewer 1998). Differences in human preferences may however be observed based on personal origin. In Germany, people living in regions dominated by broadleaved forest have stronger preference for this forest type than people from other regions (Schraml and Volz 2009). Mixed forest represents an important part of forests in Germany, due to forest management initiated in the 1980's aiming to convert coniferous monocultures to mixed stands (Knoke et al. 2008). Mixed forests are more ecologically valuable and biodiverse than coniferous forest (Cavard et al. 2011), and seem to be preferred by bank voles in our model.

Second, areas with a high exposure or high hazard can be described using the response curves. Warm summers, mild winters with a thinner snow cover and low wind speed favoured human infection risk. These outcomes are probably related to human outdoor activities and consequently to direct exposure to PUUV. Bank voles were also favoured by mild winters with thin snow cover but, conversely, were not favoured by high maximum summer temperature and low wind speed. The hazard is therefore lower in areas with hot summers and low wind speed, where exposure is expected to be more important. Hot and dry areas may be less productive areas in terms of seeding, an important food source for bank voles (Gurnell 1993), even though bank voles habitat (e.g. broadleaved forest) may prevent such conditions. Food is available for longer where the growing season is

extended and this seems to be reflected in our bank vole model. Similarly, in the human PUUV case model, a longer growing season that increases hazard is probably also increasing exposure by human outdoor activities.

Soil humidity and NE are often related to better ex-vivo virus survival. In the model, bank voles were found in environments with finer soil texture, indicating that soil humidity is also impacting hazard through bank vole presence. The impact of soil humidity on human PUUV cases probably reflects hazard rather than exposure but studies on human preferences are needed to confirm this hypothesis.

Land use influenced both models. The percentage of coniferous forest was negatively correlated to human PUUV cases and had a nonlinear effect in the bank vole model, pointing again at differences between hazard and exposure. The proportion of broadleaved and mixed forests had a positive impact on both hazard and exposure, representing the bank vole habitat and affecting human recreational choices. Rural counties with more built-up areas were expected to be more at risk of PUUV infection in human, potentially because of closer contact between humans and bank vole habitat. Buffers with more built-up and agricultural areas had a lower probability of bank vole presence, reflecting the hazard. Also, more built-up areas in forest ecotones were related to a lower probability of bank vole presence, and possibly to a lower abundance. Interestingly, in the human PUUV case model, the proportion of built-up area around forest was negatively correlated to PUUV infections although such a setting may promote contact

between humans and voles. One explanation could be that, at the county level, this variable reflects hazard. It is possible that with finer spatial detail (e.g. PUUV cases reported by infection), this relation would be positive because of the increasing exposure.

Caution was necessary when comparing human and bank vole models. Indeed, even if the temporal scale is the same, the spatial scale distribution differs between models. The variables included in the human PUUV case model were averaged within the county while the variables for the bank vole model were averaged in a 3km buffer. Also, the spatial distribution of bank voles is nested in the human PUUV cases distribution. Precipitation, dry days, forest contiguity, and forest shape were not discussed here as their ranges were too different between the two sets of data.

The impacts of several environmental variables can be assessed on hazard and exposure thanks to the human PUUV case and the bank vole model. The hazard is however not totally described by the bank vole distribution as we focused on bank vole presence irrespective of the presence of PUUV in the rodent host. Infections of bank voles should be a better proxy for hazard. When the results of both models were dissimilar, it is easy to link the effect of one variable to hazard or exposure but when the results were similar, proportion of the hazard or exposure due to the variables can only be hypothesized.

4.4.2. The complexity of Germany

Germany presents much more complexity than other European countries regarding hantavirus and host ecology (see Chapter 5). The country is located at the interface of Dobrava-Belgrad, Tula and Puumala hantavirus (Olsson et al. 2010). All these hantavirus species are present in Germany and are hosted by different rodent species, PUUV being the most prevalent (Heyman et al. 2012). Also, diverse ecoregions and climatic conditions are encountered in Germany. The human PUUV case presence/absence data shows a strong spatial structure (Figure 8). Several variables follow a gradient over the country but there is no large-scale gradient reflecting this human PUUV cases spatial structure or even, a set of phenomena (e.g. climatic gradient). Therefore, many local differences exist and the landscape scale of study seems indicated. In the bank vole model, predicted probabilities were however clustered in two large regions of high or low probabilities. False presences were clustered in these high probabilities areas and false absences in these low probabilities areas. These two clusters were possibly the result of unidentified variables acting at a larger scale, defining larger regions which are suitable or not for the bank vole. They could also result from the different sampling effort over the years.

The complexity of Germany can be partly reflected by the interactions between the variables. For example, in the human PUUV cases model, the effect of soil water index depended on the level of relative humidity and the other way round (Figure 11). The soil water index,

representing the moisture conditions of the soil, is driven by water infiltration processes. It is possible that the relative humidity partially depends of the soil water saturation and vice-versa. The combination of soil water index and relative humidity is probably needed to reflect a moist environment. Also, interaction between wind speed and the proportion of mixed forest in the bank vole model may potentially be understood by the better resistance of mixed forest to storms and high winds than monocultures (Knoke et al. 2008). Interactions remain hard to interpret but, in the case of Germany, are essential to model all the complexity encountered on the territory.

4.5. Conclusion

Hazard and exposure of PUUV infection in humans can be unravelled when looking at both the human infections and the presence of the rodent host. This was here successfully done for PUUV in Germany.
We demonstrated that the presence of human PUUV cases can reflect hazard or exposure and sometimes, both.

Some factors had a positive or negative effect on the hazard but a mixed and complex effect on the exposure. Indeed, the precipitation, the soil humidity and the forest contiguity had a positive effect on the probability of bank vole presence and the relative air humidity, the built-up areas and built-up areas in forest ecotone had a negative effect. Some factors had the same effect on hazard and exposure. High temperature in winters, less snow cover, extended growing season,

mixed forest and broadleaved forest were thereby preferential environment for the bank vole and promoted human outdoors activities. Finally, some factors had opposite effect on hazard and exposure such as more warm days, higher summer temperature and less wind that were not beneficial for the bank voles but well for outdoor activities.

The new knowledge brought by these models may help for the identification of factors relevant to build further predictive model in Germany.

Chapter 5 Landscape and regional environmental analysis of distribution of hantavirus human cases in Europe[4]

After investigating the different tools to model zoonoses with human data, some of these are here used to address the question of the spatial scale and, for the first time, of the uneven European spatial distribution of NE.

Results indicate that, at a broad scale, the effect of one variable on disease may follow three scenarios, differently impacting the model: the variable has a linear effect, the variable has a nonlinear effect, and the effect can change depending on local specificities. The presence of NE is likely in populated regions with well-connected forests, more intense vegetation activity, low soil water content, mild summers and cold winters. In these regions, landscapes with a higher proportion of built-up areas in forest ecotones and lower minimum temperature in winter are expected to be more at risk. Climate and forest connectivity have a stronger effect at the regional level. If variables are staying at their current values, the models predict that nephropathia epidemica may know intensification but should not spread.

---

[4] Adapted from: Zeimes CB, Quoilin S, Henttonen H, Lyytikäinen O, Vapalahti O, Reynes JM, Reusken C, Swart AN, Vainio K, Hjertqvist M and, Vanwambeke SO. "Landscape and regional environmental analysis of distribution of hantavirus human cases in Europe", *Frontiers in Public Health*, in press.

Chapter 5

5.1. Introduction

In Europe, the spatial distribution of NE appears uneven. A number of studies have investigated the environmental factors influencing PUUV distribution, but never more broadly than at the national level (e.g. (Linard et al. 2007b, Zeimes et al. 2012)). While the main environmental factors are fairly well known, and hypotheses explaining the broad distribution of PUUV exist (Olsson et al. 2010, Heyman et al. 2012, Vaheri et al. 2012), no study so far has investigated the distribution of NE cases in humans or in bank voles at the continental scale. This is attempted here for Belgium, Finland, France, the Netherlands, Norway and Sweden with a rich set of environmental variables related to climate, land use, vegetation, soil and human distribution.

Looking at a broad study area means that environmental factors operating at various scales need to be considered. Our first objective is to understand the level at which variables operate and how they influence PUUV distribution. The effect of the environmental factors is tested jointly at the landscape and regional scale, using multilevel models.

Provided a regional structure is identified, investigating in further details the regional and local effects of environmental factors is valuable in understanding the spatial distribution of NE cases across Europe. Our second objective is to understand what the effects of environmental variables are across diverse European ecoregions. This

is done by using both logistic regressions on ecoregions and overall niche modelling.

5.2. Materials and methods

5.2.1. Materials

5.2.1.1. Disease record collections

Data are collected for Belgium, Finland, France, the Netherlands, Norway and Sweden. As recording protocols and time span of data availability vary across countries, the study focuses on presence or absence locations. At least five years of records are available for the countries included. A presence record is included in the database when spatial information is provided. In France, NE cases from 2003 to 2012 are reported by municipality of exposure (473 records) or by municipality of residence (289 records), and 192 records had no spatial information (Centre National de Référence des Hantavirus, Institut Pasteur). In Belgium, NE cases from 2000 to 2010 are reported by municipality of residence (Institut scientifique de Santé publique-Wetenschappelijk Instituut Volksgezondheid (WIV-ISP)). In the Netherlands, official notifications for 2008-2012 are provided by municipality of residence (Rijksinstituut voor Volksgezondheid en Milieu (RIVM)). This dataset is completed with data from a cross-sectional population serological survey of 2929 persons sampled in 2006 (RIVM) and with the RIVM diagnostic database (RIVM is one of three diagnostic centers of the country) for 2007-2011. In Norway,

cases reported from 1989 to 2012 are provided by municipality of residence (Folkehelseinstituttet (FHI)). In southern Sweden, human infections reported for 1997-2012 are collected by municipality of infection (Smittskyddsinstitutet (SMI)). In northern Sweden, 212 human infections by coordinates of infection are recorded for 1991-1998 (SMI, see Chapter 2). In Finland, the National Institute of health and Welfare collects the data at the hospital district level for 2004-2009 in some high-incidence and low-incidence regions.

The map of presences and absences is presented in Figure 12. While the resolution is homogeneous across the data (municipality for most countries), the average size of this unit varies greatly. In order to accommodate this, the data are converted into points locations in the following way. The centroid of municipalities between 80 and 150 km² is used as the absence/presence point. Small municipalities (area ≤ 80 km²) are aggregated in a 10 km resolution grid of points. Larger municipalities (area > 150 km²) are represented by points located where the population density is highest (2.5 arc-minutes resolution, Gridded Population of the World from Center for International Earth Science Information Network (CIESIN)). Geographical coordinates are used when available. In Finland, the presence and absence data are aggregated on a 10 km grid of points. In northern Sweden, 300 absences are randomly selected amongst other dwellings (see Chapter 2). There are a total of 1933 presences and 6425 absences.

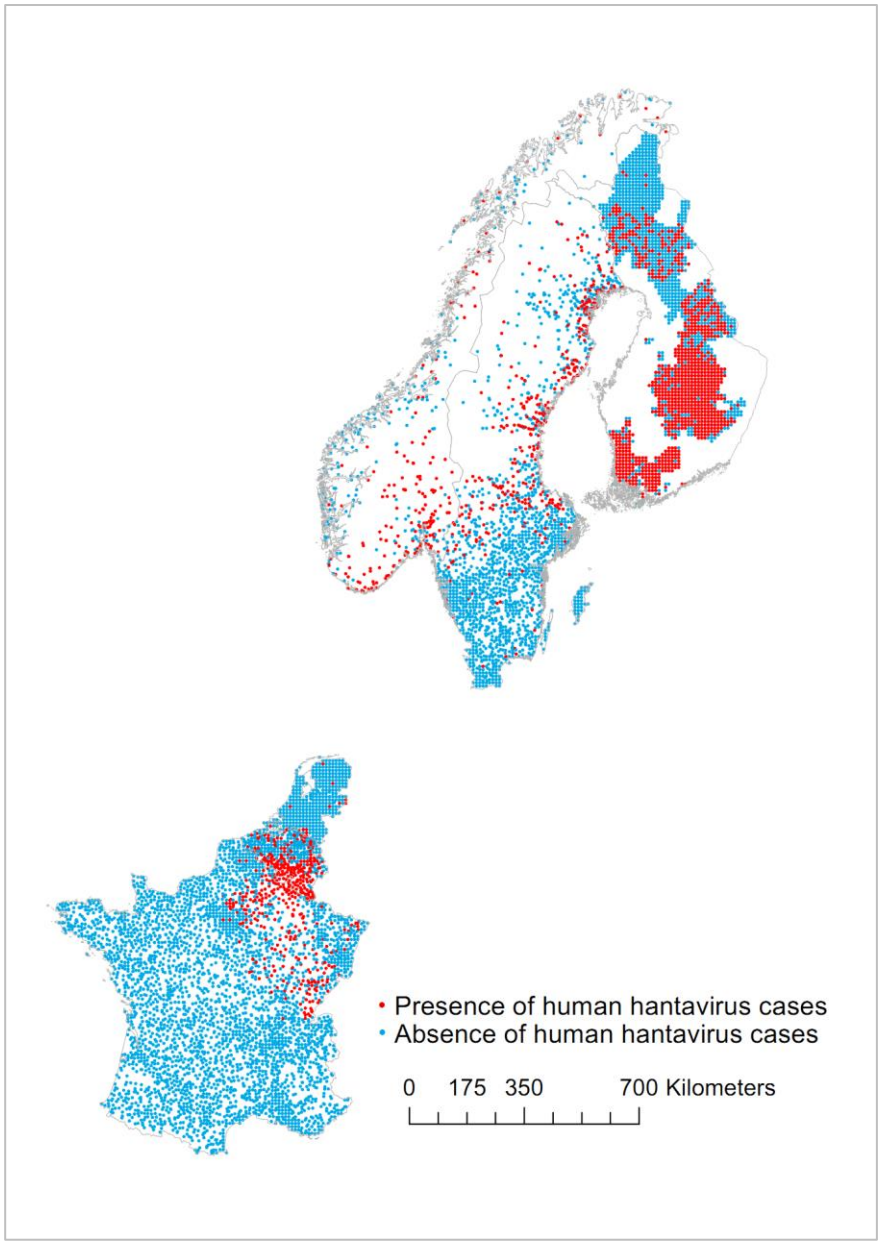


Figure 12: Map of presences and absences of human hantavirus cases in Europe

5.2.1.2. Environmental variables

Relevant environmental variables are identified based on a literature review of PUUV ecology in Europe, as summarized in Table 13. As this study focuses on the spatial distribution, seasonal or yearly temporal fluctuations are ignored and values are averaged over time. Two ecologically relevant levels are identified for the multilevel analyses: the landscape (individual) and the region (group). For operational purposes, the landscape level is defined as the area covered by a circular radius of 5km around a presence/absence record. A sample size of at least 50 groups is recommended for a good estimation of regression coefficients in multilevel regressions (Maas and Hox 2005). 50 regions were defined using Ward clustering (Ward Jr 1963) (R i386 3.1.0, "Rcmdr" package, "FactoMineR" plug-in) based on the mean annual temperature and the annual temperature range (Worldclim, resolution of 1km, monthly mean from 1950 to 2000). Environmental variables are calculated at the landscape level and at the regional level (ArcGIS 10.1 and FRAGSTATS version 4). The regional level is only used in the multilevel model. Specific hypotheses are made for the effect of environmental variable at the landscape and regional levels.

a) Climatic variables

Climatic variables are assumed to have the same effect at both levels. Higher mean annual precipitations (Worldclim, resolution of 1km, monthly mean from 1950 to 2000) are assumed to reflect higher air

humidity and should increase the probability of presence of NE cases. The mean maximum temperature in the summer (Worldclim, 1km, monthly mean from 1950 to 2000) is calculated to identify the driest and most unfavourable environments for virus ex-vivo survival. Summer is defined as the months of June, July and August. In some areas, this can reflect more human outdoors activities. The mean minimum temperature in the winter (Worldclim, 1km, monthly mean from 1950 to 2000) is added to our database. Winter is defined as the months of December, January and February. Lower minimum temperatures in winter are assumed to increase the presence of NE cases (thanks to better virus ex-vivo survival) as well as the range of minimum temperature suitable for more human outdoors activities. Snow cover is here measured by the annual mean percentage of pixel area covered by snow (MODIS, 0.05 degree, 2000 to 2008 monthly mean). Less extensive snow cover (increasing human exposure) and extensive cover (favouring bank voles and virus ex-vivo survival) should increase the probability of NE cases presence while very extensive cover (reflecting harsh condition) should decrease the probability.

b) Land use variables

We hypothesize that the proportion of forest (Corine 2006 (EEA), vector map) indicates the presence of bank voles at the landscape level and, the presence and also the abundance of bank voles at the regional level. The proportions of coniferous, broadleaved and mixed forest (Corine 2006 (EEA), vector map) are used. It is assumed that

broadleaved forest is preferred over coniferous forest but the effect of forest type may differ between the landscape and regional levels. Indeed, if a region is characterized by a large proportion of coniferous forest, the proportion of broadleaved forest could be significant at the landscape level but not at the regional level as, regionally, coniferous forests are used by default by the bank vole. The contiguity index of forest patches (Corine 2006 (EEA), vector map) is used as a proxy to represent contacts between rodents, increasing prevalence in bank voles and so, NE cases presence. At the landscape level, a higher proportion of built-up areas in a buffer of 150m around forests (Corine 2006 (EEA), vector map) would indicate a higher probability of contact between bank voles and humans. At the regional level, high levels of urbanization may conversely limit PUUV transmission.

c) Vegetation indices

The mean enhanced vegetation index (EVI) (MODIS, 0.0083 degree, 2001 to 2012 mean) and number of green days per year (MODIS, 0.005 degree, mean from 2006 to 2010) are included. Green days represent the number of days between the detection of the start of a growth cycle and the end of the cycle, when vegetation greenness decreases. These indices are hypothesized to reflect food availability and the number of litters (more food may reflect smaller territories and more nests) at the landscape level, as well as a longer breeding season at the regional level, and should thus increase the disease presence probability at both levels.

d) Soil variable

Higher soil water index (SWI) (TU-WIEN, 25km), averaged over 2007-2010, should improve the virus ex-vivo survival at the landscape level. At the regional level, we assume that a higher soil water index gives information about land use unfavourable for the bank voles (e.g. presence of marshes and bogs).

e) Human distribution variable

The probability of human NE cases increases with human population density, whether at the landscape or regional level. To reflect the presence of humans, the mean distance weighted population proximity index (Environment Research Group Oxford, 0.0083 degree, 2005) is considered. The population proximity index represents the population likely to visit a place taking into account the population in the surrounding environment (Alexander and Wint 2013). The index for a pixel is the sum of persons of all pixels within 30km that are weighted by a linear-distance value (between 1 for the closest to 0 for the furthest).

Table 13: Variables and their hypothesized effect (X) on the bank voles' abundance, virus ex-vivo survival and human repartition (numbers in brackets indicate references)

Variables	Bank voles	Virus	Human	Resolution	Units	Sources
Annual precipitation (Kallio et al. 2006, Verhagen et al. 1986)	X	X		1km, 1950-2000	mm	Worldclim
Maximum temperature in summer (Kallio et al. 2006)		X		1km, 1950-2000	° Celsius	Worldclim
Minimum temperature in winter (Linard et al. 2007b, Kallio et al. 2006, Piechotowski et al. 2008, Evander and Ahlm 2009)	X	X	X	1km, 1950-2000	° Celsius	Worldclim
Snow cover (Kallio et al. 2006, Hansson 1999, Hansson and Henttonen 1985, Hansson and Henttonen 1988, Hanski et al. 1991, Hanski et al. 2001, Korslund and Steen 2006)	X	X		0.05°, 2000-2008	Area percentage	MODIS
Proportion of forest	X			100m	Area percentage	Corine 2006 (EEA)
Proportion of coniferous forest (Olsson et al. 2005)	X			100m	Area percentage	Corine 2006 (EEA)
Proportion of broadleaved forest (Linard et al. 2007a)	X			100m	Area percentage	Corine 2006 (EEA)
Proportion of mixed forest	X			100m	Area percentage	Corine 2006 (EEA)
Forest contiguity index (Vanapeldoorn et al. 1992, Kozakiewicz et al. 2000)	X			100m	None	Corine 2006 (EEA)

Built-up areas in forest ecotones (Linard et al. 2007a, Crowcroft et al. 1999)	X		X	100m	Area percentage	Corine 2006 (EEA)
Enhanced vegetation index (EVI) (Barrios et al. 2010)	X			0.0083°, 2001-2012	None	MODIS
Number of green days (Barrios et al. 2010)	X			0.005°, 2006-2010	Number of days	MODIS
Soil water index (SWI) (Kallio et al. 2006, Verhagen et al. 1986)	X	X		25km, 2007-2010	None	TU-WIEN
Population proximity index (Alexander and Wint 2013)			X	0.0083°, 2005	Number of persons	Environment Research Group Oxford

5.2.2. Methods

5.2.2.1. At which levels do the variables operate and how do they influence NE distribution?

The spatial distribution of NE cases in Europe is first modelled using multilevel logistic regression in order to identify significant environmental variables at the landscape and regional level. The intraclass correlation coefficient (ICC), a measure of the proportion of the variance found at the regional level (Snijders and Boskers 1999, Merlo et al. 2005), is first calculated. If this value is high, a multilevel approach is recommended. A full multilevel model is subsequently fitted (Snijders and Boskers 1999).

Multilevel regressions allow including intercepts varying between regions (random intercept model) and/or slopes varying between regions (random slopes model). Variables at the landscape and regional level are tested in a model with random intercepts and a model with random intercepts and slopes ("lme4" package in R2.12.0). These models are compared with an empty model and with each other using an ANOVA test based on the AIC and the Bayesian information criterion (BIC). The final model includes the variables significant in regressions with one explanatory variable and the random intercepts and slopes which have better AIC and BIC. To avoid collinearity issues, variables with a VIF greater than 10 are removed. AUC is used to evaluate the predictive power of the model.

Disease presence probabilities, and false presences and absences are mapped.

5.2.2.2. What are the effects of environmental variables across European ecoregions?

In order to better distinguish the effects of environmental variables across Europe, logistic regressions are fitted for each ecoregion using landscape level variables. Seven ecological regions encountered in the study area are compared (European Topic Centre on Nature Protection and Biodiversity (ETC/NPB)) (Figure 13).

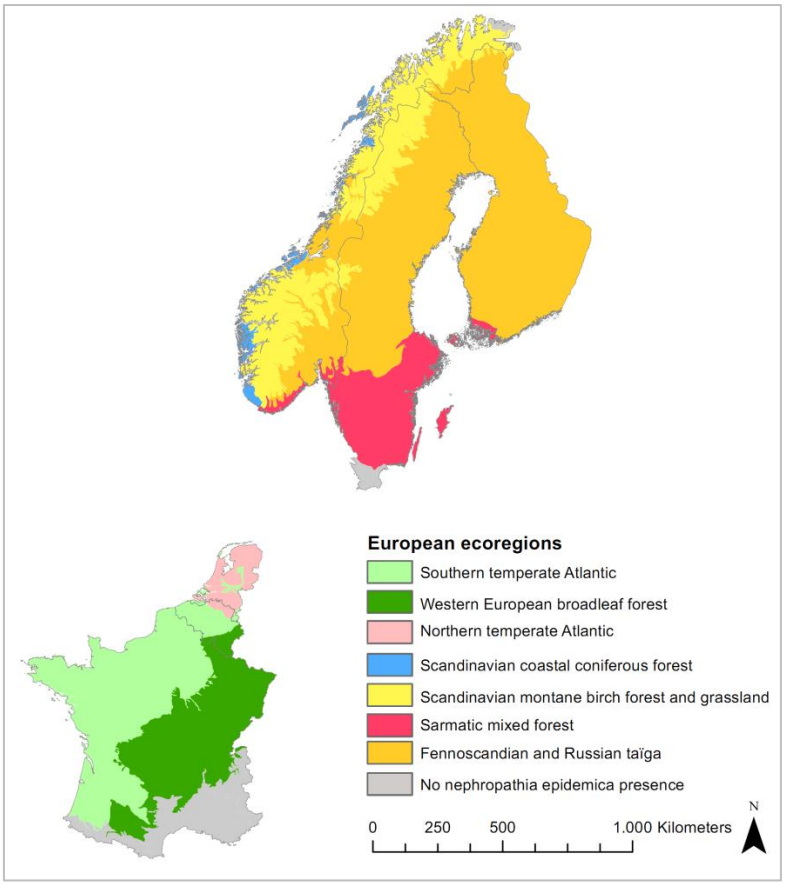


Figure 13: Digital map of European ecological regions for the study area

The effect of environmental factors is not necessarily linear and may be positive in some ecoregions, and negative in some others. Because of the broad range of values covered, this is likely to be encountered in our study area. This type of relationship is poorly addressed by classical regression, and a niche model is thus needed to explore the effect of the landscape level variables across the study area.

Boosted regression trees, which allow to model nonlinear response, are used to build a European model including all variables (“gbm”

package in R). The AUC is here based on cross-validation on 10 subsets as overfitting is a known issue with BRT. The relative importance of the variables are also calculated. Graphs of the probability of presence of the disease as a function of the value of the variable are produced. Boxplots representing the distribution of the variable in each ecological region, with colours corresponding to the map in Figure 13, are superimposed to this curve in order to jointly examine the global effect as identified by BRT and its subset in the variable range of each ecoregion. Finally, the signs of the significant coefficients (at the level of 0.05) of bivariate logistic regressions by ecoregion are added in front of each boxplot for comparison with the trend modelled by BRT in this ecoregion. An example graph is presented in Figure 14.

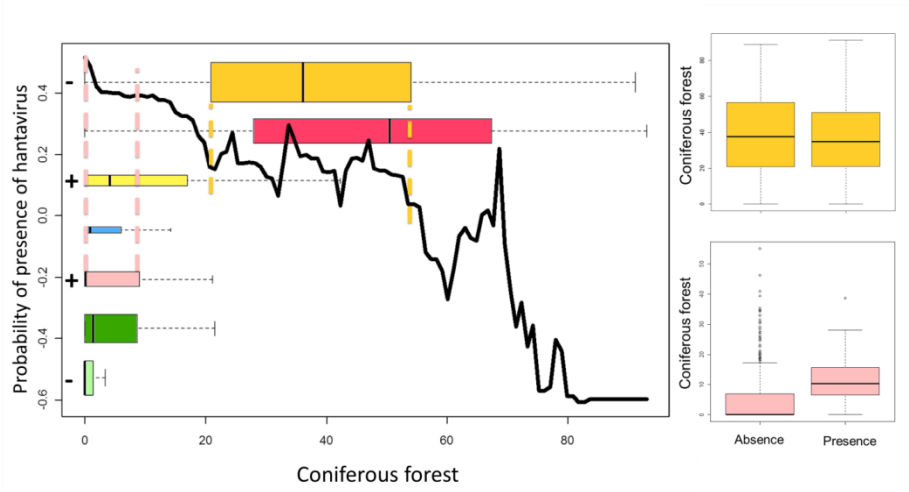


Figure 14: Black curve: relative probability of presence of hantavirus human cases according to the evolution of the variable as modelled by BRT. Coloured boxplots: distribution of the variable by ecoregions (the colours refer to figure 12 and the height of the boxplot is proportional to the number of points by ecoregion). + and – signs: sign of the logistic regression coefficient, if significant, for this variable in this ecoregion. Dashed coloured lines: boxplot extent (second to third quartile) of an ecoregion on the black curve.

5.3. Results

5.3.1. At which levels do the variables operate and how do they influence NE distribution?

The ICC shows that 57% of the variance is at the regional level. A multilevel model is therefore advised. Following the VIF analysis, snow cover, green days and the proportion of forest are removed from further analyses to avoid collinearity issues.

In bivariate analyses, all variables at the landscape and regional levels are highly significant, except the maximum temperature in summer at the landscape level.

The results of the multilevel logistic regression are presented in Table 14. The AUC is 0.895 ± 0.004 and indicates a good predictive power. Random slopes are included for the landscape level variables coniferous forest, broadleaved forest, mixed forest, contiguity of forest, built-up areas in forest ecotones and population proximity index and, at the regional level, minimum temperature in winter and EVI.

At the landscape level, the coefficient of the variable built-up areas in forest ecotones is significantly positive at the level of 0.001. The coefficient of the minimum temperature in winter is significantly negative at the level of 0.05. At the regional level, the coefficients of forest contiguity, EVI and, population proximity index are

significantly positive at the level of 0.05. Also, the maximum temperature in summer is negatively significant at the level of 0.01 and, the minimum temperature in winter and SWI at the level of 0.05.

The predicted probabilities and the false presences / absences of the model are mapped in Figure 15. The model predicts the highest concentrations of probabilities in north-eastern France, southern Belgium, inland Norway, central Sweden and, southern and middle Finland. False presences are mainly located in areas where the diseases is already established and along the Norwegian coast. A few false presences are also predicted in less expected areas: southern and central France and southern Sweden. False absences are more present at the western and southern border of the French disease distribution, in northern Belgium, in the southern Netherlands, in southern and northern Sweden and in northern Finland.

Table 14: Multilevel logistic regression (significant at the level of * 0.5, **0.01 and, * 0.001)**

	Landscape level			Regional level		
	Estimate	P-value	Random slope	Estimate	P-value	Random slope
Annual precipitations	0.00	0.02				
Maximum temperature in summer				**-0.94****	0.00	
Minimum temperature in winter	**-0.15***	0.04		**-1.62***	0.03	Yes
Coniferous forest	-0.00	0.27	Yes	-0.10	0.18	
Broadleaved forest	0.02	0.06	Yes			
Mixed forest	-0.01	0.70	Yes	-0.39	0.10	
Contiguity of forest	0.01	0.96	Yes	**37.66***	0.01	
Built-up areas in forest ecotones	**0.28*****	0.00	Yes			
EVI				**0.04***	0.03	Yes
SWI	0.00	0.21		**-0.14***	0.02	
Population proximity index	0.01	0.14	Yes	**0.11***	0.04	

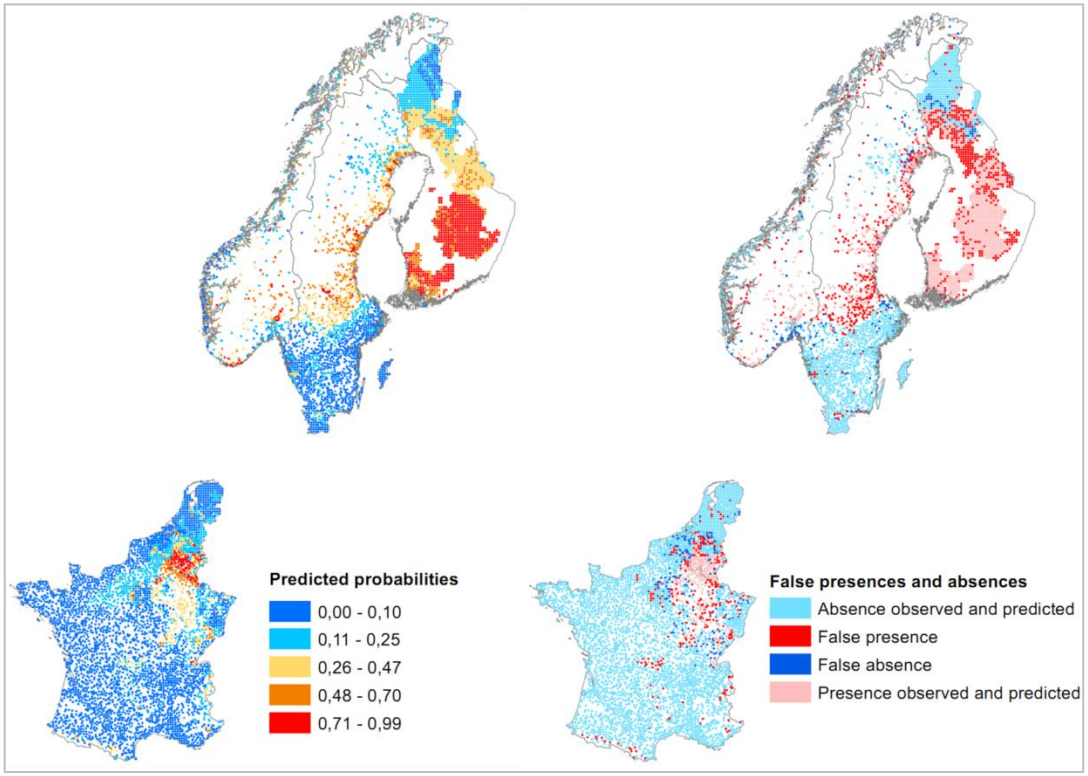


Figure 15: Map of predicted probabilities and false presences / absences of the European multilevel model (threshold for presence = 0.25, when the sensitivity equals the specificity at 84%)

5.3.2. What are the effects of environmental variables across European ecoregions?

The multilevel regression indicates that the effect of environmental factors varies across regions and random slopes were included for several variables. Logistic regressions are therefore elaborated for each ecoregion (Figure 13). When significant, forest, mixed forest, forest contiguity and built-up areas in forest ecotones are positive; EVI, green days and population proximity index are positive or negative; annual precipitation is positive except for the Scandinavian montane birch forest and grassland; the maximum temperature in summer is negative except for western European broadleaf and sarmatic mixed forest; the minimum temperature in winter is negative except for Fennoscandian and Russian taiga; snow cover and the SWI are positive except for Fennoscandian and Russian taiga; coniferous forest is negative except for the northern temperate Atlantic and the Scandinavian montane birch and broadleaved forest is positive except for northern temperate Atlantic and Fennoscandian and Russian taiga.

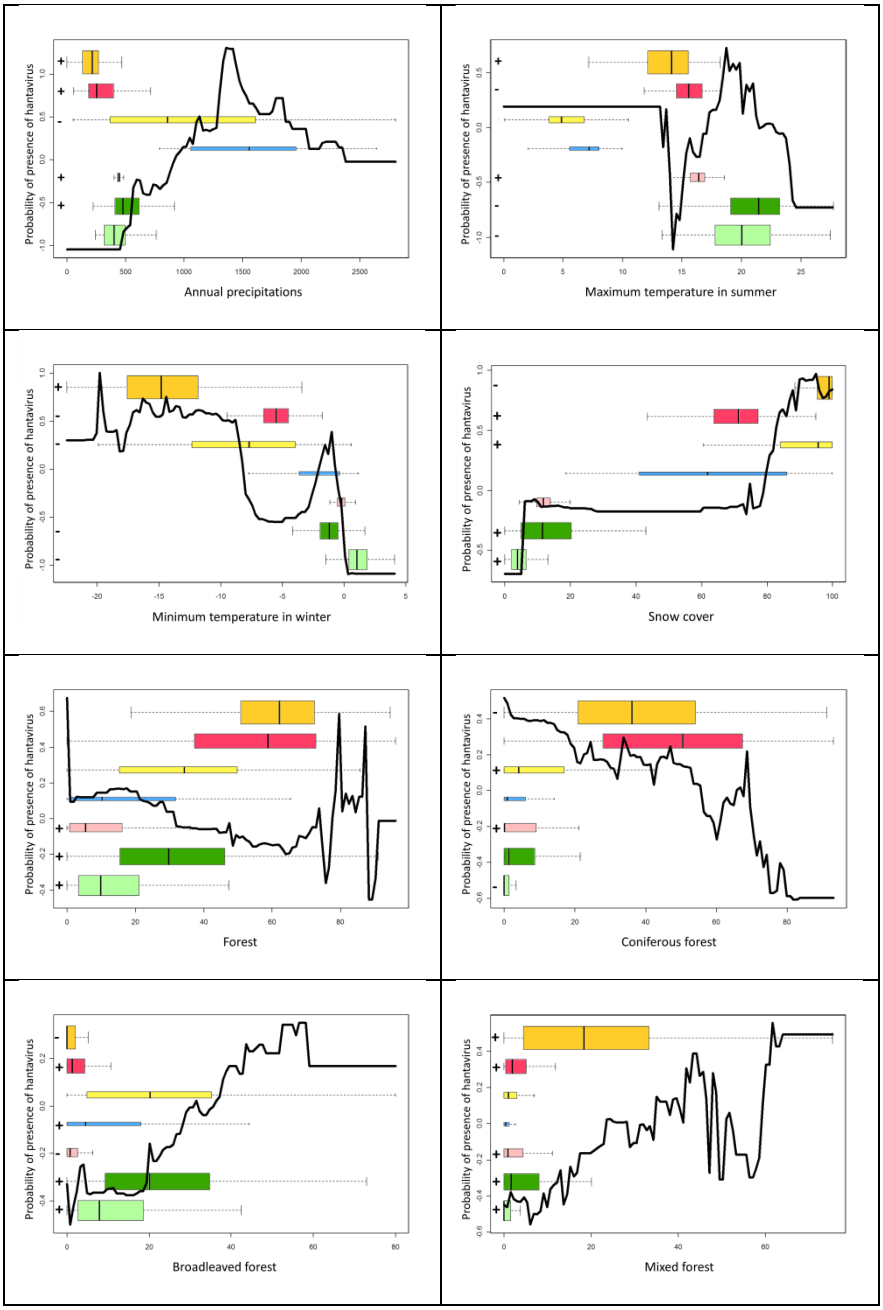
The AUC of the boosted regression trees model is 0.921 ± 0.003. The AUC on cross-validation is 0.893 ± 0.003. These indicate a very good predictive power of the model. The variables with a relative importance over 10 are the minimum temperature in winter, the snow cover, and the maximum temperature in summer.

The response curves (Figure 16) of snow cover, broadleaved forest, mixed forest, contiguity of forest, built-up areas in forest ecotones,

SWI and population proximity index show an overall positive trend. The curves of the minimum temperature in winter, forest, coniferous forest and EVI present an overall negative trend. Green days show a mixed curve. Annual precipitation and maximum temperature in summer display an increasing then decreasing trend in probabilities.

The boxplots in Figure 16 indicate that variables with a similar width of boxes between regions are the proportion of mixed forest, the built-up areas in forest ecotones and the population proximity index. Variables with boxplot extents following a gradient from South to North are minimum temperature in winter, snow cover, contiguity of forest and the SWI. The ecoregional boxplots of EVI and green days show a strong difference between western European and Fennoscandian ecoregions. Finally, the boxplot extent seems mixed between regions for the annual precipitations, the maximum temperature in summer, forest, coniferous forest and, broadleaved forest.

The general trend by ecoregions can be compared to the sign of the coefficient from the bivariate logistic regressions (Figure 16). The signs of the regression coefficient and the BRT trends are consistent for minimum temperature in winter, snow cover, mixed forest, contiguity of the forest, built-up areas in forest ecotones, green days and SWI.


Probability of presence of hantavirus
Annual precipitations
Maximum temperature in summer
Minimum temperature in winter
Snow cover
Forest
Coniferous forest
Broadleaved forest
Mixed forest

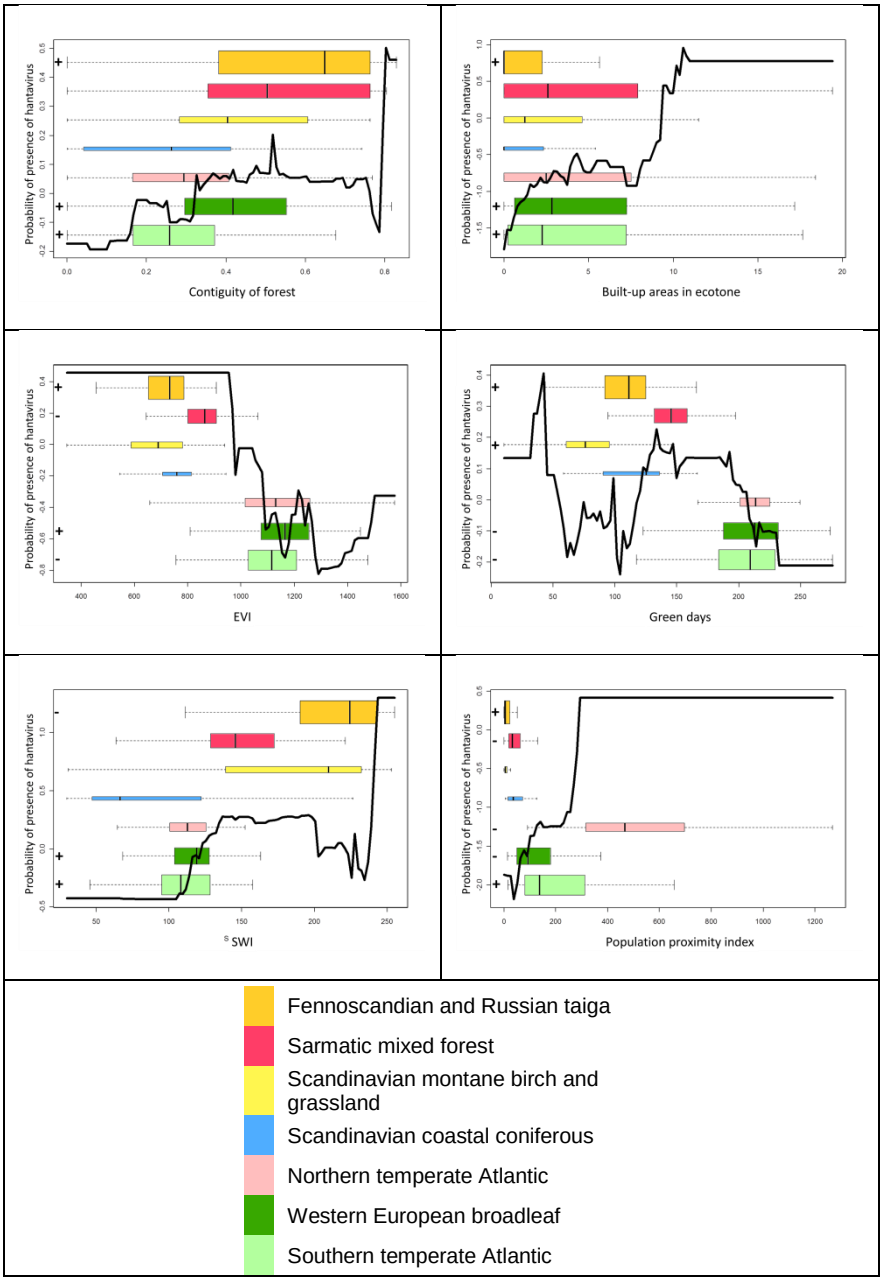


Figure 16: Response curves of the variable according to the predicted probabilities of the boosted regression trees, boxplots of the variable per ecoregion and, significant signs of the coefficient of bivariate logistic regressions per ecoregion

5.4. Discussion

5.4.1. Large-scale modelling

Both the multilevel and the BRT models have a very good predictive power. The distribution of zoonoses at a large-scale using a single database built from national databases is successfully modelled. However, large-scale models can obscure the local effect of environmental factors or how they may differ from place to place. With these models, the interpretation of the variables may generate misleading conclusion as wide range of conditions cannot be summarized into one explanation. We identify three scenarios of response to a variable that can be encountered in modelling studies at a large-scale (Figure 17). These scenarios are illustrated by variables from this study, but are relevant to other disease modelling studies.

5.4.1.1. Scenario a

The first easiest scenario implies that a variable has the same effect over the complete study area. These variables can be easily included in a large-scale model. In our study, the built-up areas in forest ecotones are assumed to increase the human exposure. This is reflected both by the ecoregional regressions and the global BRT model.

5.4.1.2. Scenario b

The second scenario, the most common, is when the variable has a nonlinear effect on the probability of the disease. Because of this, the effect of the variable varies from region to region, depending on the value of the variable encountered. For these variables, results and hypotheses from local studies can be used for places with the same variable extent. They can be used in nonlinear large-scale models. For example, looking at the snow cover BRT trend, less extensive snow cover (such as in southern Europe) increases the probability, as it relates to more contact between the virus/bank voles and humans; a more extensive snow cover increases the probability even more as that implies a larger bank voles population and a better ex-vivo virus survival; very extensive snow cover (such as in the extreme North of Europe, in Fennoscandian and Russian taiga) has a decreasing probability because of the harsh conditions for the bank voles and the increasing predator pressure due to lower prey diversity. The sign of the ecoregional regression matches these hypotheses as well as the BRT response trend.

5.4.1.3. Scenario c

The third scenario, the most difficult to identify, is when the variable does not seem to correspond to one hypothesis, whatever the extent of its distribution. Different hypotheses are needed for each region. For these variables, results from local studies cannot be applied and interpolated to other places and their interpretation in large-scale study

may lead to misleading conclusion. Taking the coniferous forest as an example, the global BRT trend is negative as is generally expected but, the sign of the regression coefficient appear significantly positive for the Northern temperate Atlantic and the Scandinavian montane birch ecoregions. In the Northern temperate Atlantic ecoregion, there are less forests and it is possible that bank voles are found in coniferous forests, lacking a better habitat. For the Scandinavian montane birch ecoregions, coniferous forests are in valleys while mountains are colonized by birches (Bryn et al. 2013). Soils on the slopes are drier and know important runoff. In those areas, the conditions in broadleaved forest are less favourable for the virus and rodents than in the coniferous forest. This is probably why coniferous forests appear to be positively related to NE presence in these regions. The effect of coniferous is thus not the result of one nonlinear process but can be explained by the local particularities of ecoregions. These particularities may results from the level of the other variables or from historical and cultural factors.

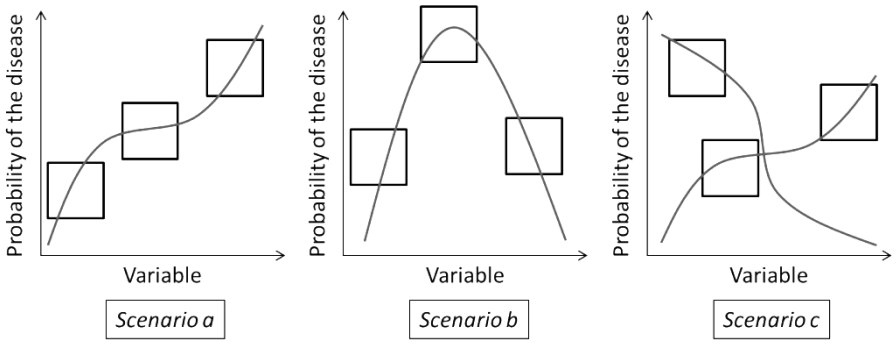


Figure 17: Three scenarios that can happen when modelling variables over a large study area, squared frames represent different subregions of a large study area

5.4.2. Model variables

So far, the distribution of NE cases has not been studied at the European level, and this study helps to answer whether factors impact the distribution in one region or at a multi-country level (see the question raised in (Heyman et al. 2012)). Our model highlights the new finding that climatic variables as well as the connectivity of forests act more at the regional level than at the landscape level. The multilevel logistic model shows that NE cases are mostly found in populated regions with connected forests, high EVI and low SWI. These regions have a low maximum temperature in the summer and a low minimum temperature in the winter (summers are mild and winters are cold). In these regions, landscapes with a higher proportion of built-up areas in forest ecotones and a lower minimum temperature in winter are expected to be more at risk. Model results are consistent with what has been already hypothesized on NE in Europe (Olsson et al. 2010, Heyman et al. 2012). Unexpected curves are nonetheless observed for EVI and green days but their relative importance is low and their effect complex to interpret (Lhermitte et al. 2011).

By combining ecoregional regressions and a global BRT model, we identified variables for which it is advised to pay attention when hypothesizing effects (scenario c): precipitations, coniferous forest, broadleaved forests and population. For the minimum temperature in winter, the snow cover, green days, the SWI, the proportion of mixed forest, the built-up areas in forest ecotones and the contiguity of

forests, our results suggest that they follow one process that can be linear or nonlinear.

5.4.3. Model predictions

The predicted probability map shows that, according to environmental data, fewer cases should be expected in southern, western, central and the extreme East of France, northern Belgium, the Netherlands, southern and northern Sweden and, northern Finland. The BRT model yields similar predictions. Also, a lot of false presences are found in areas where the disease is already recorded and false absences are located at the edge of these areas. If variables stay at their current values (this model predicts the current situation and not the future), false presences can be interpreted as suitable places to find the disease and where the disease may appear in the future while false absences can represent unsuitable environments. It seems that the recorded data already covered the potential distribution of NE cases. A spread of the disease should not occur but intensification is possible.

5.4.4. Limitation of the models

As we learned here, effects can vary between scales and, even if the landscape scale is interesting, a finer scale may bring further details and reflect more particularities. Exact geographical coordinates are rarely known or disclosed, due to privacy rules but also due to the nature of the hantavirus pathogenesis, and the size of the

administrative units may be too big to make proper models. NE cases are not reported in the same way everywhere (between and within countries). Cases can for example be identified from antibodies from serological surveys or by the identification of clinical symptoms. A bias may result from the difference of these reports. Inclusion of a variable accounting for these differences could solve this issue but is particularly challenging as distinction of the way of records is not always known. Cases are most often reported by residence or where the disease is identified, rather than the true location of infection. Therefore, some presences are reported in unexpected places, like in southern Sweden where PUUV is reportedly absent. These presences were nevertheless included in the database and were all predicted as absences. However, in this area, some false presences are predicted near these false absences. It could indicate that the recorded presences do indicate local transmission. The models presented here are not designed to predict the future distribution of PUUV but we advise that the confirmation of the absence or presence of cases must be assessed at the false presence locations (e.g. through serosurvey among exposed humans). In southern and central France, absence of confirmed case will probably results from the fact that places are too far and isolated from the current distribution to imagine that the disease will reach these places. The other false presences places like southern Sweden, the Netherlands and the Norwegian coast, even if located in lower probability areas, should be carefully considered in the future for monitoring and surveillance.

5.5. Conclusion

This study successfully models the pan-European NE distribution using a compilation of national disease data source and European level environmental databases. Europe covers diverse landscapes and ecoregions and, even if some variables recurrently appear in the literature for various areas, the underlying hypotheses can be different. This study highlights that these differences can be the result of local specificities or of the nonlinearity of the processes. When the effect of a factor varies locally, its inclusion in a large-scale model is compromised.

According to both multilevel logistic and nonlinear model, the distribution of PUUV seems limited in space by environmental variables that are identified here. Our model does not predict an important spread of the disease, assuming current conditions, but possibly intensification in places where the disease is already present. Also, some lower probabilities areas such as the South of Sweden, the Netherlands and the Norwegian inner coast should be further studied.

Chapter 6 Conclusion

In this conclusion, the main findings of the thesis are first reviewed and, then, the limitations of the modelling techniques and plans for further researches are explored.

6.1. Contributions of the thesis

The main contribution of this thesis may be classified in four categories. The first one is related to the creation of a new framework to investigate the risk of zoonoses. The second and third categories give recommandations on the different modelling tools and on the spatial scale issue. The fourth category reviews the important results in term of identification of factors affecting the spatial distribution of tick-borne encephalitis and hantavirus infections.

6.1.1. Hazard and exposure framework

The hazard and exposure framework presented in the introduction demonstrated its efficiency in identifying the effect of landscape factors in a more specific and interpretable fashion in Chapter 3. The hazard is defined as the pathogen transmission cycle in the wild and the exposure, as people entering infected landscape. We showed that applying this framework leads to better models, even if looking only at one component is sufficient to have a good predictive power.

Moreover, this framework helps to understand zoonosis risk by investigating the exposure and the hazard. These two components are too often studied separately. The geographer's perspective helps to have a broad study angle that covers diverse fields relevant to the application of the framework. Indeed, the hazard is better depicted by ecology and, more specifically, by landscape ecology, and exposure is often addressed by social sciences, such as in studies of touristic preferences and accessibility studies. All of these inputs are important in the study of zoonoses because they imply both human and wildlife. The work on zoonoses is "anthropocentric" as it focuses on diseases that affect humans. Nonetheless, before occurring in human population, the disease is first affecting animals. In Chapter 4, we had the chance to combine both human and bank vole data. It helped us to understand which variables are important at which step of the disease transmission system. In this chapter, we successfully depicted and unravelled the concept of hazard and exposure. We also investigated in depth how the environment can affect zoonoses differently according to the part of the system investigated.

In conclusion, applying the concept of hazard and exposure to shape the risk of zoonoses helps to do better models but also to better understand the disease transmission system. Focusing on one aspect may yield good results but the interpretation may be incomplete or fallacious. A unified approach between epidemiology and ecology is therefore important. Moreover, broadening the research horizon by looking at the touristic literature is strongly encouraged as most of zoonoses in Europe are caught during outdoors leisure time.

6.1.2. Modelling tools

With the help of our framework, relevant questions can be raised that may allow to reach conclusions meaningful in a broader context, including risk management. Doing models to make a distribution map is nowadays not enough. Advances in techniques, either spatial analyses or statistical modelling tools, allow going a step further and answer new questions. Laboratory and field researches are useful to discover a set of factors affecting the pathogen or the host but, with models, we can investigate the effect of other factors that cannot be reproduced in a lab or that cannot be controlled on the field. With empirical modelling and machine learning, the overall complexity of reality is approached by preliminary assumptions that are tested to meet specific goals. In this thesis, each study started on the demand of collaborators or on our own curiosity with the aim to better understand why the disease is there and not there.

From these questions, data are collected and potential candidate variables are identified through the literature. Considering these inputs and according to the study objectives, a choice of modelling technique(s) is needed. In Chapter 2, we saw particular methods that can be applied to human presence data. Each method shows pros and cons that can be more or less important or problematic depending on the desired output. The choice of the right modelling techniques is consequently not trivial. Getting familiar with the various tools may be necessary before reaching a decision of using a particular

modelling method and, if needed, to adapt them, e.g. by combining different approaches.

The usefulness of BRTs as explanatory models was also highlighted. BRT has for major advantage to approach the reality in its overall complexity and to produce nonlinear response curves. As it was discussed, these curves should be an important tool to interpret the global trends of the variables but not to quantify their impact on the probabilities.

6.1.3. Spatial scale issue

Spatial scale appears through this entire thesis as a special issue itself. For example, in Chapter 3 on tick-borne encephalitis in Sweden, the distance to the sea was used to represent the attractiveness of the landscape while in Chapter 2 on hantavirus in Sweden, the distance to the sea was considered at a larger scale and reflected climatic buffer effects. A same proxy variable, looking at different scale, can thus reflect a diversity of environmental and spatial effects.

Another effect of the spatial scale is identified in Chapter 4 on hantavirus in Germany. The probability map of the bank vole model was clustered in areas of high and low presence probabilities. The false absences and the false presences were respectively located in these areas. Variables included in the model were calculated at the landscape scale and these false presences and absences could consequently be explained by missing variable(s) acting at a regional

scale. Nonetheless, a landscape approach to study Germany seems appropriate as the country shows spatial complexity with a lot of environmental differences from one place to another.

Approaches combining different levels are therefore recommended. This is highlighted in Chapter 5 where local and regional scales were studied together. This chapter gave also recommendations for large-scale modelling by identifying three scenarios. The first one, in which a variable has the same effect on the entire study area, does not have particular consequence. The second one, in which a variable has a nonlinear effect, involves the use of nonlinear model. The last one, in which a variable responds to different processes according to the places, implies that the variable cannot be used for large-scale modelling, unless it is included at the proper local group in multilevel model.

6.1.4. Factors influencing the spatial distribution of zoonoses

Variable effects, as identified by BRT models, are summarized in Table 15. Chapter 3 revealed that, in Sweden, TBE was mostly found in accessible landscapes close to the sea which presented a diversity of land uses, numerous holiday houses, waterbodies, open areas in forest ecotones and, well-connected broadleaved and mixed forests with game species. Oak, birch and pine forests with a complex shape, numerous clear-cuts and tree heights variation were preferred over spruces. PCA revealed that roe deer are not found in the same areas as

the other major game species. The controversial hypothesized dilution effect related to this particular deer species is based on the negative relationship between roe deer and TBE. This may nevertheless result from any other unsuitable aspects of the environment where roe deer are found.

The TBE disease transmission system involves forest-dwelling species such as rodents. In Sweden, some factors are therefore comparable with hantavirus infections in the North. Nonetheless, in Chapter 2, interpretation for the distance to the sea differed and spruces were predicted as suitable for hantavirus. This can be explained because, in the North of Sweden, spruce forests are older, less managed and with more undergrowth while in the South, they are more heavily managed.

Hantavirus distribution in Europe was found in landscapes with built-up areas in forest ecotones that were located in regions densely populated with well-connected broadleaved and mixed forests, more intense vegetation activity, low soil water content, mild summers and cold winters. The study done in Germany also revealed that the wind speed should not be too important, going against our first idea that it should favours the indirect transmission process. Wind speed is probably more related to outdoors activities.

Table 15: Summary of variables responses in all models

		TBE in Sweden	Hanta Germany		Hanta Sweden	Hanta Europe
			Human case	Bank vole		
Climate	Precipitation		+	+		+ and -
	Relative humidity		mixed	-		
	Dry days		mixed	mixed		
	Warm days		+	-		
	Max T° in summer		+	-		+ and -
	min T° in winter		+	+		-
	Snow depth/cover		-	-	-	+
	Wind speed		-	+		
	Growing season length		+	+		mixed
	NDVI/EVI					-
Land use	Built-up areas		-	-		
	Agricultural areas		mixed	-		
	Waterbodies	+	-	-		
	Distance to water course	- and +				
	Distance to the sea	- and +			-	
	Water course				+	
	Peat bogs				-	
Forest	Forest	- and +			+	-
	Mixed forest	+	+	+		+
	Broadleaved forest	+	+	+		+
	Coniferous forest	-	-	mixed		-
	Oak	+				
	Birch	+				
	Pine	+			mixed	
	Spruce	-			+ and -	
Forest index	Forest shape index	+	Null	+	-	
	Forest proximity/contiguity index	+	+	mixed	+	+

		TBE in Sweden	Hanta Germany		Hanta Sweden	Hanta Europe
			Human case	Bank vole		
Forest index	Min distance between forests		-	-	Null	
	Shannon index	+				
	Clear-cuts	+				
	Mean trees height	-				
	Standard deviation of trees height	+				
	Open area in ecotone	- and +				
	Built-up in ecotone		-	-		+
Wildlife	Roe deer	-				
	Wildlife (red/fallow deer and wild boar)	- and +				
Soil	SWI		mixed	mixed		+
	Soil Texture		+	+		
Elevation	Elevation		+	+	-	
	Slope		+	+		
Human	Population		+		+	+
	Roads in forest	+				
	Holiday houses	+				
	Distance to holiday houses				-	
	Accessibility/ road	+			+	
	Worker in agriculture/forestry/fishery		-			

6.2. Models limitations

Limitations can result from each step of the modelling process: the availability of input data, the choice of the model complexity, the

nature of empirical models, the models evaluation and, the interpretation of the results.

First, models may be limited because of their nature but also because of the availability of the input data. In Chapter 1, we saw that presences are not always accurate and that real absences are complex to record. Also, explanatory variables are most often proxies that are chosen according to the availability of data and their resolution. Fortunately, spatial data are more and more freely available and their accuracy is improving thanks to, among else, remote sensing.

The second type of limitations comes from the modelling process. Models were first used to simplify the reality but are becoming more and more complex. Do we really need this complexity? According to Jiménez-Valverde (2008), complex methods are more suitable to predict the realized distribution while simpler methods only reflect the potential distribution of species. However, good model should be a balance of accuracy, transparency and flexibility (Keeling and Rohani 2008). With complex machine learning model, the transparency and flexibility is not fully completed. A good compromise must therefore be done between ease and complexity.

Third, the identification of potential variables included are often limited by the understanding of the processes occurring behind a spatial pattern and, empirical models based on presence and absence cannot give information on the abundance of the host or vector (even

if environments predicted as highly suitable probably correspond to an higher abundance) (Hess et al. 2002).

Four, AUC that were used as a measure of the predictive power of our models also know limitations (Lobo et al. 2008). AUC varies greatly according to the number and the extent of absences included in the model, it does not give information on the spatial distribution of the errors, it does not asses the goodness-of-fit of the model and, it equally considers false presence and false absence. Kappa statistics used in Chapter 2 is also controversial (Pontius Jr and Millones 2011). Model accuracy is therefore complex to measure. The level of accuracy required for a model will depend of the use of the model. When mapping the risk of a rare zoonosis, a certain level of accuracy is needed to avoid unrequired field prevention campaign while, when mapping the risk for a severe disease, false presences will be of great interest (Fielding and Bell 1997).

Finally, the model interpretation is often complex. Outcomes interactions can also be counterintuitive and unpredictable (Hess et al. 2002). Maps are often seen as pretty pictures and no more, while they contains complex digitalized biological knowledge (Hess et al. 2002).

To conclude this section, we can only agree with Box that "*Essentially, all models are wrong, but some are useful."* (Box 1979).

6.3. Further perspectives

This thesis brings its stone to the building of knowledge on zoonoses and spatial modelling. Further researches are still needed to complete the poor understanding that we have on wildlife, diseases and interactions between humans and environment. We hope that our framework may help to guide further researches and identify new relevant questions.

Interactions between fields are essential to yield results. Two examples of the benefits coming from partnerships of ongoing studies are presented here. The first one builds bridges between medical geography and public health and the second one, between medical geography and population genetics.

Public health and medical geography connections can be done at several stages of the research and in both directions. At an earlier stage, public health risk perception studies can identify research questions from people who are the most in need of information. Moreover, based on these studies, preliminary assumptions can be done for some variables. It can also help identifying factors that were not yet studied. The next step is the model building. Outputs are then mapped in order to reply to those who wanted to have the information. It can be public health practitioners and, in this case, results should be kept as accurate as possible with predictive model that can provide adaptive maps based, for example, on climatic data from the previous years. It can also be general public who is more asking for easy to

understand information. Maps can then be used as a communication tool and adapted to highlight what is asked. We are currently working on this for hantavirus infections in Germany with the help of the social scientist, Kerstin Dressel.

Medical geography has also much to bring to population genetic field and vice-versa. Some genes of the host are responsible for a greater tolerance to the disease, allowing the persistence of the pathogen in the organism and, consequently, in the time. These genes are selected through time and this selection is driven by several factors, including environmental variables. Landscape features can consequently be calculated and included in population genetic models. For example, the spatial distribution of NE in Europe is partially represented by the genetic distribution of bank vole *Myodes glareolus* (Vapalahti et al. 2003) and probably related to environmental variables. This is currently studied in collaboration with Nathalie Charbonnel.

These two studies are only two examples of many possible collaborations between different fields. The combination of multiple fields offers promising perspectives to address the complexity of zoonoses issues.

References

Ahlm, C., A. Thelin, F. Elgh, P. Juto, E. L. Stiernstrom, S. Holmberg & A. Tarnvik (1998) Prevalence of antibodies specific to Puumala virus among farmers in Sweden. *Scandinavian Journal of Work Environment & Health,* 24**,** 104-108.

Akaike, H. (1987) Factor analysis and AIC. *Psychometrika,* 52**,** 317-332.

Alexander, N. S. & W. Wint (2013) Projected Population Proximity Indices (30km) for 2005, 2030 & 2050. *Open Health Data,* 1**,** e2.

Allan, B. F., F. Keesing & R. S. Ostfeld (2003) Effect of forest fragmentation on Lyme disease risk. *Conservation Biology,* 17**,** 267-272.

Anderson, P. R., M. Dudík, S. Ferrier, A. Guisan, J. R. Hijmans, F. Huettmann, R. J. Leathwick, A. Lehmann, J. Li & G. L. Lohmann (2006) Novel methods improve prediction of species' distributions from occurrence data. *Ecography,* 29**,** 129-151.

Andreo, V., G. Glass, T. Shields, C. Provensal & J. Polop (2011) Modeling potential distribution of oligoryzomys longicaudatus, the andes virus (Genus: Hantavirus) reservoir, in Argentina. *EcoHealth,* 8**,** 332-348.

Bahn, V. & B. J. McGill (2013) Testing the predictive performance of distribution models. *Oikos,* 122**,** 321-331.

Barrios, J. M., W. W. Verstraeten, P. Maes, J. Clement, J. M. Aerts, S. A. Haredasht, J. Wambacq, K. Lagrou, G. Ducoffre, M. van Ranst, D. Berckmans & P. Coppin (2010) Satellite derived forest phenology and its relation with nephropathia epidemica in Belgium. *International Journal of Environmental Research and Public Health,* 7**,** 2486-2500.

Beale, C. M., J. J. Lennon, J. M. Yearsley, M. J. Brewer & D. A. Elston (2010) Regression analysis of spatial data. *Ecology letters,* 13**,** 246-264.

Bishop, D., C. Calisher, J. Casals, M. Chumakov, S. Gaidamovich, C. Hannoun, D. Lvov, I. Marshall, N. Okerblom, R. Pettersson, J. Porterfield, P. Russell, R. Shope & E. Westaway (1980) BUNYAVIRIDAE. *Intervirology,* 14**,** 125-143.

Bonada, N., S. Dolédec & B. Statzner (2012) Spatial autocorrelation patterns of stream invertebrates: exogenous and endogenous factors. *Journal of Biogeography,* 39**,** 56-68.

Boone, I., C. Wagner-Wiening, D. Reil, J. Jacob, U. M. Rosenfeld, R. G. Ulrich, D. Lohr & G. Pfaff (2012) Rise in the number of notified human hantavirus infections since October 2011 in Baden-Württemberg, Germany. *Eurosurveillance,* 17.

Bormane, A., I. Lucenko, A. Duks, V. Mavtchoutko, R. Ranka, K. Salmina & V. Baumanis (2004) Vectors of tick-borne diseases and epidemiological situation in Latvia in 1993-2002. *International Journal of Medical Microbiology, Supplement,* 293**,** 36-47.

Bostedt, G. & L. Mattsson (1995) The value of forests for tourism in Sweden. *Annals of Tourism Research,* 22**,** 671-680.

Box, G. E. 1979. Robustness in the strategy of scientific model building. In *Robustness in Statistics,* ed. R. L. L. a. G. N. Wilkinson. New York: Academic Press.

Boyce, M., P. Vernier, S. Nielsen & F. Schmiegelow (2002) Evaluating resource selection functions. *Ecological Modelling,* 157**,** 281-300.

Breiman, L. (2001) Statistical modeling: The two cultures (with comments and a rejoinder by the author). *Statistical Science,* 16**,** 199-231.

Brotons, L., W. Thuiller, M. Araujo & A. Hirzel (2004) Presence-absence versus presence-only modelling methods for predicting bird habitat suitability. *Ecography,* 27**,** 437-448.

Brownstein, J. S., T. R. Holford & D. Fish (2003) A climate-based model predicts the spatial distribution of the Lyme disease vector Ixodes scapularis in the United States. *Environmental health perspectives,* 111**,** 1152.

Brownstein, J. S., D. K. Skelly, T. R. Holford & D. Fish (2005) Forest fragmentation predicts local scale heterogeneity of Lyme disease risk. *Oecologia,* 146**,** 469-475.

Brummer-Korvenkontio, M., O. Vapalahti, H. Henttonen, P. Koskela, P. Kuusisto & A. Vaheri (1999) Epidemiological study of nephropathia epidemica in Finland 1989-96. *Scandinavian Journal of Infectious Diseases,* 31**,** 427-435.

Bryn, A., P. Dourojeanni, L. Ø. Hemsing & S. O'Donnell (2013) A high-resolution GIS null model of potential forest expansion

following land use changes in Norway. *Scandinavian Journal of Forest Research,* 28**,** 81-98.

Cagnacci, F., L. Bolzoni, R. Rosà, G. Carpi, H. C. Hauffe, M. Valent, V. Tagliapietra, M. Kazimirova, J. Koci, M. Stanko, M. Lukan, H. Henttonen & A. Rizzoli (2012) Effects of deer density on tick infestation of rodents and the hazard of tick-borne encephalitis. I: Empirical assessment. *International Journal for Parasitology,* 42**,** 365-372.

Carbajo, A. & U. Pardiñas (2007) Spatial distribution model of a hantavirus reservoir, the long-tailed colilargo (Oligoryzomys longicaudatus), in Argentina. *Journal of Mammalogy,* 88**,** 1555-1568.

Carlsson, N. O., J. M. Jeschke, N. Holmqvist & J. Kindberg (2010) Long-term data on invaders: when the fox is away, the mink will play. *Biological Invasions,* 12**,** 633-641.

Cavard, X., S. E. Macdonald, Y. Bergeron & H. Y. H. Chen (2011) Importance of mixedwoods for biodiversity conservation: Evidence for understory plants, songbirds, soil fauna, and ectomycorrhizae in northern forests. *Environmental Reviews,* 19**,** 142-161.

Chetkiewicz, C.-L., C. St. Clair & M. Boyce (2006) Corridors for Conservation: Integrating Pattern and Process. *Annual Review of Ecology, Evolution and Systematics,* 37**,** 317-342.

Clement, J., N. Lameire, E. Keyaerts, P. Maes & M. Van Ranst (2003) Hantavirus infections in Europe. *Lancet Infectious Diseases,* 3**,** 752-753.

Cleve, C., J. Perrine, B. Holzman & E. Hines (2011) Addressing biased occurrence data in predicting potential Sierra Nevada red fox habitat for survey prioritization. *Endangered Species Research,* 14**,** 179-191.

Cortinas, M. R., M. A. Guerra, C. J. Jones & U. Kitron (2002) Detection, characterization, and prediction of tick-borne disease foci. *International Journal of Medical Microbiology,* 291**,** 11-20.

Crooks, E. & S. E. Randolph (2006) Walking by Ixodes ricinus ticks: intrinsic and extrinsic factors determine the attraction of moisture or host odour. *Journal of Experimental Biology,* 209**,** 2138-2142.

Crowcroft, N. S., A. Infuso, D. Lief, B. Le Guenno, J. C. Desenclos, F. Van Loock & J. Clé Ment (1999) Risk factors for human

hantavirus infection: Franco-Belgian collaborative case-control study during 1995-6 epidemic. *British Medical Journal,* 318**,** 1737-1738.

Daniel, M., J. Kolár, P. Zeman, K. Pavelka & J. Sádlo (1998) Predictive map of Ixodes ricinus high-incidence habitats and a tick-borne encephalitis risk assessment using satellite data. *Experimental & applied acarology,* 22**,** 417-433.

Daszak, P., A. A. Cunningham & A. D. Hyatt (2000) Emerging infectious diseases of wildlife - Threats to biodiversity and human health. *Science,* 287**,** 443-449.

Dearing, M. & L. Dizney (2010) Ecology of hantavirus in a changing world. *Year in Ecology and Conservation Biology 2010,* 1195**,** 99-112.

Despommier, D., B. R. Ellis & B. A. Wilcox (2006) The role of ecotones in emerging infectious diseases. *EcoHealth,* 3**,** 281-289.

Donalisio, M. & A. Peterson (2011) Environmental factors affecting transmission risk for hantaviruses in forested portions of southern Brazil. *Acta Tropica,* 119**,** 125-130.

Dormann, C. F. (2007) Effects of incorporating spatial autocorrelation into the analysis of species distribution data. *Global Ecology and Biogeography,* 16**,** 129-138.

Dormann, C. F., J. Elith, S. Bacher, C. Buchmann, G. Carl, G. Carré, J. R. G. Marquéz, B. Gruber, B. Lafourcade, P. J. Leitão, T. Münkemüller, C. McClean, P. E. Osborne, B. Reineking, B. Schröder, A. K. Skidmore, D. Zurell & S. Lautenbach (2013) Collinearity: A review of methods to deal with it and a simulation study evaluating their performance. *Ecography,* 36**,** 027-046.

Dumpis, U., D. Crook & J. Oksi (1999) Tick-borne encephalitis. *Clinical Infectious Diseases,* 28**,** 882-890.

Edwards, D. M., M. Jay, F. S. Jensen, B. Lucas, M. Marzano, C. Montagné, A. Peace & G. Weiss (2012) Public preferences across Europe for different forest stand types as sites for recreation. *Ecology and Society,* 17.

Eisen, L., R. Eisen & R. Lane (2006) Geographical distribution patterns and habitat suitability models for presence of host-seeking ixodid ticks in dense woodlands of Mendocino County, California. *Journal of Medical Entomology,* 43**,** 415-427.

Elith, J., H. Graham, P. Anderson, M. Dudík, S. Ferrier, A. Guisan, J. Hijmans, F. Huettmann, R. Leathwick, A. Lehmann, J. Li, G. Lohmann, A. Loiselle, G. Manion, C. Moritz, M. Nakamura, Y. Nakazawa, M. McC.Overton, A. Townsend Peterson, J. Phillips, K. Richardson, R. Scachetti-Pereira, E. Schapire, J. Soberón, S. Williams, S. Wisz & E. Zimmermann (2006) Novel methods improve prediction of species' distributions from occurrence data. *Ecography,* 29**,** 129-151.

Elith, J. & J. Leathwick. 2014. Boosted Regression Trees for ecological modeling.

Elith, J., J. Leathwick & T. Hastie (2008) A working guide to boosted regression trees. *Journal of Animal Ecology,* 77**,** 802-813.

Elith, J., S. Phillips, T. Hastie, M. Dudik, Y. Chee & C. Yates (2011) A statistical explanation of MaxEnt for ecologists. *Diversity and Distributions,* 17**,** 43-57.

Essbauer, S. S., J. Schmidt-Chanasit, E. L. Madeja, W. Wegener, R. Friedrich, R. Petraityte, K. Sasnauskas, J. Jacob, J. Koch, G. Dobler, F. J. Conraths, M. Pfeffer, C. Pitra & R. G. Ulrich (2007) Nephropathia epidemica in metropolitan area, Germany [12]. *Emerging Infectious Diseases,* 13**,** 1271-1273.

Estrada-Peña, A. (2001) Distribution, abundance, and habitat preferences of Ixodes ricinus (Acari: Ixodidae) in northern Spain. *Journal of Medical Entomology,* 38**,** 361-370.

Evander, M. & C. Ahlm (2009) Milder winters in northern Scandinavia may contribute to larger outbreaks of haemorrhagic fever virus. *Global Health Action,* 2**,** 2020.

Ewers, R. & R. Didham (2006) Confounding factors in the detection of species responses to habitat fragmentation. *Biological Reviews,* 81**,** 117-142.

Faber, M., T. Wollny, M. Schlegel, K. M. Wanka, J. Thiel, C. Frank, D. Rimek, R. G. Ulrich & K. Stark (2013) Puumala Virus Outbreak in Western Thuringia, Germany, 2010: Epidemiology and Strain Identification. *Zoonoses and Public Health,* 60**,** 549-554.

Faber, M. S., R. G. Ulrich, C. Frank, S. O. Brockmann, G. M. Pfaff, J. Jacob, D. H. Krüger & K. Stark (2010) Steep rise in notified hantavirus infections in Germany, April 2010. *Eurosurveillance,* 15.

FAO. 1999. Principles and guidelines for the conduct of microbiological risk assessment. Codex Alimentarius: Food and hygiene basic texts. ed. F. a. A. O. o. t. U. Nations. Rome.

Fauci, A. S. (2005) Emerging and reemerging infectious diseases: The perpetual challenge. *Academic Medicine,* 80**,** 1079-1085.

Fawcett, T. (2006) An introduction to ROC analysis. *Pattern Recognition Letters,* 27**,** 861-874.

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

Forman, R. 1995. Land Mosaics: The ecology of landscapes and regions. In *Cambridge University Press*. Cambridge.

Gastón, A. & J. García-Viñas (2011) Modelling species distributions with penalised logistic regressions: A comparison with maximum entropy models. *Ecological Modelling,* 222**,** 2037-2041.

Gray, J. S. (2002) Biology of Ixodes species ticks in relation to tick-borne zoonoses. *Wiener Klinische Wochenschrift,* 114**,** 473-478.

Groen, J., M. Gerding, J. Jordans, J. Clement, J. Nieuwenhuijs & A. Osterhaus (1995) Hantavirus infections in The Netherlands: epidemiology and disease. *Epidemiology and Infection,* 114**,** 373-383.

Guerra, M., E. Walker, C. Jones, S. Paskewitz, M. R. Cortinas, A. Stancil, L. Beck, M. Bobo & U. Kitron (2002) Predicting the Risk of Lyme Disease: Habitat Suitability for *Ixodes scapularis* in the North Central United States.

Guisan, A. & N. E. Zimmermann (2000) Predictive habitat distribution models in ecology. *Ecological Modelling,* 135**,** 147-186.

Guivier, E., M. Galan, Y. Chaval, A. Xuéreb, A. Ribas Salvador, M. L. Poulle, L. Voutilainen, H. Henttonen, N. Charbonnel & J. F. Cosson (2011) Landscape genetics highlights the role of bank vole metapopulation dynamics in the epidemiology of Puumala hantavirus. *Molecular Ecology,* 20**,** 3569-3583.

Gurnell, J. (1993) Tree seed production and food conditions for rodents in an oak wood in Southern England. *Forestry,* 66**,** 291-315.

Gustafson, R., M. Forsgren, A. Gardulf, M. Granström & B. Svenungsson (1993) Antibody prevalence and clinical

manifestations of Lyme borreliosis and tick-borne encephalitis in Swedish orienteers. *Scandinavian Journal of Infectious Diseases,* 25**,** 605-611.

Haglund, M., M. Forsgren, G. Lindh & L. Lindquist (1996) A 10-year follow-up study of tick-borne encephalitis in the Stockholm area and a review of the literature: need for a vaccination strategy. *Scandinavian Journal of Infectious Diseases,* 28**,** 217-224.

Hanski, I., L. Hansson & H. Henttonen (1991) SPECIALIST PREDATORS, GENERALIST PREDATORS, AND THE MICROTINE RODENT CYCLE. *Journal of Animal Ecology,* 60**,** 353-367.

Hanski, I., H. Henttonen, E. Korpimaki, L. Oksanen & P. Turchin (2001) Small-rodent dynamics and predation. *Ecology,* 82**,** 1505-1520.

Hansson, L. (1999) Intraspecific variation in dynamics: Small rodents between food and predation in changing landscapes. *Oikos,* 86**,** 159-169.

Hansson, L. & H. Henttonen (1985) GRADIENTS IN DENSITY VARIATIONS OF SMALL RODENTS - THE IMPORTANCE OF LATITUDE AND SNOW COVER. *Oecologia,* 67**,** 394-402.

Hansson, L. & H. Henttonen (1988) RODENT DYNAMICS AS COMMUNITY PROCESSES. *Trends in Ecology & Evolution,* 3**,** 195-200.

Hart, C. & M. Bennett (1999) Hantavirus infections: epidemiology and pathogenesis. *Microbes and Infection,* 1**,** 1229-1237.

Hess, G. R., S. E. Randolph, P. Arneberg, C. Chemini, C. Furlanello, J. Harwood, M. G. Roberts & J. Swinton. 2002. Spatial Aspects of Disease Dynamics. In *The ecology of wildlife diseases,* ed. O. U. P. Oxford, 102-118.

Heyman, P., C. Ceianu, I. Christova, N. Tordo, M. Beersma, M. Alves, A. Lundkvist, M. Hukic, A. Papa, A. Tenorio, H. Zelena, S. Essbauer, I. Visontai, I. Golovljova, J. Connell, L. Nicoletti, M. Van Esbroeck, S. Dudman, S. Aberle, T. Avsic-Zupanc, G. Korukluoglu, A. Nowakowska, B. Klempa, R. Ulrich, S. Bino, O. Engler, M. Opp & A. Vaheri (2011) A five-year perspective on the situation of haemorrhagic fever with renal syndrome and status of the hantavirus reservoirs in Europe, 2005-2010. *Eurosurveillance,* 16**,** 15-22.

Heyman, P., B. R. Thoma, J. L. Marié, C. Cochez & S. S. Essbauer (2012) In search for factors that drive hantavirus epidemics. *Frontiers in Physiology,* 3 JUL**,** 237.

Hosmer, D. & S. Lemeshow. 1989. *Applied logistic regression*. New York: Wiley.

Imholt, C., D. Reil, J. A. Eccard, D. Jacob, N. Hempelmann & J. Jacob (2014) Quantifying the past and future impact of climate on outbreak patterns of bank voles (Myodes glareolus). *Pest Management Science*.

Jaenson, T. G. T., L. Eisen, P. Comstedt, H. A. Mejlon, E. Lindgren, S. BergstrÖm & B. Olsen (2009) Risk indicators for the tick Ixodes ricinus and Borrelia burgdorferi sensu lato in Sweden. *Medical and Veterinary Entomology,* 23**,** 226-237.

Jaenson, T. G. T., M. Hjertqvist, T. Bergström & Å. Lundkvist (2012a) Why is tick-borne encephalitis increasing? A review of the key factors causing the increasing incidence of human TBE in Swedena. *Parasites and Vectors,* 5.

Jaenson, T. G. T., D. G. E. Jaenson, L. Eisen, E. Petersson & E. Lindgren (2012b) Changes in the geographical distribution and abundance of the tick Ixodes ricinus during the past 30 years in Sweden. *Parasites and Vectors,* 5.

Jimenez-Valverde, A., J. Lobo & J. Hortal (2008) Not as good as they seem: the importance of concepts in species distribution modelling. *Diversity and Distributions,* 14**,** 885-890.

Johnson, C., S. Nielsen, E. Merrill, T. McDonald & M. Boyce (2006) Resource selection functions based on use-availability data: Theoretical motivation and evaluation methods. *Journal of Wildlife Management,* 70**,** 347-357.

Jones, K. E., N. G. Patel, M. A. Levy, A. Storeygard, D. Balk, J. L. Gittleman & P. Daszak (2008) Global trends in emerging infectious diseases. *Nature,* 451**,** 990-993.

Jore, S., S. O. Vanwambeke, H. Viljugrein, K. Isaksen, A. B. Kristoffersen, Z. Woldehiwet, B. Johansen, E. Brun, H. Brun-Hansen & S. Westermann (2014) Climate and environmental change drives Ixodes ricinus geographical expansion at the northern range margin. *Parasit Vectors,* 7.

Kallio, E., J. Klingstrom, E. Gustafsson, T. Manni, A. Vaheri, H. Henttonen, O. Vapalahti & A. Lundkvist (2006) Prolonged survival of Puumala hantavirus outside the host: evidence for

indirect transmission via the environment. *Journal of General Virology,* 87**,** 2127-2134.

Keeling, M. J. & P. Rohani. 2008. *Modeling infectious diseases in humans and animals*. Princeton University Press.

Kirkwood, B. & J. Sterne. 2003. *Essential Medical Statistics*. Oxford: Blackwell Scientific.

Knoke, T., C. Ammer, B. Stimm & R. Mosandl (2008) Admixing broadleaved to coniferous tree species: A review on yield, ecological stability and economics. *European Journal of Forest Research,* 127**,** 89-101.

Korslund, L. & H. Steen (2006) Small rodent winter survival: Snow conditions limit access to food resources. *Journal of Animal Ecology,* 75**,** 156-166.

Kozakiewicz, M., R. Van Apeldoorn, P. Bergers, T. Gortat & A. Kozakiewicz (2000) Landscape approach to bank vole ecology. *Polish Journal of Ecology,* 48**,** 149-161.

Kriz, B., M. Daniel, C. Benes & M. Maly (2014) The Role of Game (Wild Boar and Roe Deer) in the Spread of Tick-Borne Encephalitis in the Czech Republic. *Vector-Borne and Zoonotic Diseases,* 14**,** 801-807.

Kruse, H., A. M. Kirkemo & K. Handeland (2004) Wildlife as source of zoonotic infections. *Emerging Infectious Diseases,* 10**,** 2067-2072.

La Manna, L., S. Matteucci & T. Kitzberger (2011) Modelling Phytophthora disease risk in Austrocedrus chilensis forests of Patagonia. *European Journal of Forest Research***,** 1-15.

Labuda, M., O. Kozuch, E. Zuffová, E. Elecková, R. S. Hails & P. A. Nuttall (1997) Tick-borne encephalitis virus transmission between ticks cofeeding on specific immune natural rodent hosts. *Virology,* 235**,** 138-143.

Lambin, E., A. Tran, S. Vanwambeke, C. Linard & V. Soti (2010) Pathogenic landscapes: Interactions between land, people, disease vectors, and their animal hosts. *International Journal of Health Geographics,* 9.

Leduc, J. (1989) EPIDEMIOLOGY OF HEMORRHAGIC-FEVER VIRUSES. *Reviews of Infectious Diseases,* 11**,** S730-S735.

Lhermitte, S., J. Verbesselt, W. W. Verstraeten & P. Coppin (2011) A comparison of time series similarity measures for classification and change detection of ecosystem dynamics. *Remote Sensing of Environment,* 115**,** 3129-3152.

Linard, C., P. Lamarque, P. Heyman, G. Ducoffre, V. Luyasu, K. Tersago, S. O. Vanwambeke & E. F. Lambin (2007a) Determinants of the geographic distribution of Puumala virus and Lyme borreliosis infections in Belgium. *International Journal of Health Geographics,* 6**,** 20.

Linard, C., K. Tersago, H. Leirs & E. F. Lambin (2007b) Environmental conditions and Puumala virus transmission in Belgium. *International Journal of Health Geographics,* 6**,** 55.

Lindgren, E., L. Tälleklint & T. Polfeldt (2000) Impact of climatic change on the northern latitude limit and population density of the disease-transmitting European tick Ixodes ricinus. *Environmental health perspectives,* 108**,** 119.

Lindhjem, H. (2007) 20 years of stated preference valuation of non-timber benefits from Fennoscandian forests: A meta-analysis. *Journal of Forest Economics,* 12**,** 251-277.

Lindström, A. & T. G. T. Jaenson (2003) Distribution of the common tick, Ixodes ricinus (Acari: Ixodidae), in different vegetation types in southern Sweden. *Journal of Medical Entomology,* 40**,** 375-378.

Lobo, J., A. Jimenez-Valverde & J. Hortal (2010) The uncertain nature of absences and their importance in species distribution modelling. *Ecography,* 33**,** 103-114.

Lobo, J. M., A. Jiménez-valverde & R. Real (2008) AUC: A misleading measure of the performance of predictive distribution models. *Global Ecology and Biogeography,* 17**,** 145-151.

Lundkvist, A., A. Wallensten, S. Vene & M. Hjertqvist (2011) Tick-borne encephalitis increasing in Sweden, 2011. *Eurosurveillance,* 16.

Maas, C. J. M. & J. J. Hox (2005) Sufficient sample sizes for multilevel modeling. *Methodology,* 1**,** 86-92.

Marano, N., P. Arguin, M. Pappaioanou, B. Chomel, E. Schelling, V. Martin, J. C. Butler, C. B. Beard & L. King (2006) International attention for zoonotic infections. *Emerging Infectious Diseases,* 12**,** 1813-1815.

McCallum, H. 2008. Infectious disease ecology : the effects of ecosystems on disease and of disease on ecosystems. In *Landscape Structure, Disturbance, and Disease Dynamics,* ed. P. U. Press, 100-122. Princeton, New Jersey.

McCullagh, P. & J. A. Nelder. 1989. *Generalized linear models*. London ; New York: Chapman and Hall.

Medlock, J. M., K. M. Hansford, A. Bormane, M. Derdakova, A. Estrada-Peña, J. C. George, I. Golovljova, T. G. T. Jaenson, J. K. Jensen, P. M. Jensen, M. Kazimirova, J. A. Oteo, A. Papa, K. Pfister, O. Plantard, S. E. Randolph, A. Rizzoli, M. M. Santos-Silva, H. Sprong, L. Vial, G. Hendrickx, H. Zeller & W. Van Bortel (2013) Driving forces for changes in geographical distribution of Ixodes ricinus ticks in Europe. *Parasites and Vectors,* 6.

Merlo, J., B. Chaix, M. Yang, J. Lynch & L. Råstam (2005) A brief conceptual tutorial of multilevel analysis in social epidemiology: Linking the statistical concept of clustering to the idea of contextual phenomenon. *Journal of Epidemiology and Community Health,* 59**,** 443-449.

Müller, D. K. (2006) The attractiveness of second home areas in Sweden: A quantitative analysis. *Current Issues in Tourism,* 9**,** 335-350.

Nelson, G. S. (1960) Schistosome infections as zoonoses in Africa. *Transactions of the Royal Society of Tropical Medicine and Hygiene,* 54**,** 301-316.

Nielsen, A. B., E. Heyman & G. Richnau (2012) Liked, disliked and unseen forest attributes: Relation to modes of viewing and cognitive constructs. *Journal of Environmental Management,* 113**,** 456-466.

Olsson, G., F. Dalerum, B. Hornfeldt, F. Elgh, T. Palo, P. Juto & C. Ahlm (2003) Human hantavirus infections, Sweden. *Emerging Infectious Diseases,* 9**,** 1395-1401.

Olsson, G., H. Leirs & H. Henttonen (2010) Hantaviruses and Their Hosts in Europe: Reservoirs Here and There, But Not Everywhere? *Vector-Borne and Zoonotic Diseases,* 10**,** 549-561.

Olsson, G., N. White, J. Hjalten & C. Ahlm (2005) Habitat factors associated with bank voles (Clethrionomys glareolus) and concomitant hantavirus in northern Sweden. *Vector-Borne and Zoonotic Diseases,* 5**,** 315-323.

Ostfeld, R., G. Glass & F. Keesing (2005) Spatial epidemiology: an emerging (or re-emerging) discipline. *Trends in Ecology & Evolution,* 20**,** 328-336.

Panagos, P., M. Van Liedekerke, A. Jones & L. Montanarella (2012) European Soil Data Centre: Response to European policy support and public data requirements. *Land Use Policy,* 29**,** 329-338.

Pangrácová, L., M. Derdáková, L. Pekárik, I. Hviščová, B. Víchová, M. Stanko, H. Hlavatá & B. Peťko (2013) Ixodes ricinus abundance and its infection with the tick-borne pathogens in urban and suburban areas of Eastern Slovakia. *Parasites & Vectors,* 6**,** 238.

Parola, P. & D. Raoult (2001) Ticks and tickborne bacterial diseases in humans: An emerging infectious threat. *Clinical Infectious Diseases,* 32**,** 897-928.

Patz, J. A., T. K. Graczyk, N. Geller & A. Y. Vittor (2000) Effects of environmental change on emerging parasitic diseases. *International Journal for Parasitology,* 30**,** 1395-1405.

Pearce, J. & S. Ferrier (2000) Evaluating the predictive performance of habitat models developed using logistic regression. *Ecological Modelling,* 133**,** 225-245.

Phillips, S. (2012) A brief tutorial on Maxent. *Lessons in Conservation,* 3**,** 107-135.

Phillips, S., R. Anderson & R. Schapire (2006) Maximum entropy modeling of species geographic distributions. *Ecological Modelling,* 190**,** 231-259.

Piechotowski, I., S. Brockmann, C. Schwarz, C. Winter, U. Ranft & G. Pfaff (2008) Emergence of hantavirus in South Germany: rodents, climate and human infections. *Parasitology Research,* 103**,** S131-S137.

Pontius Jr, R. G. & M. Millones (2011) Death to Kappa: birth of quantity disagreement and allocation disagreement for accuracy assessment. *International Journal of Remote Sensing,* 32**,** 4407-4429.

Randolph, S. E. (2008) Tick-borne encephalitis virus, ticks and humans: Short-term and long-term dynamics. *Current Opinion in Infectious Diseases,* 21**,** 462-467.

Randolph, S. E. (2011) Transmission of tick-borne pathogens between co-feeding ticks: Milan Labuda's enduring paradigm. *Ticks and Tick-borne Diseases,* 2**,** 179-182.

Randolph, S. E., L. Gern & P. A. Nuttall (1996) Co-feeding ticks: Epidemiological significance for tick-borne pathogen transmission. *Parasitology Today,* 12**,** 472-479.

Randolph, S. E., D. Miklisová, J. Lysy, D. J. Rogers & M. Labuda (1999) Incidence from coincidence: Patterns of tick infestations on rodents facilitate transmission of tick-borne encephalitis virus. *Parasitology,* 118**,** 177-186.

Randolph, S. E. & K. Storey (1999) Impact of microclimate on immature tick-rodent host interactions (Acari: Ixodidae): Implications for parasite transmission. *Journal of Medical Entomology,* 36**,** 741-748.

Ridgeway, G. (2007) Generalized Boosted Models: A guide to the gbm package. *Update,* 1.

Rizzoli, A., H. C. Hauffe, V. Tagliapietra, M. Neteler & R. Rosà (2009) Forest structure and roe deer abundance predict tick-borne encephalitis risk in Italy. *PLoS ONE,* 4**,** e4336.

Robertson, J. N., J. S. Gray & P. Stewart (2000) Tick bite and Lyme borreliosis risk at a recreational site in England. *European Journal of Epidemiology,* 16**,** 647-652.

Robinson, W. S. (2009) Ecological correlations and the behavior of individuals. *International journal of epidemiology,* 38**,** 337-341.

Rodgers, S. E., C. P. Zolnik & T. N. Mather (2007) Duration of exposure to suboptimal atmospheric moisture affects nymphal blacklegged tick survival. *Journal of Medical Entomology,* 44**,** 372-375.

Rogers, D. 2006. Models for vectors and vector-borne diseases. In *Advances in Parasitology, Vol 62: Global Mapping of Infectious Diseases: Methods, Examples and Emerging Applications,* eds. S. I. Hay, A. Graham & D. J. Rogers, 1-35. San Diego: Elsevier Academic Press Inc.

Rosà, R., A. Pugliese, R. Norman & P. J. Hudson (2003) Thresholds for disease persistence in models for tick-borne infections including non-viraemic transmission, extended feeding and tick aggregation. *Journal of Theoretical Biology,* 224**,** 359-376.

Sauvage, F., M. Langlais & D. Pontier (2007) Predicting the emergence of human hantavirus disease using a combination of viral dynamics and rodent demographic patterns. *Epidemiology and Infection,* 135**,** 46-56.

Sauvage, F., M. Langlais, N. Yoccoz & D. Pontier (2003) Modelling hantavirus in fluctuating populations of bank voles: the role of

indirect transmission on virus persistence. *Journal of Animal Ecology,* 72**,** 1-13.

Schraml, U. & K.-R. Volz. 2009. Do species matter? Valuable broadleaves as an object of public perception and policy. In *Valuable Broadleaved Forests in Europe.,* ed. H. Spiecker, 213–236. S. Brill, Leiden, Boston, Köln.

Schriewer, K. (1998) Die Wahrnehmung des Waldes im Wandel. *Vokus,* 2**,** 4-17.

Schwarz, A. C., U. Ranft, I. Piechotowski, J. E. Childs & S. O. Brockmann (2009) Risk factors for human infection with Puumala virus, southwestern Germany. *Emerging Infectious Diseases,* 15**,** 1032-1039.

Shannon, C. & W. Weaver. 1949. *The mathematical theory of communication*. Urbana.

Shmueli, G. (2010) To explain or to predict? *Statistical Science***,** 289-310.

Silvennoinen, H., J. Alho, O. Kolehmainen & T. Pukkala (2001) Prediction models of landscape preferences at the forest stand level. *Landscape and Urban Planning,* 56**,** 11-20.

Snijders, T. A. B. & R. J. Boskers. 1999. *Multilevel Analysis. An Introduction to Basic and Advanced Multilevel Modeling*. London.

Soulsby, E. J. J. 1974. Parasitic zoonoses, clinical and experimental studies. ed. A. Press. New York.

Stefanoff, P., M. Rosinska, S. Samuels, D. J. White, D. L. Morse & S. E. Randolph (2012) A National Case-Control Study Identifies Human Socio-Economic Status and Activities as Risk Factors for Tick-Borne Encephalitis in Poland. *PLoS ONE,* 7**,** e45511.

Šumilo, D., A. Bormane, L. Asokliene, I. Lucenko, V. Vasilenko & S. Randolph (2006) Tick-borne encephalitis in the Baltic States: identifying risk factors in space and time. *International Journal of Medical Microbiology,* 296**,** 76-79.

Šumilo, D., A. Bormane, L. Asokliene, V. Vasilenko, I. Golovljova, T. Avsic-Zupanc, Z. Hubalek & S. E. Randolph (2008) Socio-economic factors in the differential upsurge of tick-borne encephalitis in central and Eastern Europe. *Reviews in Medical Virology,* 18**,** 81-95.

Süss, J., C. Schrader, U. Falk & N. Wohanka (2004) Tick-borne encephalitis (TBE) in Germany—epidemiological data, development of risk areas and virus prevalence in field-

collected ticks and in ticks removed from humans. *International Journal of Medical Microbiology Supplements,* 293**,** 69-79.

Tack, W., M. Madder, L. Baeten, P. De Frenne & K. Verheyen (2012) The abundance of Ixodes ricinus ticks depends on tree species composition and shrub cover. *Parasitology,* 139**,** 1273-1281.

Tsoar, A., O. Allouche, O. Steinitz, D. Rotem & R. Kadmon (2007) A comparative evaluation of presence-only methods for modelling species distribution. *Diversity and Distributions,* 13**,** 397-405.

Vaheri, A., H. Henttonen, L. Voutilainen, J. Mustonen, T. Sironen & O. Vapalahti (2012) Hantavirus infections in Europe and their impact on public health. *Reviews in Medical Virology***,** n/a-n/a.

Vaheri, A., H. Henttonen, L. Voutilainen, J. Mustonen, T. Sironen & O. Vapalahti (2013) Hantavirus infections in Europe and their impact on public health. *Reviews in Medical Virology,* 23**,** 35-49.

Vanapeldoorn, R., W. Oostenbrink, A. Vanwinden & F. Vanderzee (1992) Effects of Habitat Fragmentation on the Bank Vole, Clethrionomys-Glareolus, in an Agricultural Landscape. *Oikos,* 65**,** 265-274.

Vanwambeke, S. O., D. Šumilo, A. Bormane, E. F. Lambin & S. E. Randolph (2010) Landscape predictors of tick-borne encephalitis in Latvia: Land cover, land use, and land ownership. *Vector-Borne and Zoonotic Diseases,* 10**,** 497-506.

Vapalahti, K., M. Paunio, M. Brummer-Korvenkontio, A. Vaheri & O. Vapalahti (1999) Puumala virus infections in Finland: Increased occupational risk for farmers. *American Journal of Epidemiology,* 149**,** 1142-1151.

Vapalahti, O., J. Mustonen, A. Lundkvist, H. Henttonen, A. Plyusnin & A. Vaheri (2003) Hantavirus infections in Europe. *Lancet Infectious Diseases,* 3**,** 753-754.

Verhagen, R., H. Leirs, E. Tkachenko & G. Van der Groen (1986) Ecological and epidemiological data on Hantavirus in bank vole populations in Belgium. *Archives of Virology,* 91**,** 193-205.

Ward Jr, J. H. (1963) Hierarchical grouping to optimize an objective function. *Journal of the American statistical association,* 58**,** 236-244.

WHO. 1959. Zoonoses: Second report of the Joint WHO/FAO Expert Committee. ed. W. h. Organization.

WHO. 2004. Report of the WHO/FAO/OIE joint consultation on emerging zoonotic disease. ed. W. H. Organization. Geneva, Swisterland.

Wiens, T., B. Dale, M. Boyce & G. Kershaw (2008) Three way k-fold cross-validation of resource selection functions. *Ecological Modelling,* 212**,** 244-255.

Zeimes, C. B., G. E. Olsson, C. Ahlm & S. O. Vanwambeke (2012) Modelling zoonotic diseases in humans: comparison of methods for hantavirus in Sweden. *International Journal of Health Geographics,* 11**,** 39.

Zeimes, C. B., G. E. Olsson, M. Hjertqvist & S. O. Vanwambeke (2014) Shaping zoonosis risk: landscape ecology vs. landscape attractiveness for people, the case of tick-borne encephalitis in Sweden. *Parasites & Vectors,* 7**,** 1-10.

List of figures

Figure 1: Places for zoonoses transmission 4
Figure 2: Human PUUV infections in Sweden 33
Figure 3: Comparison between logistic regression, boosted regression tree, cross-validated logistic regression and MaxEnt model 50
Figure 4: Human disease cases of tick-borne encephalitis in Sweden 66
Figure 5: Graphs of each variable according to the fitted function of the global model 81
Figure 6: Resulting maps of the hazard, exposure and global models (based on 1998-2007 TBE records) and TBE records in 2011 83
Figure 7: Graphs of clear-cuts and mixed forest according to the fitted function of the reduced model. Blue scale: adjusted scale. Dark scale: common scale at all variables graphs. 85
Figure 8: Presence and absence of human PUUV cases and bank voles between 2001-2013 in Germany 97
Figure 9: Probability and false presence/absence maps from boosted regression trees on human Puumala virus infection cases and bank vole presence 104
Figure 10: Response curves from the boosted regression trees on the probability of human hantavirus and bank vole presence 110
Figure 11: Interaction graphs for the probability of presence of human PUUV hantavirus cases based on soil water index and the relative humidity and, the probability of bank vole presence based on wind speed and the proportion of mixed forest 112

Figure 12: Map of presences and absences of human hantavirus cases in Europe 125
Figure 13: Digital map of European ecological regions for the study area 134
Figure 14: Black curve: relative probability of presence of hantavirus human cases according to the evolution of the variable as modelled by BRT. Coloured boxplots: distribution of the variable by ecoregions (colours refer to figure 12 and the height of the boxplot is proportional to the number of points by ecoregion). + and – signs: sign of the logistic regression coefficient, if significant, for this variable in this ecoregion. Dashed color lines: boxplot extent (second to third quartile) of an ecoregion on the black curve. 135
Figure 15: Map of predicted probabilities and false presences / absences of the European multilevel model (threshold for presence = 0.25, when the sensitivity equals the specificity at 84%)..... 139
Figure 16: Response curves of the variable according to the predicted probabilities of the boosted regression trees, boxplots of the variable per ecoregion and, significant signs of the coefficient of bivariate logistic regressions per ecoregion 143
Figure 17: Three scenarios that can happen when modelling variables over a large study area, squared frames represent different subregions of a large study area..... 146

List of tables

Table 1: Review of factors influencing hazard and exposure 13
Table 2: Factors influencing hazard and exposure of hantavirus infections 18
Table 3: Precision and accuracy of human zoonose records 20
Table 4: Variables and hypothesized relationships with the abundance of bank voles, ex-vivo virus survival and human presence (* variable log-transformed) 36
Table 5: Models obtained by logistic regression and cross-validated logistic regression method 47
Table 6: AUC of partial models based on variables related to each element 53
Table 7: General advantages and disadvantages of logistic regression, boosted regression trees, cross-validated logistic regression and Maxent model based on methodologies accessible to beginners 60
Table 8: Variables selected in the hazard or exposure model 72
Table 9: Relative importance of variables introduced in the hazard and in the exposure boosted regression trees 79
Table 10: Relative importances of variables included in the reduced and global model 84
Table 11: List of variables included in the models with their units, sources and temporal and spatial resolution 100
Table 12: Relative importance of the variable in the human hantavirus case model and in the bank vole model, colours gradient represent the magnitude of the relative importance 105

Table 13: Variables and their hypothesized effect (X) on the bank voles' abundance, virus ex-vivo survival and human repartition (numbers in brackets indicate references) 130
Table 14: Multilevel logistic regression (significant at the level of * 0.5, **0.01 and, *** 0.001) 138
Table 15: Summary of variables responses in all models 157

Data table

		Data	Resolution	Units	Source	Weblink
World	Human	Population density	0.0083°	pop/km²	Gridded Population of the World	http://sedac.ciesin.columbia.edu/data/collection/gpw-v3
	Elevation	Elevation	1km	m	Global Digital Elevation Model	http://gdem.ersdac.jspacesystems.or.jp/
	Meteo	Annual temperature	1km	°C	Worldclim	http://www.worldclim.org/current
		Annual temperature range		°C		
		Precipitation		mm		
		Maximum temperature		°C		
		Minimum temperature		°C		
		Snow cover	0.05°	Area %	MODIS	http://nsidc.org/data/mod10cm
Europe	Land-cover	Forest	100m	NA	CORINE Land Cover	http://www.eea.europa.eu/data-and-maps/data/corine-land-cover-2006-raster-3
		Broad-leaved forest				
		Coniferous				
		Mixed forest				
		Urban areas				
		Water bodies				
		Open areas				

		Precipitation		mm		
		Relative humidity		%		
		Dry days		Days		
		Warm days		Days		
	Meteo	Maximum temperature	Station coordinates	°C	European Climate Assessment & Dataset	http://eca.knmi.nl/utils/mapserver/indices.php#bottom
		Minimum temperature		°C		
		Snow depth		cm		
		Wind speed		m/s		
		Growing season length		Days		
	Soil	Soil water index	25km	Relative units	TU-WIEN	https://rs.geo.tuwien.ac.at/products/da94d222-5306-5380-8d33-8a452dc6d173/153232/
	Human	Population proximity index	0.0083°	Number	Environment Research Group Oxford	http://openhealthdata.metajnl.com/article/view/jophd.ab/4
Sweden	Forest	Pine	30m	m^3/ha	Swedish University of Agricultural Sciences Skogskarta	http://www.slu.se/en/collaborative-centres-and-projects/swedish-national-forest-inventory/services-and-products/interactive-services/knn-sweden/
		Spruce				
		Oak				
		Birch				
		Height		m		
	Land-	Peat bogs	Shapefile	NA	Lantmäteriet database	www.lantmateriet.se/en/Maps-and-geographic-information/
		Forest				

	cover	Roads Water courses				
	Holiday houses	Holiday houses	Shapefile		Statistiska Centralbyran	http://www.scb.se/
	Meteo	Snow depth Snow duration	Station coordinates	cm Days	Swedish Meteorological and Hydrological Institute	http://opendata-catalog.smhi.se/explore/
	Bag records	Animals shot	Hunting district	Number of shot animals	Dr Jonas Kindberg, Swedish Association for Hunting and Wildlife Management	Personal communication
	Soil	Soil texture	100m	Type of soil	Geological Survey of Sweden	http://maps2.sgu.se/kartgenerator/maporder_en.html
Germany	Human	Population Agriculture, forestry and fishery workers	Municipality	Number Number/1000 inhabitant	Statistisches Bundesamt	https://www.regionalstatistik.de/genesis/online/data