Mapping tick dynamics and tick bite risk using data-driven approaches and volunteered observations

Irene Garcia-Martí

MAPPING TICK DYNAMICS AND TICK BITE RISK USING DATA-DRIVEN APPROACHES AND VOLUNTEERED OBSERVATIONS

DISSERTATION

to obtain
the degree of doctor at the University of Twente,
on the authority of the rector magnificus,
prof. dr. T. T. M. Palstra,
on account of the decision of the Doctorate Board,
to be publicly defended
on Friday, September 27, 2019 at 12.45

by

Irene Garcia-Martí
born on November 24, 1984
in Onda, Spain

This dissertation is approved by:

prof. dr. R. Zurita-Milla (promoter)

ITC dissertation number 367
ITC, P.O. Box 217, 7500 AE Enschede, The Netherlands

ISBN:	978-90-365-4859-5
DOI:	http://dx.doi.org/10.3990/1.9789036548595
Printed by:	ITC Print Department, Enschede, The Netherlands

ITC
UNIVERSITY OF TWENTE.
FACULTY OF GEO-INFORMATION SCIENCE AND EARTH OBSERVATION

Graduation committee

Chair / Secretary	
prof. dr. ir. A. Veldkamp	University of Twente
Promoter	
prof. dr. R. Zurita-Milla	University of Twente
Members	
prof. dr. M. J. Kraak	University of Twente
prof. dr. A. Nelson	University of Twente
prof. dr. S. Vanwambeke	Catholic University of Louvain
dr. ir. R. J. A. van Lammeren	Wageningen University

To my grandparents, Dolores, Pura & Salvador, who understood the real value of Education.

—Would you tell me, please, which way I ought to go from here?— asked Alice. —That depends a good deal on where you want to get to.—said the Cheshire Cat. —I don't much care where... —Then it doesn't much matter which way you go. —...so long as I get somewhere. —Oh, you are sure to do that, only if you walk long enough.

— Lewis Carroll, *Alice in Wonderland*

Acknowledgements

Research is a thrilling and challenging world that is in constant touch with the uncertain and the unknown. In a similar spirit as the old cartographic adventures that shaped our global society, the PhD explorer embarks in a journey with the hope of expanding human knowledge, and also being aware of the multiple (scientific) dangers that lie ahead. Perseverance and resourcefulness are two skills that are intensively trained during doctoral research, and these lead to the successful conclusion of this exploratory adventure, a story that you are now reading. It would be naïve to think that research in general is a one-person effort, because the exploration of this vast *"terra incognita"* requires the expert advice and support of personal and professional networks. I would like to make use of this opportunity to thank them.

I would like to start by expressing my deepest gratitude to Raúl Zurita-Milla, my supervisor, promotor and Jedi master in this scientific adventure. Raúl, thanks for being a supportive, respectful, and understanding supervisor. I believe that after these years working with you, I am now a resourceful professional, ready to tackle complex analytical problems always respecting the scientific method. Also, thanks for showing me the importance of being a dedicated and committed researcher. I have profound respect for the scientist you are, and I hope that I got from you some of these positive traits. Thanks for your guidance during this long journey and your patience at showing me the intricacies of the machine learning universe.

I extend my gratitude to Arno Swart, from the Dutch Institute for Public Health and the Environment (RIVM). I am glad and thankful that this research kept you on board during these years. I appreciate your openness regarding the application of data-driven techniques at modelling risk of disease, and your witty remarks about statistics. Thanks for your candid criticism and your active collaboration regarding our publications; no doubt your contributions increased substantially the quality of our works.

Thanks to Menno-Jan Kraak for accepting me as a PhD candidate in the department of Geo-Information Processing (GIP), for his interest in my research, and the valuable comments provided during research meetings. Thanks to Jolanda Kuipers for her good will and energy at helping out with the administration at multiple times during this period. Thanks to my corridor colleagues, Frank Ostermann, Rolf de By, Ellen-Wien Augustijn, and Lyande Eelderink, for creating such a pleasant work atmosphere. A special shout-out for my PhD candidate fellows also working with Raul: Hamed Mehdi Poor, Norhakim Yusof, Azar Zafari, and Xiaoling Wu: we made it, folks! My best wishes for your future endeavors. Also, thanks to all the staff members of the GIP department, for the multiple times I was at your door seeking your advice.

This research has been possible thanks to the contribution of thousands of anonymous citizens coordinated by researchers from Wageningen University & Research (WUR), and the RIVM. Thanks to Arnold van Vliet and Willem Takken, from WUR, and Cees van den Wijngaard and Margriet Harms, from RIVM, for providing the volunteered data collections that have been at the core of this research. I am also grateful for your interest in this research and for providing perspectives from ecology and epidemiology that have been helpful during the development of this research. In addition, thanks to the thousands of anonymous citizens that contributed with their observations on tick bites and tick activity.

During these years I have been lucky of having as friends a group of wonderful people that I can call my Netherlands family: Parya, André, Ana, Andrés, Emma, Vero, and Luis, you guys have been an incredible moral support in this long journey. I cherish each of the moments that we spent together, and I am very grateful for the positive atmosphere the countless occasions we were together. Thanks for the laughter and the memories that we now share. Also, thanks to my extended family, Tatjana, Alby, Sheila, Abhishek, Gustavo, Valentina, Rosa, Eduardo, and Manuel, for also helping at making the ITC a louder and merrier place to work in. We definitely had great fun together during the evening shift, so thanks for the good vibes.

Besides these great friends from the Netherlands, I would like to express my gratitude to all my favorite people in Spain: thanks to all my friends and family in Onda for always having encouraging and kind words towards my research, and also for finding quality time for me in each of my visits. You make me feel at home, and when you are abroad this is specially appreciated. Thanks to Gonzalo for believing in my work and professional capabilities. My special thanks to Mireia, Maria C., Noelia, Mauri and Maria F., for always being there, lifting my spirits, and making me feel their presence and their friendship from afar.

Last but definitely not least, I would like to express my gratitude to my parents, Vicent and Inma, and my brother, Guillem, for their unconditional support during these years. I always knew I was very fortunate of having such a loving and caring family, but I had to come this far to fully appreciate your greatness. Thanks for the happy atmosphere at home. Thanks for teaching me to respond with perseverance to adversity. Thanks for encouraging me to pursue my geek academic and professional goals with a smile and optimism for the future.

To all of you who have supported me during these years, this story is also yours, thank you.

Contents

List of Figures

List of Tables

Introduction 1

1.1 Background: the (re) emergence of vector-borne diseases

Vectors are living agents capable of transmitting pathogens causing infectious diseases to humans, animals, and plants (Last, 2001). Vector-borne diseases (VBD) are a major threat compromising public health, food security, and economic activities around the globe (WHO, 2014). The treatment of a VBD has associated an economic burden for citizens and the subsequent medical costs for public health systems (WHO, 2017). In addition, patients suffering a VBD also might have a temporal or chronic level of disability, which prevents them from working and supporting their households (WHO, 2014). For instance, an average episode of dengue requiring hospitalization represents 15 – 19 lost days (WHO, 2009), whereas an episode of mild Lyme borreliosis (LB) can take 5 weeks for a patient to recover (van den Wijngaard et al., 2015).

Currently, the World Health Organization (WHO) has identified nine types of vectors (i.e. mosquitoes, ticks, sandflies, triatomine bugs, black flies, tsetse flies, mites, snails, and lice) that can cause, at least, 16 major vector-borne diseases in humans. Major tick-borne diseases comprise two bacterial infections (i.e. Lyme borreliosis, tick-borne encephalitis) and one viral infection (i.e. Crimea-Congo haemorrhagic fever), although there are several minor tick-borne diseases (e.g. rickettsial diseases, human babesiosis, relapsing fever) with local importance (WHO, 2017). Livestock and crops are not free of the risk of acquiring VBD: ticks can infect cattle with bovine babesiosis, aphids transmit citrus *tristeza* virus to orange trees, whereas sap-feeding insects infect with the *xylella fastidiosa* bacterium grapevines or olive trees. Crops and cattle infected with VBDs might suffer damage or health conditions that compromise annual productivity, with the consequential cost for the agriculture and livestock sectors (NASEM, 2016).

Scientists and medical doctors started to discover the vector-borne transmission of pathogens after 1877, and by 1910, more than 10 VBD were already identified (Gubler, 1998). After the end of World War II, governments and public health organizations started massive programmes of vector eradication, especially focused on mosquito and lice abatement, by carrying out campaigns of insecticide spraying that dramatically decreased the vector populations —in spite of its toxicity to humans, wildlife and nature (Vos et al., 2000; WHO, 2015)— and consequentially, the incidence of VBD plunged. These programmes were a success from the point of view of disease control, but after two decades of implementation they were abandoned, because VBD were no longer a public health issue (Gubler, 1998).

The cease of the fumigation campaigns was not the only cause of the global (re) emergence of VBD. Historically, most of VBD have been confined to distinct geographical areas, but the global changes human societies triggered, also created opportunities for vectors and pathogens to geographically expand to new areas (WHO, 2014). Therefore, after the cease of these campaigns, major global modifiers such as climate change, human developments and demography, socio-economic exchanges, and human outdoor recreational activities, proved its relevant role at reintroducing vectors and facilitating the transmission of VBD.

The onset of the 1970s started with the identification of carbon emissions produced by a plethora of anthropogenic activities (e.g. burning fossil fuels, increasing transportation intensity, deforestation), as a trigger leading to a global increase of temperatures (Sawyer, 1972). Increasing temperatures exacerbate the intensity and incidence of extreme weather events (e.g. drought, flood, storms), and this in turn, might affect human health by causing an excessive mortality during heat waves, increasing the risk of respiratory disease due to pollution, or even shifting the geographic distribution of VBD (Haines et al., 2006; Medlock et al., 2013). By the end of the decade, this hypothesis was deemed as plausible by the World Meteorological Organization (WMO), who warned about the vulnerability of citizens in front of global warming (WMO, 1979).

The stage of human development of a society, is also a factor of resilience (or vulnerability) when a global change occurs: developed societies with well-consolidated health systems —and a low endemicity of VBD (WHO, 2016)— were able to better combat the upsurge of VBD from 1980s onwards, whereas the health systems of developing societies were overwhelmed by its resurgence (NASEM, 2016; WHO, 2017). Regarding demography, in the last five decades there have been two major types of movements of people. First, population growth, especially in developing societies, resulted in the major settlement of people in urban areas, often by means of an unplanned urbanization (Gubler, 1998). Second, in developed societies, a process of counter-urbanization occurred, in which part of citizens abandoned crowded urban centers to settle in peri-urban areas, thus bringing humans in closer contact with nature (Zeman and Benes, 2014). Both situations, increased and prompted suitable conditions for vectors to thrive, and for VBD to cause outbreaks (Chakravarti and Kumaria, 2005; Okwa et al., 2009; Randolph et al., 2008), which pose recurrent challenges to the health systems.

Global traffic and trade have substantially increased in the past 50 years: passenger flights have consistently grow 9% annually since 1960s (Tatem et al., 2006), and shipping traffic has experienced a fourfold increase since 1992 (Tournadre, 2014). Airplane traffic has facilitated the propagation of pathogens by transporting aboard infected humans or vectors in their adult stage, whereas vector eggs can be found in cargo shipping, which has helped introducing VBD like malaria, dengue, or chikungunya in non-endemic regions (Guzman et al., 2016; Tatem et al., 2012). In addition, the increasing societal adoption of healthier lifestyles by dedicating leisure time to physical or sportive activities outdoors, means that more humans are exposed to tick-borne diseases, such as LB (Sandifer et al., 2015), in suburban forests (Paul et al., 2016), urban parks (Hansford et al., 2017), and even private gardens (Mulder et al., 2013).

Today's globalized economy has accelerated the international flow of citizens, commodities, and livestock, which has expanded the geographic range of VBD and has allowed pathogens and vectors to colonize new regions (Lemon, Stanley M. et al., 2008). Since the (re) emergence of VBD, West Nile virus has been established in United States, chikungunya fever has resurged in Asia and Africa, dengue has appeared in Europe, whereas Zika virus has travelled all the way from Uganda to Brazil where it caused a major outbreak (Bouzid et al., 2014; Gubler et al., 2017; Musso et al., 2015). This globalized context poses new challenges to specialists and researchers to effectively plan campaigns that reduce the effect of new outbreaks.

1.2 Lyme borreliosis: a complex ecological problem

Ticks are pervasive ectoparasites globally present (except in regions with an extreme climatology) that are adapted to survive in a wide range of environmental conditions (Cumming, 2002; Vesco et al., 2011). Ticks are hematophagous arthropods, which means that they need to feed from human or animal hosts to complete their life cycle (i.e. egg, larvae, nymph, adult) (Lindgren and Jaenson, 2006). There are multiple tick species, capable of infecting with different pathogens humans, livestock, wildlife or pets (Uspensky, 2017). Ticks of genus Ixodes (also known as 'hard ticks') are capable of co-transmitting spirochetal and rickettsial bacteria, flaviviruses and protozoan parasites that cause different diseases in humans (Diuk-Wasser et al., 2016).

The relationship between a tick and LB was first reported in 1909. In that year, a Swedish dermatologist described an expanding skin lesion in an elderly patient, following a tick bite (Dammin, 1989). This skin lesion is known as erythema migrans (EM), and is a common early-stage manifestation of LB which can develop into severe forms of LB (e.g. neuroborreliosis, arthritis) if left untreated (van den Wijngaard et al., 2017). However, LB is one of the latest incorporations to the list of VBD, since the causative agent was not investigated until 1975. That year, a cluster of children arthritis and carditis in the village of Lyme (Connecticut, USA) attracted the attention of public health specialists. Scientists carried out a thorough investigation and several years later, the agent causing Lyme disease was found (Burgdorfer et al., 1982). The Borrelia burgdorferi complex is formed by different types of spirochetal bacteria (e.g. *B. burgdorferi*, *B. afzelii*, *B. garinii*) that cause similar symptoms of LB in humans (Diuk-Wasser et al., 2016).

Since the identification of the borrelia pathogens, scientists and clinicians have reported that the incidence of LB has steadily increased in, at least, nine European countries (Medlock et al., 2013), Canada (Ogden et al., 2014), and the USA (Schwartz et al., 2017). However, in recent years, sub European sentinel networks of general practitioners have identified the first signs of stabilization (Altpeter et al., 2013; Bleyenheuft et al., 2015; Vandenesch et al., 2014). In the Netherlands, tick bite consultations in general practitioners (GP) tripled during the period 1994-2009, from 191 to 564 cases per 100,000 inhabitants, whereas the incidence of EM experienced a similar rise, growing from 39 to 134 cases per 100,000 inhabitants (Hofhuis et al., 2015a). Similarly to other European countries, LB incidence in the Netherlands is showing the first signs of stabilization (Hofhuis et al., 2016). Yet, each year there are roughly 25,000 Dutch citizens that are diagnosed with LB. Most of them respond well to the antibiotic treatment, but there is a minority of patients reporting persisting symptoms after treatment, that can lead to disabling symptoms and increase the disease burden (van den Wijngaard et al., 2017).

LB infections are the realization of a complex ecological system involving the interaction of several biotic (e.g. environment, wildlife) and abiotic factors (e.g. weather, landscape) (Ostfeld, 2012). Ticks are the vehicle that pathogens utilize to infect new organisms, hence, it is of utmost importance to monitor tick dynamics to be able to identify hazardous locations for LB infection. Nevertheless, ticks are not the only factor to consider at estimating the risk of tick bites, since this calculation requires the inclusion of human exposure metrics in a location. Because of the global changes mentioned in Section 1.1, there range of ticks and humans has expanded, subsequently increasing the chances of a human-tick encounter while carrying out outdoor activities.

The geographical range of ticks has experienced an latitudinal and altitudinal expansion in the last decades as reported by scientists in Norway (Jore et al., 2011), Sweden (Jaenson et al., 2012) and Canad (Clow et al., 2017). This is due to two sequential factors: increasing global temperatures have turned unsuitable habitats for the tick life cycle into suitable regions, and subsequently, different wildlife species (e.g. rodents, ungulates, birds) have expanded their range, thus introducing ticks in new locations (Medlock et al., 2013). In addition, ticks are particularly vulnerable to weather conditions, since their high surface-to-volume ratio is prone to water losses through their exoskeleton, conditions that make them to desiccate and die (Ostfeld, 2012). Temperature determines the start of the questing season, or the survival chances throughout the winter season (Ogden et al., 2006; Randolph et al., 2008). Precipitation and atmospheric water levels (e.g. evapotranspiration, saturation deficit) are important to keep optimal levels of humidity at the ground level, which is crucial for the development of new tick populations and determine tick activity (Berger et al., 2014a; Mather et al., 1996; Randolph and Storey, 1999).

Similarly, the geographical range of humans has experienced an expansion. Concretely in Europe, intense human activities have led to a massive modification of the landscape. As a result, the area of cities has expanded, due to the development of low-density residential areas at the outskirts of cities (EEA, 2006). Urban sprawl has a remarkable effect on the human population distributions, since it brings urban settlers in closer contact with nature and the countryside (EEA, 2011). As a response of the expanded human range, several bird (e.g. thrushes) and mammal species (e.g. rodents, foxes, raccoons) have adapted their ethology to be able to live at the interface between forests and urban regions (e.g. more food, less predators) (Uspensky, 2017), but this also means that the pathogens that wildlife species carry are closer to residential areas. In addition, the progressive adoption of healthier lifestyles encourages citizens to spend more time outdoors carrying out leisure or sportive activites, but this behaviour could also lead to a higher exposure to tick-borne diseases (Mulder et al., 2013; Hall et al., 2017).

As seen, LB is an elusive public health threat due to the ubiquity of ticks and humans, and the wide range of biotic and abiotic factors involved in the ecologic system. In addition, traditional acquisition methods such as satellites, simulation models, or sensor networks are not able to monitor such fine-grained phenomena. In this context, citizen science initiatives can engage the general public in a wide array of environmental and public health monitoring activities. These activities result in very local observations and enable taking the pulse of LB and other VBF at unprecedented spatio-temporal scales.

1.3 The role of citizen science at monitoring tick-borne diseases

Citizen science (CS) is the non-professional involvement of volunteers in the scientific process, whether it is in the data collection phase or other phases of research (Gold et al., 2018). CS can be applied to any field of expertise (e.g. astronomy, ecology, meteorology, water and air quality), and some of these projects have gathered enough compromise from citizens to last more than a century. As an example, ornithologists in the USA and UK started organizing yearly bird counts in 1900 and 1932 (Craglia and Shanley, 2015), respectively, only surpassed in antiquity by the first meteorology cooperative programme, initiated in USA in 1890 (Fiebrich, 2009). Thus, long before the term CS was popularized, enthusiasts of science could see the potential of joining efforts at monitoring diverse large-scale or fine-grained environmental phenomena.

The rise of Web 2.0 technologies (O'Reilly, 2007) applied to geography boosted the number of applications that are based in citizen's location. Global society has witnessed the emergence of new technologies such as web mapping, location-based services, geotagging or geoblogging, and its widespread use required the creation or the re-formulation of terms in order to refer those developments properly (Elwood, 2008a,b). The continuous growth and popularity of Web 2.0 technologies among citizens, naturally led to a new way of digital collaboration between users and data acquisition based in crowdsourcing for a specific purpose. Crowdsourcing is an activity where the massive participation of citizens through communities of users is desired in order to accomplish collectively something perceived as a greater good, whose output might be exploited by other individuals, public or private entities (Haklay, 2010).

Citizen's participation is at the core of "Volunteered Geographic Information" (VGI) (Goodchild, 2007a), a term that intends to approach the efforts made by the crowd in location-based projects to the geospatial domain. The author argues that humanity as a collective possesses a huge amount of knowledge about the Earth surface and its properties (e.g. local toponyms, status of cultural heritage, conditions or road pavement in a city), therefore, enabling citizens with electronic devices to digitize this knowledge makes possible the creation of a massive collection of raw data, that subsequently can be introduced in scientific analysis, web services or geoprocesses (Goodchild, 2007b).

The idea of "humans-as-sensors" promoted by Goodchild, has been implemented in a plethora of CS initiatives in the past decade, and across multiple disciplines. Citizens reporting on fine-grained phenomena such as the occurrence of pollinators (e.g. BeeSpotter), birds (e.g. eBird), or wildlife in general (e.g. Waarneming, 'observation' in Dutch), contribute monitoring the pulse of distributions of living organisms. Human-made structures or toponyms around the globe have been thoroughly mapped (e.g. OpenStreetMap) and identified (e.g. GeoNames) by volunteers, and it is even possible for citizens to report the changes of agricultural land use and urban dynamics (e.g. LandSense), or contribute monitoring the weather (e.g. Weather Observations Website, WOW) using a personal automatic weather station. The data provided by these platforms might contribute to a wide range of applications, from monitoring the species migration or distributions at the national or continental scale (La Sorte et al., 2017), to help studying urban heat islands (Chapman et al., 2017), assessing the synchronicity of phenological events (Mehdipoor et al., 2018b), or even assisting emergency managers when natural disasters occur (Haworth, 2016).

CS can be used to advance scientific discovery and knowledge, build a sense of community, inform policy and environmental management, or to educate and rise awareness among a target citizen group (Craglia and Shanley, 2015). These are desirable topics to comply with to advance towards a more informed and participatory decision making process (Gold et al., 2018). CS initiatives have gained a remarkable attention in the scientific scope, this is why the number of publications using VGI has experienced a substantial increase (Kullenberg and Kasperowski, 2016) in multiple fields of environmental sciences and Earth Observation (See et al., 2016a,b) since the early 2000s. This is in general a positive trend, since the inclusion of VGI sources in scientific workflows can provide an unprecedented spatio-temporal resolution at monitoring complex and elusive environmental phenomena. Nevertheless, VGI is not exempt of several issues and challenges that require attention.

Data quality control is a dicey challenge to work with when dealing with VGI collections. By default, there are some general rules proposed by (Goodchild and Li, 2012) to increase the quality of VGI observations, in which it is necessary to implement a validation workflow considering three dimensions: crowdsourcing, social and geographical. The verification of these three dimensions implies that a new observation, if valid, should have been reported by other users, approved by a group of trusted moderators, and consistent with the surrounding observations. However, although this procedure is reasonable, the complexity of the phenomena under study in CS projects might require a more elaborated validation procedure. For an instance, in (Zhao and Sui, 2017) the authors engineer a procedure to detect location spoofing in Twitter data by using a time-aware Bayesian analysis, in (Mehdipoor et al., 2015) the authors develop a workflow to detect temporally inconsistent volunteered observations, whereas the eBird project (Sullivan et al., 2009) applies a thorough checklist of filters verifying whether a bird species observation is out of range or season. VGI quality is also related with the level of expertise of each individual contributor (Yang et al., 2016). For example, volunteers helping at classifying wildlife species might not have enough skills to distinguish between two types of deer (Kosmala et al., 2016), or citizens with limited access to technology might not have a basic technical profile (e.g. map literacy, fluent use of digital devices) enabling them to introduce new observations correctly in a database (Su et al., 2017).

The representativeness of the phenomenon under study (Zhang and Zhu, 2018) and the inequality in data coverage (reporting bias) (Su et al., 2017) are also related with the number of contributors to the project and their skills. Oftentimes, mitigating the effects of these factors requires the development of a custom-made procedure. For an instance, researchers in the eBird project propose an adaptative spatio-temporal model capable of accommodating spatial bias and the density of reports in a region by training local models that are subsequently integrated in a larger one (Fink et al., 2010, 2013). Other researchers filter clusters of repetitive observations (Boria et al., 2014; Varela et al., 2014) or they provide a weight to observations matching a set of criteria (Zhu et al., 2015), with the intention of mitigating reporting bias. These two factors are also related with the number of contributors to the project in a region, and their skills. In Haklay et al. (2009) the authors discuss that a small group of 15 contributors per square kilometre can map reality with a good positional accuracy for the OSM project, to the point that this dataset is comparable in quality to Ordnance Survey (UK) in densely populated areas (Haklay, 2010). Thus, positional errors in VGI data collections might not be randomly distributed, but depending on the professional or amateur skills on the contributor (Craglia and Shanley, 2015; Yang et al., 2016).

Other well-known issues of VGI include sustaining the commitment of users in the long term and the privacy or confidentiality of data (Craglia and Shanley, 2015). The authors discuss that to keep a CS project alive in the long-term it is necessary to understand the motivation of the users and aligning the objectives of the project with the expectations of the users. Regarding privacy, the authors recommend documenting a project properly, so the procedure can be reproduced and be understood from different disciplines and backgrounds. Note however, that some projects might be tied to confidentiality and privacy clauses, since the phenomenon under study can be sensitive (e.g. monitoring endangered species) or the contributors do not wish to be identified (e.g. personal data, living habits) (Mooney et al., 2017).

Life and environmental sciences tend to coalesce the majority of CS initiatives (Gold et al., 2018), but in the field of health geographics there are not so many examples that have studied how volunteered data could help at monitoring VBD. Some existing study cases include the incorporation of volunteered data to devise new indicators predicting dengue in Malaysia (Mokraoui et al., 2018) or the mapping of Chagas disease in Texas with the help of citizens submitting triatomine bugs for analysis (Curtis-Robles et al., 2015). Focusing in tick-borne diseases, there have been some initiatives by public research institutes in which they have launched campaigns of analysing submitted ticks to check for the pathogens they are carrying (Nieto et al., 2018). Other scientists have followed a less conventional approach, in which roadkill animals are analysed to find out the tick-borne pathogens they carry (Szekeres et al., 2018). In addition, public health organizations gather data from general practitioners (GP) every few years to assess the number of tick-borne consultations (Hofhuis et al., 2016). The limitations of these valuable efforts are that public campaigns to analyse ticks can provide vast amounts of data at the national scale, but they tend to be “one-time efforts”, which are costly to maintain in time. Analysing roadkill animals might provide accurate information on pathogens at the local scale, however, it is not straightforward to scale the results up to the national scale to get a general overview on the status of a disease. Finally, the assessment of GP data can provide accurate results on the incidence of tick-borne diseases, however, these massive studies gathering data from thousands of GP are commissioned every few years, so they are unable to account the intra annual variation of a disease. In this context it seems desirable to find a CS initiative capable of monitoring tick bites and tick dynamics at a fine spatio-temporal resolution, so it is possible to assess for each location in the country the probability of getting a tick bite.

In 2006, Wageningen University started collecting volunteer tick bites through the educational phenology platform Natuurkalender (NK; ‘nature’s calendar’, www.natuurkalender.nl), gathering nearly 10,000 volunteered tick bites in six years. This pioneering project attracted the attention of the Dutch National Institute for Public Health and the Environment (RIVM) and in

2012, the platform Tekenradar (TR; 'tick radar', www.tekenradar.nl) was launched together with Wageningen University. TR is a web platform especially conceived to inform citizens about the risk and prevention of tick bites and at the same time a citizen science platform to collect volunteer tick bites and erythema migrans observations. These projects have attracted enough media attention over the years to engage citizens at contributing, on a volunteered basis, tick bite reports to the platforms. The result of this engagement with citizens has produced over 50,000 volunteered tick bite reports in the Netherlands. This unique collection of observations enables multiple possibilities at monitoring and modelling elusive public health threats, such as tick bites. To the best of our knowledge, these platforms constitute the first citizen science projects that specifically focus on ticks and tick-borne diseases. Also in 2006, a group of scientists from Wageningen University started a countrywide investigation to assess the factors influencing the risk of LB (Gassner et al., 2011). This study comprised the monthly sampling of 24 forested locations in the Netherlands to count ticks in each of their life stages (i.e. larvae, nymph, adult). To do so, a group of trained volunteers would sample a transect of forest using a method called blanket dragging and turning the blanket every 25m to count the number of ticks attached. This project ran from 2006 – 2016 and created a unique collection of volunteered data measuring tick dynamics.

These two volunteered collections of data on tick dynamics and tick bites reports were available at the beginning of this PhD thesis for research purposes. Note that these data sources are not free of the problems associated to VGI mentioned before, since they present some of the expected traits of volunteered collections, such as loose structure, positional accuracies, reporting bias, and a variable quality of the observations (Mehdipoor et al., 2015; Senaratne et al., 2017; Welvaert and Caley, 2016). Albeit these expected issues, these data collections have a sufficient quality to be included in several scientific workflows. Hence, the modelling of these data collections enable the possibility of mapping at a fine spatio-temporal resolution tick hazard, human exposure to tick bites, and tick bite risk.

1.4 Spatio-temporal modelling of hazard, exposure, and risk with data-driven models

In the field of risk assessment, risk (R) is often modelled as a function of hazard (H), exposure (E), and vulnerability (V). The relationship between the four variables can be conceptualized as $R = HxExV$ (Braks et al., 2016; UNDRR, 2016). The dictionary of epidemiology (Last, 2001) defines risk as the *"probability that an individual will become ill or die within a stated period of time [. . .]"*, hazard is the *"inherent capability of an agent [. . .] to have an adversely health effect"*, whereas the exposure refers to the *"proximity and/or contact with a source of a disease agent in such a manner that effective transmission of the agent or harmful effects of the agent may occur"*.

Vulnerability is defined by the UN (UNDRR, 2016) as *"the conditions determined by physical, social, economic and environmental factors [...] increasing the susceptibility to the impacts of hazards"*. The combination of the abovementioned risk assessment principle with the definitions used in epidemiology, provide an analytical framework that we used throughout the development of this PhD thesis. In this research, we understand R as the "risk of tick bite", H as "tick dynamics", and E as "human exposure". The V component has not been considered in this research, due to the unavailability or incompleteness of occupational or human behavioural data collections. Nevertheless, we expect vulnerability to be fairly constant, since citizens tend not to take preventive measures against ticks (e.g. chemical repellent, protective clothes), thus becoming vulnerable. Therefore, the remaining of this PhD dissertation describes data-driven approaches that enable the calculation of each of these components and integrating them in a single tick bite risk variable.

In the past decades, there has been a remarkable effort in the fields of biology, ecology, and epidemiology to model tick-borne diseases, especially LB. Modelling these three components requires the inclusion of the spatial and temporal dimensions, since they are inherently dependent on the location and time. Previous works in literature modelling different VBDs have attempted at quantifying the H component including the spatial dimension explicitly or implicitly. In (Berger et al., 2014a; Linard et al., 2007) the authors explicity conceive space as a dimension from which a number of parameters (i.e. real-world traits) characterizing local or global effects can be derived, which are subsequently modelled with classical statistical methods, whereas in (Kala et al., 2017) the authors implicitly define the spatial dependency using a geographic weighted regression. Other researchers have attempted to simultaneously model space and time to find clusters of spatio-temporal co-ocurrence of disease outbreaks (Kanaroglou et al., 2015; Yang et al., 2017). Although these are valid approaches, there is an intrinsic limitation in it: classical statistical models tend to have difficulties at finding and understanding the non-linear interactions within the elements of the zoonotic cycle (i.e. ticks, pathogens, environment, humans) (Ostfeld, 2012). In this context, the use of machine learning methods to model spatio-temporal phenomena might overcome these hurdles, since these methods naturally deal with non-linear phenomena and high-dimensional problems. In this thesis we have performed classification, regression, and frequent pattern analyses using machine learning algorithms. The objective of these analyses was to investigate and calculate each of the components of R using machine learning methods.

1.5 Societal and environmental relevance

VBDs pose a substantial burden on public health systems and households, which translates in medical costs, working days lost to illness, and potential long-term sequels for the patient (WHO, 2014). Measuring the burden and cost of a disease is not a straightforward task, since these conditions vary for each year, country (or region) affected, and the intensity and persistence of an outbreak (Murray et al., 2012). In public health there is a measure to quantify the burden of a disease in patients: the disability-adjusted life years (DALY), which refers to the number of life-years lost due to poor health or disability per population unit (World Bank, 1993), often 100,000 citizens.

VBD have a variable cost and burden depending on the factors mentioned before: In the USA, Chagas had an estimated cost of $464 per patient and 0.51 DALYs (Lee et al., 2013), whereas in South America the disease burden ranges 25 -125 DALYs (Stanaway and Roth, 2015). Dengue outbreaks in Southeast Asia during the period 2001-2010 cost $610M - $1,384M with an estimated disease burden of 21-52 DALYs (Shepard et al., 2013). Chikungunya in India supposed a cost of $5.5M and 4.5 DALYs (Krishnamoorthy et al., 2009), whereas in the Caribbean region the disease burden ranges from 0.25 – 911 DALYs (Cardona-Ospina et al., 2015). Focusing in LB, a study in the USA shows that the economical of treating short-term to persisting symptoms, ranges between $464 - $1380 (Zhang et al., 2006). In the Netherlands, the treatment of LB costs approximately €5,700 per patient, with a yearly lump sum of €20M euros in total, and an estimated disease burden of 10.55 DALYs (van den Wijngaard et al., 2017).

For all the above, we think that the monitoring of VBD in general, and LB in concrete, has a remarkable societal relevance, especially to contribute to three the UN sustainable development goals (SDG) [1]: “good health and well-being”, “climate action”, and “life on land”. LB might not have the dubious honor of being a well-known VBD causing tens of millions of infections per year globally, yet this silent disease has a substantial burden. Ticks are organisms with a reduced motility that require a complex ecological system around them to survive. This means that if we can use VGI data to, both, understand the relationship between ticks and the environment, and the environment and humans, then we are able to devise novel and more effective map products for tick hazard, human exposure and tick bite risk that can help at designing tick-borne prevention campaigns.

In this research we worked at developing such map products to help public health professionals, and to inform citizenship on the risk of getting a tick bite. Tick hazard maps can be useful for ecologists and biologists to further study how tick dynamics is influenced by atmospheric conditions or wildlife. Human exposure maps could help public health specialists to identify recreational locations that are massively visited by citizens and consequently, design a tick prevention campaign in the closest neighbourhoods or municipalities. In addition, public health specialists could use this map to jointly work with forest managers to implement measures of tick habitat manipulation (e.g. clear shrubs, add dry substrates) to make forests a safer environment for visitors. Citizens and public health specialists could benefit of a tick bite risk map, since the former group would be aware of

[1] United Nations SDG

the locations to avoid or where extra caution is needed, whereas the latter group could target new locations to inform about the risk of LB. Regarding distribution channels, we think that citizens could benefit of a mobile application alerting users when they are entering a risky location for LB infection. In addition, the models developed in this research could be implemented and deployed in other organizations, so they can be used, studied, and further improved them to match the necessities of each professional group. We hope that this thesis could help as a basis for different professionals to work towards better tick bite prevention campaigns, so in the next years we witness a decrease in the number of LB cases.

Linking our research with other VBD, we hope that the lessons learned during the development of this PhD thesis can help at monitoring other diseases in a similar manner. We think that it might be of interest for other researchers learning how to combine VGI with environmental data to subsequently modelling it with machine learning methods, since these methods can understand the non-linearity of zoonotic cycles and enable the possibility of devising new map products regarding hazard, exposure, and risk of disease. We envision that the popularization of these methods among the health geographics researchers and public health specialists, could help planning large-scale VBD prevention campaigns, so that the number of infections per year decrease substantially. This is especially important in developing countries, since VBD tend to exacerbate poverty and socio-economic differences.

1.6 Research objectives

Human societies are witnessing an era in which major global changes are occurring at an accelerated pace. As a response, these changes might have a negative impact in our daily lives, hence it seems reasonable to dedicate efforts at monitoring them to create more resilient societies. In the context of VBDs in general and LB in particular, this quest requires the investigation of a plethora of phenomena at the finest spatio-temporal resolution possible.

The objective of this research is to investigate innovative methods to advance the modelling of tick dynamics and tick bite risk, by simultaneously modelling volunteered data and a wide array of heterogeneous geodata collections with data-driven methods. This methodology is important to gain knowledge on where and when these negative impacts will occur, and might help professionals to mitigate these pernicious effects. We operationalize this main objective by investigating data-driven approaches to model the R, H, and E components. The following research questions (RQ) introduce the research questions that vertebrate this PhD research:

- **RQ1:** How to develop a data-driven approach combining volunteered and environmental geodata, which is capable of capturing tick dynamics and assess the major drivers of tick activity across time-scales?

- **RQ2:** How to use data mining methods to identify spatio-temporal patterns linked to tick bites and verify that these patterns, stemming from volunteered observations, are intrinsic to the phenomenon under study?
- **RQ3:** How to devise a novel indicator of human exposure to ticks, enabling the geographical identification of clusters of high exposure?
- **RQ4:** How to integrate hazard and exposure metrics to devise a tick bite risk model, capable of handling the skewness and zero-inflation inherent to the volunteered tick bite reports?

1.7 Thesis outline

This chapter contains a thorough description of the main building blocks that vertebrate this thesis (i.e. VBDs, LB cycle, citizen science, data-driven methods), and we highlight our contributions or innovations to each of them. We also include a section in which we discuss the environmental and the societal relevance of this thesis. This chapter also includes the main research objective and research questions of this thesis.

Chapter 2 presents the description of a data-driven model capable of predicting daily tick dynamics. This analysis required the integration of an array of environmental variables (i.e. weather, tick habitat, satellite-derived vegetation indices, land cover, mast years) at different time-scales, to better understand the impact that long-term and short-term variables have on tick activity. We modify a well-known ensemble learning algorithm to enable it to yield temporally-aware predictions based on the day of the year.

Chapter 3 introduces an extensive exploratory data analysis that identifies the most recurrent human and environmental patterns found in a volunteered tick bites dataset. We enrich the volunteer dataset with multiple environmental and human variables, which are modelled with a frequent pattern mining algorithm. We also assess whether the tick bites collection is representative of the phenomenon under study, by generating an artificial dataset and comparing whether the patterns of the original tick bites dataset can be reproduced by random spatio-temporal sampling.

Chapter 4 presents a novel map representing human exposure to tick bites in forested areas in the Netherlands. This map is the result of combining the tick dynamics model developed in Chapter 2 with the volunteered tick bites. We demonstrate that the risk of tick bite is strongly influenced by human behavior, rather than the tick dynamics in a location. With this map, we are able to identify at the national level locations where citizens are exposed to ticks, such as urban parks, popular recreational sites, or suburban forests.

Chapter 5 develops a tick bite risk model integrating tick hazard and human exposure to tick bites. We take the tick dynamics model developed in Chapter 2 and we devise a series of human exposure indicators, based on accessibility and landscape attractiveness metrics. We modify a well-known ensemble learning algorithm to enable it modelling imbalanced data collections, by combining a segmentation task with count data models (i.e. Poisson family). In this way, we are able to predict tick bite risk for the Netherlands and identify risky locations for disease transmission.

Chapter 6 summarizes the main findings from Chapters 2 to 5. We provide a reflection on the relevance of the contributions of this thesis, and we answer the research objectives posed in Chapter 1. In addition, we provide recommendations and guidelines for future research.

2 Modelling and mapping tick dynamics using volunteered observations

2.1 Introduction

Tick populations and tick-borne infections like Lyme borreliosis have steadily increased since the mid-1990s. This concurrent increase has been observed in various European countries (Heyman et al., 2010; Jaenson et al., 2009), in the US (Subak, 2003) and in Canada (Ogden et al., 2014). In the Netherlands, periodic national studies among general practitioners (GPs), revealed a consistent two-decade rising trend in the number of tick bites consultations and Lyme borreliosis diagnoses (Hofhuis et al., 2015b), that only showed a first sign of stabilization recently. Still, more than 20,000 people per year develop Lyme borreliosis in the Netherlands and its disease burden is substantial, especially in patients who develop chronical symptoms (Hofhuis et al., 2016).

Scientists of different fields have investigated this global increase of tick populations and tick-borne infections, converging upon two main causes: global environmental changes are altering the spatio-temporal dynamics of ticks (Medlock et al., 2013; Sprong et al., 2012) and socio-economic changes are changing the spatial patterns of human populations around urbanized areas, increasing the human exposure to ticks (Randolph, 2013; Randolph et al., 2008; Zeman and Benes, 2014). Tick dynamics are complex ecological processes driven by numerous factors (i.e. wildlife, weather, vegetation, landscape). Understanding the interactions between these factors and tick dynamics is crucial to develop models capable of forecasting the incidence and distribution of ticks and tick-borne diseases (Estrada-Peña and de la Fuente, 2016; Ostfeld, 2012).

Models predicting the spatio-temporal distribution of ticks are needed to implement control measures which mitigate future disease infections (Cianci

This chapter is based on: (Garcia-Martí et al., 2017)
Available at: https://doi.org/10.1186/s12942-017-0114-8

et al., 2015; Hartemink et al., 2015) or help managing public health risks (Medlock et al., 2013). However, the development of such models is not straightforward due to several issues. First, it is unclear what the best set of environmental predictors are. Past studies have found correlations between different combinations of biotic and abiotic factors and tick dynamics, but the spatio-temporal scale of these experiments is diverse enough to pose difficulties in drawing general conclusions. For instance, Berger et al. (2014b,a) found a link between relative humidity and the seasonal abundance of ticks at the regional level. Dantas-Torres and Otranto (2013), found weak correlations at local scale between monthly temperature, evapotranspiration and saturation deficit with tick abundances, whereas Randolph and Storey (1999) found links (in laboratory conditions) between the saturation deficit and the number of questing ticks. Second, it is often unclear at what time scales the different predictors operate. Previous studies have found linear correlations between tick abundances and environmental predictors at multiple temporal scales (Berger et al., 2014b; Tack et al., 2012). However, the temporal sparsity of the tick sampling or the use of short-term time series question if these correlations are scalable to long-term time series at the country level. Third, tick dynamics are complex phenomena that traditionally have been modelled with linear methods. Two of the well-known disadvantages of classical linear methods is that they are not capable of finding non-linear interactions between variables (except when explicitly included a-priori), and do not properly handle large numbers of predictors (e.g. due to collinearity). However, such data are a reality when modelling complex natural phenomena.

In this work, we address the above-mentioned issues by modelling nine years of monthly data on Active Questing Ticks (AQT) collected by volunteers on 15 different locations in the Netherlands. This modelling exercise includes a wide array of (a)biotic predictors and, by applying an ensemble regression method (i.e. Random Forest), we aim at identifying the most important variables to model AQT at multiple time-scales. Building such AQT dynamic model allows us to explore and map tick's seasonality across the Netherlands. We envision applications of this model in the fields of environmental and ecological research, nature management and public health, which hopefully will reduce the incidence of Lyme disease.

2.2 Ticks and environment

2.2.1 Tick sampling

Ticks are blood sucking arthropods capable of transmitting a wide variety of pathogens (e.g. bacteria, viruses) which cause disease in humans (Heyman et al., 2010). Deciduous or mixed forests in temperate and humid regions, which are inhabited by different mammalian species (e.g. deer, rodents), create optimal habitats sustaining ticks life cycle (Ostfeld, 2012). Ticks quest at the top of vegetation or litter layer, waiting for a human or animal host to attach and feed. This behavior is used to determine tick populations in a particular location. To do so, two manual monitoring techniques are used: flagging and dragging. Flagging consists on sweeping a squared cloth attached to a pole on one side upon the litter or vegetation layers, whereas dragging consists in attaching the previous material to a rope, which the investigator can pull along the study area (Rulison et al., 2013). In both cases, ticks that are touched by the cloth attach to it, allowing researchers to count the number of ticks in its different life stages (i.e. larvae, nymph, or adult). Both techniques have been widely used in small scale biological studies to acquire raw data on tick counts that can be later incorporated in a scientific workflow (Dantas-Torres and Otranto, 2013; Estrada-Peña, 2001; Estrada-Peña et al., 2013; Gassner et al., 2011; Randolph, 2000).

2.2.2 Environmental factors

Ticks are particularly susceptible to environmental conditions because of their high surface-to-volume ratio, which makes them experience water losses through their exoskeleton, and their lack of thermal inertia, which makes them vulnerable to extreme weather conditions (Ostfeld, 2012). The following sub sections list the environmental variables used in our work and sketch their impact on tick dynamics.

2.2.2.1 Weather data

Temperature determines the start of the questing season, tick population development rate and the chances of survival through the winter season (Ogden et al., 2006; Randolph et al., 2008; Wu et al., 2010). Precipitation and relative humidity are crucial to sustain tick populations in nature. Precipitation is necessary during the summer season (Jore et al., 2014), but extreme precipitation events (i.e. drought and heavy rain) may prevent the development of new tick populations (Ostfeld, 2012). Long-lasting and adverse humidity conditions have been linked to an increased mortality among nymphal ticks and this, in turn, may decrease the total number of cases of Lyme disease (Berger et al., 2014a). Some studies suggest that nymphal ticks can desiccate within 48 hours if the humidity conditions at ground level are sub-optimal (Berger et al., 2014b). Additionally, relative humidity and temperature can be used to calculate the saturation deficit and vapor pressure. Saturation deficit has been used in a previous and thorough study to understand the role of humidity in tick survival (Randolph and Storey, 1999) and vapor pressure has been identified as a major indicator of tick habitat suitability (Brownstein et al., 2003). In some studies, evapotranspiration has been used as a proxy for vapor pressure deficit (Ruiz-Fons et al., 2012).

Weather datasets are publicly available at the online data center of the Royal Netherlands Meteorological Institute (KNMI)[1]. We downloaded daily gridded layers of temperature, precipitation, evapotranspiration and relative humidity for the period 2005-2014. From temperature and relative humidity, we obtained saturation deficit and vapour pressure (Murray, 1967; Randolph

[1] https://data.knmi.nl/datasets

and Storey, 1999). The temporal resolution of the weather datasets and the tick sampling is different, since the former are available at daily temporal resolution, whereas the latter is carried out on a unique day each month. To match both resolutions, it is necessary to aggregate the weather variables to a coarser temporal scale in a way that reflect the impact later caused on the tick count.

2.2.2.2 Vegetation data from satellites

Ticks are sensitive to local environmental conditions, such as the thickness of forest canopy or soil moisture at the ground level (Medlock et al., 2013). Earth observation satellites allow the monitoring of these environmental conditions over large areas. In this work, we used three vegetation indices to characterize local environmental conditions: the Normalized Difference Vegetation Index (NDVI), the Enhanced Vegetation Index (EVI) and the Normalized Difference Water Index (NDWI). Previous studies have demonstrated that fluctuations in NDVI, which has traditionally been used to measure the greenness and the density of vegetation, correlate well with fluctuations in the number of nymphs and adult ticks and that NDVI can be used as a proxy to find suitable tick habitats (Estrada-Peña, 2001; Randolph, 2000). More recent studies show that novel vegetation indices like EVI or NDWI are better estimators of tick populations (Barrios González, 2013) and Lyme disease incidence (Ozdenerol, 2015).

Vegetation indices are publicly available in the Google Earth Engine (GEE) platform [2] [3]. GEE is a free image processing cloud platform for environmental analysis, which aggregates and integrates products coming from different Earth observation sensors, such as the Moderate-Resolution Imaging Spectroradiometer (MODIS). MODIS provides daily global imagery at 250, 500 and 1000 meters of spatial resolution. However, due to the persistent cloud coverage over the Netherlands we used MODIS composite products. In particular, we used the MCD43A4 product, which provides the NDVI, EVI and NDWI indices derived from the daily surface reflectance at a pixel size of 500 meters, using data of the previous 16 days. It is important to note that this product is released every eight days, so there is a 50% of temporal overlap between each composite, meaning that the vegetation signal will contain smooth changes.

2.2.2.3 Land cover, tick habitat and mast years

Land cover is another important factor in the field of tick ecology because it influences tick survival and determines the chances of human-tick contact. Ticks prefer habitats where the vegetation prevents reaching desiccation conditions and where hosts (e.g. deer, rodents, mice) species are present. Complex landscapes, in which multiple land covers are intertwined in a small area unit, increase the probability of contact between ticks and their

[2] https://code.earthengine.google.com/

[3] https://earthengine.google.com/

human or animal hosts (Hartemink and Takken, 2016; Lambin et al., 2010; Li et al., 2015; Tran and Tran, 2016). For land cover we use the 7th release of the national land cover database or LGN (Landelijk Grondgebruik Nederland)[4]. This database was produced in 2012 and contains information for 39 classes at 25 m.

The sampling sites are located in forested areas with specific types of vegetation (i.e. deciduous and coniferous forest, grasses, and bushlands). The plant associations in these sites contribute determining the presence of wildlife species in each location, by providing forage or shelter, and subsequently, tick populations move with them. Previous studies have demonstrated that deciduous forests present higher abundances of AQT than coniferous forest, and also that a dense shrub layer has a positive effect on tick populations (Tack et al., 2013). Gassner et al. (2011) gives a thorough description of the plant associations and habitat characteristics found in the surroundings of each transect of the flagging sites.

2.3 Data

This work relies on a unique dataset of tick dynamics collected by volunteers in the context of a project of participatory modelling. This dataset was enriched with a set of environmental variables extracted for each sampling location. For this, we collected and preprocessed weather and satellite data, and included biological data regarding habitat and mast years. The remaining of this section first contains a description of the volunteered tick counts data (Section 2.3.1), and then we explain the process of feature engineering carried out to create a series of predictors that characterize tick dynamics as monitored by volunteers (Section 2.3.2).

2.3.1 Volunteered tick counts reports

In the context of the Dutch phenological network Nature's Calendar[5] every month since July 2006, a group of volunteers sampled AQT on 24 forest sites. This joint effort aimed to quantify and understand the spatial and temporal dynamics of ticks and the Borrelia bacteria that can cause Lyme disease (Gassner et al., 2011). Out of the 24 sites participating in the research project, we were able to include data from 15 sites, which represent a total of 3,073 observations collected by volunteers. We excluded the sites in which the sampling stopped in an early stage of the project, or the site was sparsely sampled in time. At each site, volunteers sampled two transects, separated from each other several hundred meters. Ticks were collected using a technique called "dragging", in which the volunteer drags a 1m ^{2} cloth over the low vegetation of each transect for 100m, turning the cloth every 25m to count the number of larvae, nymphs and adult ticks. This study focuses on the nymphs because they pose the highest risk for humans

[4] http://tinyurl.com/j47m2ol

[5] www.natuurkalender.nl

to get a tick bite. Figure 2.1 shows the raw number of nymphs per transect and per month. The number of AQT across all sites present strong spatial and temporal variations: 1) Some transects present a more continuous and recurrent shape, whereas others have an erratic tick count (e.g. Gieten vs. Bilthoven); 2) Some transects produce very different yields, from low tick counts to high peaks (e.g. Veldhoven vs. Eijsden); 3) Transects within a sampling site may yield a different number of ticks, even though they are close in space and sampled on the same day (e.g Montferland). The reasons of these strong local and seasonal variations are still poorly understood, but previous works have found clear links between tick populations and the abundance of small mammals in the area (Ostfeld et al., 2006), mast years (Jones et al., 1998) or warming weather conditions (Jore et al., 2014; Subak, 2003), which are major influences over tick dynamics, as seen in Section 2.2.2.

Volunteered projects have proved useful to acquire information at a timely and fine spatial scale, but the quality and the amount of uncertainty of such data collections is difficult to measure (Kamel Boulos, 2005; Kamel Boulos et al., 2011; Goodchild and Li, 2012). A visual inspection of Figure 2.1 shows that the monthly tick counts signal presents an irregular and noisy shape. A closer descriptive analysis of the raw data reveals that out of 3,073 records in the dataset, around one third of the samples are zeros, and a small proportion of samples present high peaks. Zero AQT means that a volunteer visited a site for tick sampling on a particular date and no ticks were caught questing, whereas a peaky AQT means the ticks were very active on that day.

To assess the potential impact that zero and peaky AQT may have in our modelling process, we created four versions of the original dataset, which vary in the amount of zeroes and peaky observations. In two datasets we removed all samples with a zero AQT within the tick season (i.e. 1st March until 31st October) and half of the samples with a zero AQT outside the season. This creates a group with two datasets with a reduced amount of zeros, and a second group with two datasets which are not modified with respect to the original. After this step, we applied a smoothing process to only one of the datasets of each group. We chose a Savitzky-Golay filter to mitigate the effect of peaky AQT in the modelling process, whereas the other dataset was kept with the original AQT signal. In this way, the modelling process accounts for the possible effect of extreme observations to fit the AQT signal, and helps distinguishing whether varying levels of noise is hampering the learning process of the chosen modelling algorithm.

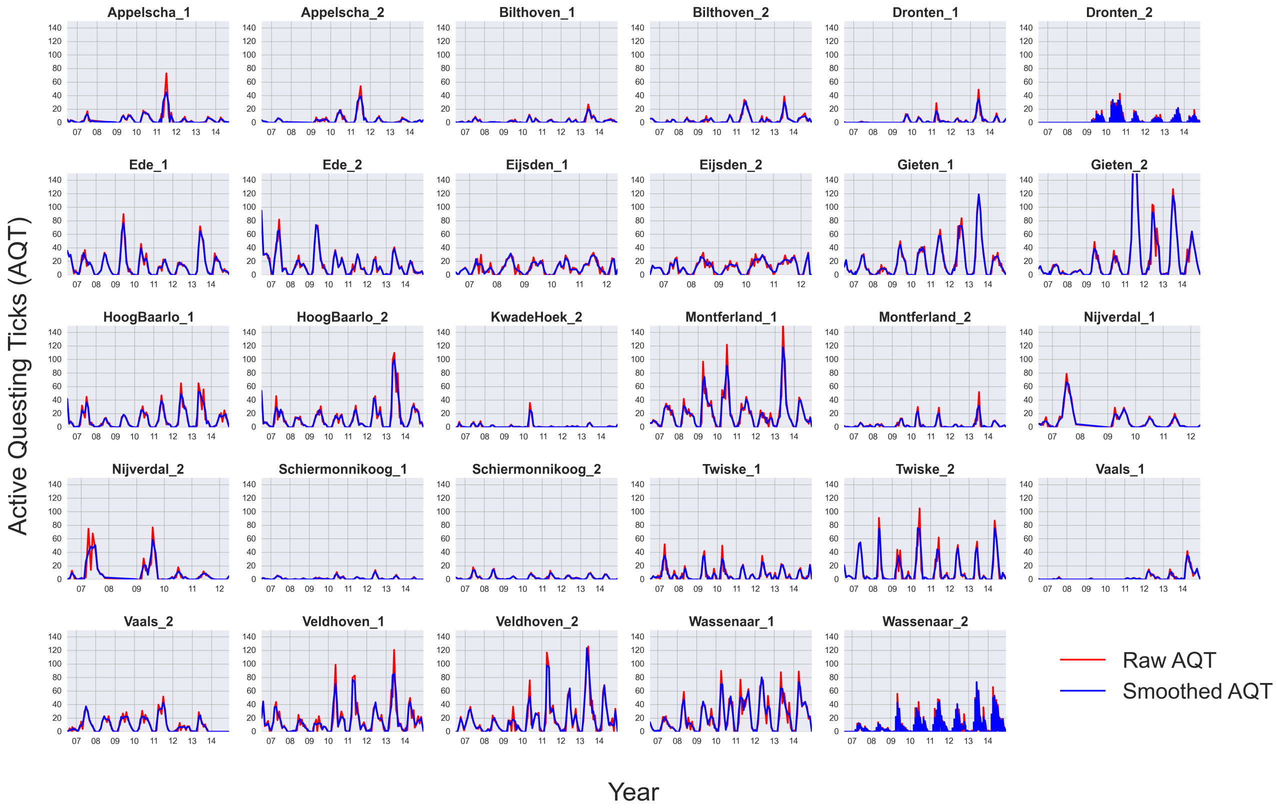

Figure 2.1: Monthly time-series (2006-2014) of active questing ticks (AQT) per transect. Each subplot shows both the number of ticks counted by the volunteer (red) and the Savizky-Golay smoothed version of this signal (blue).

2.3.2 Characterizing the environment

Feature engineering is a common process in the machine learning field to obtain new predictors from original data sources, which incorporate the knowledge of a domain to create predictive models. In our case, we obtained a set of features, based in the theoretical grounds described in Sections 2.2.1 and 2.2.2, which aim to that aim to characterize the environmental conditions in each tick sampling site. Thus, this work uses 101 features (Table 2.2) classified in five types: weather, remote-sensed vegetation, land cover, habitat and mast. Weather and vegetation features contain a value aggregated in a particular time window. Land cover, habitat and mast features contain the value of land cover in a point, the type of tick habitat in the sampling sites, and the strength of a mast year for three tree species, respectively. The remaining of this section describes how the features associated to each type were obtained from the original data sources.

Because of the lack of consensus in the literature on the optimal temporal unit(s) to model AQT, we created a suite of features by aggregating each weather variable (i.e. minimum and maximum temperature, precipitation, evapotranspiration, relative humidity, saturation deficit, and vapour pressure deficit) at multiple temporal scales. These temporal scales are defined by the number of days before the date of the tick sampling. The reason for doing this is straightforward: we assume the tick count produced today, depends on past weather conditions. Therefore, for each tick sampling date we calculated weather features using a range of 1 to 7 days before the sampling date (i.e. fine temporal units), and of 14, 30, 90 and 365 days (i.e. coarse temporal units). This procedure leads to 11 features per weather variable, adding up a total of 77 features (indices 16-92, type W).

Using GEE, we averaged the 3 to 4 images available per month to reduce the impact of clouds. Then, using the coordinates of each of the flagging sites, we obtained three (NDVI, EVI and NDWI) time-series summarizing the evolution of vegetation indices since 2005. To remove further noise in these time series, we decomposed each of them into their seasonal, trend and noise components. We kept the seasonal component and obtained the minimum value and range (i.e. width between the minimum and maximum values) per transect and vegetation index. This procedure creates 6 vegetation features (indices 93-98, type V) that condense the general vegetation and moisture conditions in the site over the time-series.

For the land cover, we reduced the number of classes to 12 due to two reasons: 1) the flagging sites are located only in certain types of land cover (e.g. deciduous, grasslands); 2) several land cover types are unrelated to the tick ecology (e.g. sweet water, saltmarshes) or can be aggregated to a coarser level (e.g. types of crop to agricultural land), thus can be unified in a single category. After re-classifying the LGN, the product was resampled to 500 and 1000 meters of spatial resolution using a majority filter. This process allows to account for the surroundings of each flagging site, and reduces the chances of the flagging site to be placed in a noisy pixel at

ID	Feature Name	Short description	Type
1	Litter	Thickness of litter layer	H
2	Moss	Coverage on a 1–10 scale of the moss layer	H
3	Herb	Coverage on a 1–10 scale of the herb layer	H
4	Brush	Coverage on a 1–10 scale of the brush layer	H
5	Tree	Coverage on a 1–10 scale of the tree layer	H
6	BioLC	Land cover as described in (Gassner et al., 2011)	H
7	Oak-Y	Strength of mast year in a oak forests	M
8	AOak-Y	Strength of mast year in American oak forests	M
9	Beech-Y	Strength of mast year in European beech forests	M
16	tmin-X	Avg. min. temperature in a time window	W
17	tmax-X	Avg max. temperature in a time window	W
18	prec-X	Avg. precipitation in a time window	W
19	rv-X	Avg. evapotranspiration in a time window	W
20	rh-X	Avg. relative humidity in a time window	W
21	sd-X	Avg. saturation deficit in a time window	W
22	vp-X	Avg. vapour pressure in a time window	W
93	min_ndvi	Min. NDVI value for a location in a year	V
94	range_ndvi	Range for NDVI value for a location in a year	V
95	min_evi	Minimum EVI value for a location in a year	V
96	range_evi	Range for EVI value for a location in a year	V
97	min_ndwi	Minimum NDWI value for a location in a year	V
98	range_ndvi	Range for NDWI value for a location in a year	V
99	lc25m	Land cover in a location at 25m spatial res.	L
100	lc500m	Land cover in a location at 500m spatial res.	L
101	lc1km	Land cover in a location at 1km spatial res.	L

Table 2.2: List of features involved in the current analysis. The features belong to the following categories: tick habitat (H), mast years (M), weather (W), vegetation (V) and land cover (L). The weather features are calculated at 11 temporal aggregations, so there are 77 weather features in total. The X character is replaced by a number between 1-7 for short-term temporal aggregations or by a 14, 30, 90 or 365 in the case of bi-weekly, monthly, seasonal or yearly temporal aggregation, respectively. Mast features have a Y character that will be replaced by a number between 0-2 in function of the mast year they are referring to. In total, there are 101 features involved in this work.

25m resolution. We obtained the value of the land cover for each of the flagging sites at these three different spatial resolutions (indices 99-101, type L). The strength of the mast year of the year of the observation, as well as the strength of the previous two years (indices 7-15, Type M) is included in our work, because tick dynamics might have a delayed response to mast years. Finally, the habitat characteristics per transect are described using 5 variables: the thickness of litter layer and the amount of moss, herbal, brush and tree layers, which are encoded in 5 features (indices 1-6, Type H).

2.4 Modelling AQT with Random Forest

Random Forest (RF, (Breiman, 2001)) is an ensemble learning method that can be used both for classification and regression problems. Ensemble methods rely on the creation of a committee of experts, which work on solving a real-world problem while minimize the chances of taking a poor decision. In the case of RF, the ensemble is formed by a group of weak learners called decision trees, which are combined to create a robust decision ensemble.

RF is a combination of the bagging growing scheme (Breiman, 1996) and the Random Subspace Method (Ho, 1998). These two sources of randomness contribute to create an ensemble with very different trees that lead to high variance predictions when tested individually (Louppe et al., 2013). Bagging allows RF to see multiple variations of the input data, whereas the RSM introduces randomness in the samples and features presented to each tree during the learning phase. This process creates an ensemble of trees, which is capable of adapting to the tick dynamics phenomenon, and yield predictions with great robustness and stability (Rodriguez-Galiano et al., 2014).

The mechanism used by RF to grow decision trees for regression problems, such as modelling AQT as a function of environmental features, is conceptually simple. For each tree (B), N bootstrap samples (with replacement) are drawn from the available training data. This subsample is used to grow a unique decision tree (T_b) by recursively partitioning the N samples until a stop condition is reached, namely: 1) all the samples within a node have the same target response target; 2) the samples in the node are homogeneous with respect to the selected features; 3) a heuristic, such as the maximum depth of the tree, is reached. If none of these conditions are met, the algorithm grows the tree by selecting the best feature and split point among a given and random subset of training features, where best means that it minimizes the Mean Squared Error (MSE). This process creates two child nodes, and the available samples are assigned to them considering the split criteria (e.g. samples with values for feature m larger than the split point value go to the left child node). This procedure is repeated until a full forest with B trees is grown.

After completing the training phase, RF predicts unseen samples by averaging the predictions of the B trees. This reduce the variance and the generalization error of the predictions. In fact, the generalization error converges as the number of trees increases, thus reducing the chances of overfitting data (Breiman, 2001; Rodriguez-Galiano et al., 2014).

A key characteristic of RF is that it provides a measure of the importance of the features involved in the modelling. This is done by averaging the reduction in MSE associated to the use of each variable in each of the nodes/trees that form the ensemble (Louppe et al., 2013). In our work, we exploit this characteristic to understand the main drivers of tick dynamics. Thus, the ranking of features provided by RF gives an idea about what the most relevant (or irrelevant) features are, regardless of the dimensionality of the problem. This is particularly suitable to understand the complex and non-linear interactions found in biological and environmental systems.

RF, like most data-driven regression methods, are not time-aware models. This means that its standard application to regression problems involving (seasonal) time-series, such as the AQT dataset, can lead to sub-optimal results. The reason for this is that the trees in the RF ensemble are trained with random subsets of the training set, where each data sample belongs to a particular date. Thus, RF is trained to predict single snapshots and remains unaware of the temporal continuity of the time-series.

In this work we overcome this limitation by introducing time-awareness in RF. To do so, we transformed the AQT counts into monthly Z-scores by: 1) grouping the 9 years of observations according to the month when they were collected; 2) calculating monthly means and standard deviations, after removing extreme observations from each group so that the Z-scores are not biased. In this context, extreme observations are those that report AQT counts above the 3rd quartile or below the 1st quartile of the monthly values; 3) creating monthly Z-scores, by subtracting the monthly mean from each observation, and dividing the result by the corresponding monthly standard deviation. In this way, we ensure that samples collected during the same month have a constrained and normalized range of AQT counts.

With this monthly normalization we train RF to understand which factors increase or decrease AQT with respect to the long-term average, instead of modelling the absolute number of ticks recorded in a particular location and month. Moreover, by predicting monthly Z-scores we help RF to understand the temporality of the data and hope to get more realistic seasonal dynamics than by using the classical (single snapshot) RF model.

The general set-up of the RF models was as follows: 1) we reserved 70% of the data for training, and the remaining 30% was used for testing the model. Samples were randomly assigned to the training and test subsets; 2) To account for the randomness of RF (different features and samples used in each tree/run), we executed the models 10 times (keeping the training and test samples constant) and the error metrics and feature importance were averaged; 3) we use two well-known statistical metrics to validate our results, the root mean squared error (RMSE) and the normalized RMSE (NRMSE). Note that the error metrics were obtained after de-normalizing the Z-scored signal. Finally, we used the trained model to prepare maps illustrating its performance in a country-wide scenario.

2.5 Experiments

The process described in Section 2.3.1 creates four versions of the original dataset with a varying number of zero and peaky AQT, and the process of feature engineering from Section 2.3.2 enriches each of the volunteered observations with 101 features. With this set-up, we designed the tree experiments explained in the next sub sections,whose goal is: 1) to assess the impact of noisy observations and selecting the best model capable of capturing AQT dynamics; 2) to evaluate the most important features to model AQT at different time scales; 3) to create AQT map for forested areas in the Netherlands.

2.5.1 Model selection by assessing the impact of noisy AQT

We modelled the four versions of the volunteered AQT dataset with our time-aware version of RF. Figure 2.2 shows the general performance of the models. To ease the interpretation of these results, three elements are included: 1) a 1:1 line showing the ideal predictions; 2) a grey band showing one standard deviation from the mean of the observations; 3) a grey box containing the selected statistical metrics for this experiment. The visual inspection of the four plots shows that the two experiments using raw data perform poorly when compared to the two experiments using smoothed data. The models built with raw data have the highest errors in terms of RMSE and NRMSE and also present a higher dispersion of the predictions, indicating that these models did not properly capture the peaky AQT observations.

A close inspection of the NRMSE metric reveals that RF models with smoothed data present very similar performances, regardless of the number of zeros left in the AQT dataset. This suggests that smoothed models can capture the conditions yielding low AQT, but peaky AQT may be actually hampering the modelling process. This is clearly visible when inspecting the points falling outside the gray band in the bottom subplots: a certain number of high AQT true observations could not be captured by the model, thus producing a lower prediction than the true value. We selected the model for next experiments based on the lowest RMSE and NRMSE metrics, thus, out of the four models, we picked the one keeping zero AQT and smoothing the peaky AQT with the Savitzky-Golay filter.

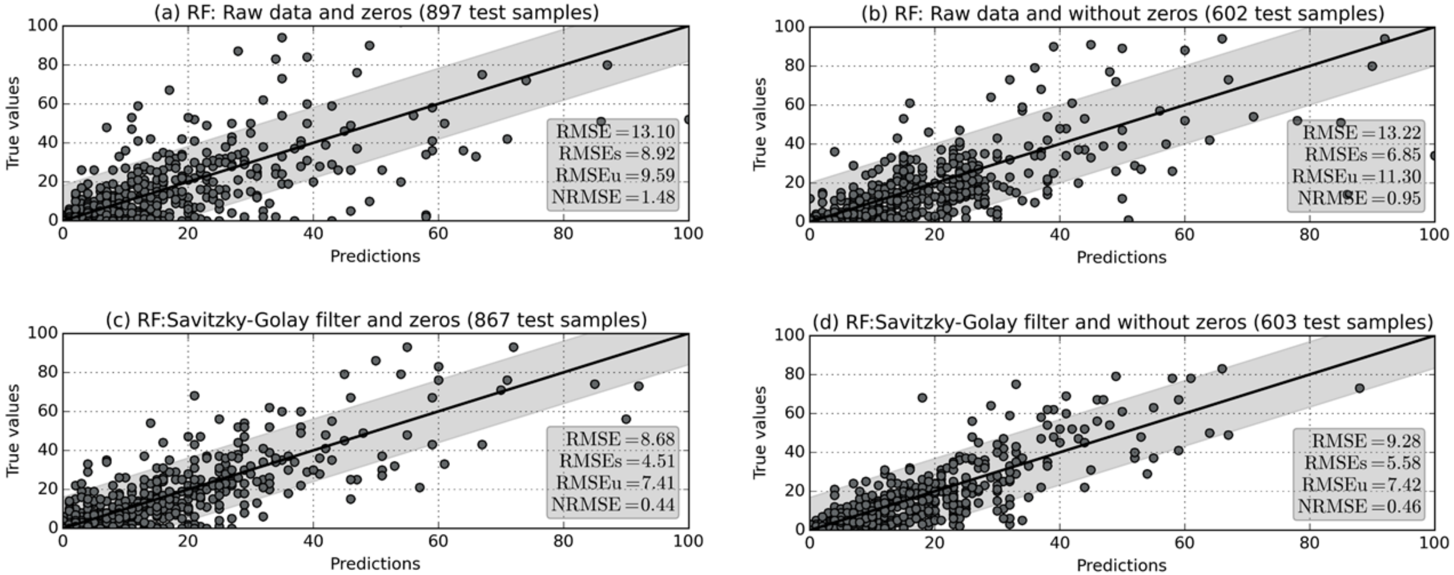

Figure 2.2: Performance of RF in each of the four selected scenarios. The X-axis represents the predictions yielded by the model and the Y-axis the true values measured by volunteers. To assess the quality of the volunteered AQT dataset, we have tested RF with varying levels of zero AQT and peaky AQT. As seen, the model has more difficulties in capturing peaks than zeros. Thus, the selected model for the following experiments is the model at the bottom left, because it presents the lowest metrics.

Table 2.4 presents the feature importance of the top 10 features for the selected RF model. To ease the interpretation of results, we restrict the ranking of the feature importance to the top 10 most prominent out of 101. As seen in this table, the modelled phenomenon is driven by a combination of several weather variables and a vegetation one. The two most explanatory features are the annual evapotranspiration (i.e. ev-365) and the monthly relative humidity (i.e. rh-30). Temperature, which has been traditionally spotted in tick modelling studies as a major driver of tick dynamics, only appears once (as tmax-365) and with a relatively low importance. In this experiment, water-related features perform better than temperature. Note that evapotranspiration and relative humidity do appear several times in the ranking (i.e. ev-90, rh-365), suggesting that in a context with multiple atmospheric variables, water-levels are again more important than temperature. It is also important to highlight that variables about mast years or tick habitat do not appear in the top ten. This could be because they are static (i.e. one value for the whole study period) and, hence, unable to explain the temporal and spatial variation in seen in the AQT dataset.

To further evaluate the usefulness of our RF-based model to predict AQT, we split the test samples according to their associated transect (cf. Figure 2.3). The goodness of the fitting (R^2) between the predictions and the real smoothed AQT values varies between 0.19 and 0.94, indicating that the performance of the model strongly depends on each transect. In Figure 4 (left) we sort the R2 values to provide a better depiction of the performance of the model per transect. Based on these results, we note that the model presents a moderate-to-strong R2 (i.e. $0.7 < R2 < 1$) for roughly half of the sites. This means that these transects better respond to weather variables than the remaining transects, in which AQT may be driven by variables, such as wildlife, not included in the current model. Note that for transects within the same site (e.g. Vaals, Montferland) the goodness of fit is very different, revealing the very local nature of AQT. Figure 2.4 (right) shows the geographic representation of the transects. Symbols in green represent the transects better responding to weather variables, whereas red symbols represent the poorly fitted transects. The visual inspection of this figure shows no strong spatial pattern (e.g. north-south gradient).

Position	Feature	Importance
1	ev-365	15
2	rh-30	11
3	tmax-365	7
4	prec-90	4
5	prec-3	4
6	ev-90	3
7	rh-365	3
8	tmin-365	2
9	prec-365	2
10	tmax-90	2

Table 2.4: Ranking of the top ten most important features (out of 101) for the selected RF model. The sum of the feature importance for all features provided by RF equals to 1, but to ease the interpretation of results we multiplied it by a hundred to have natural numbers. As seen, features involving atmospheric water levels (i.e. evapotranspiration and relative humidity) are found to be important to predict tick activity, since they appear several times in the current ranking.

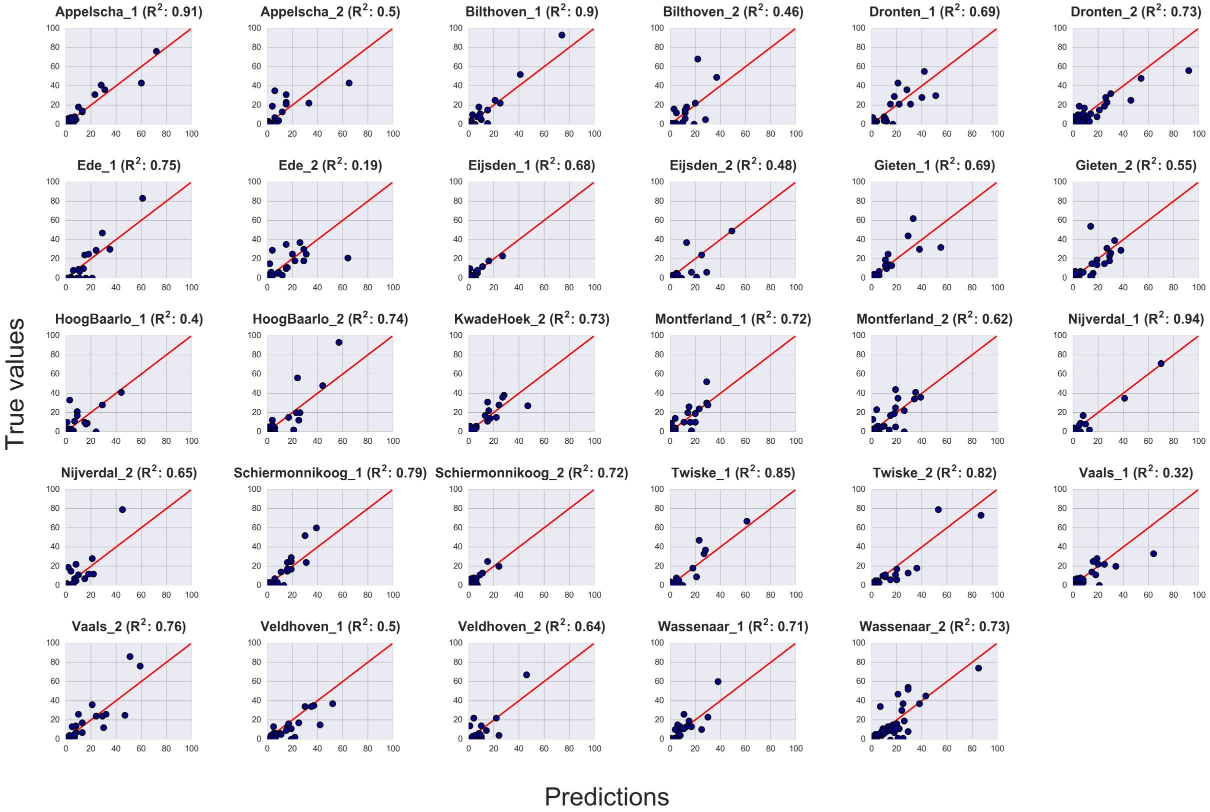

Figure 2.3: Performance of the selected RF for each flagging site. The X-axis represents the prediction yielded by the model and the Y-axis the true values measured by the volunteers. The R2 value is provided in the title of each subplot.

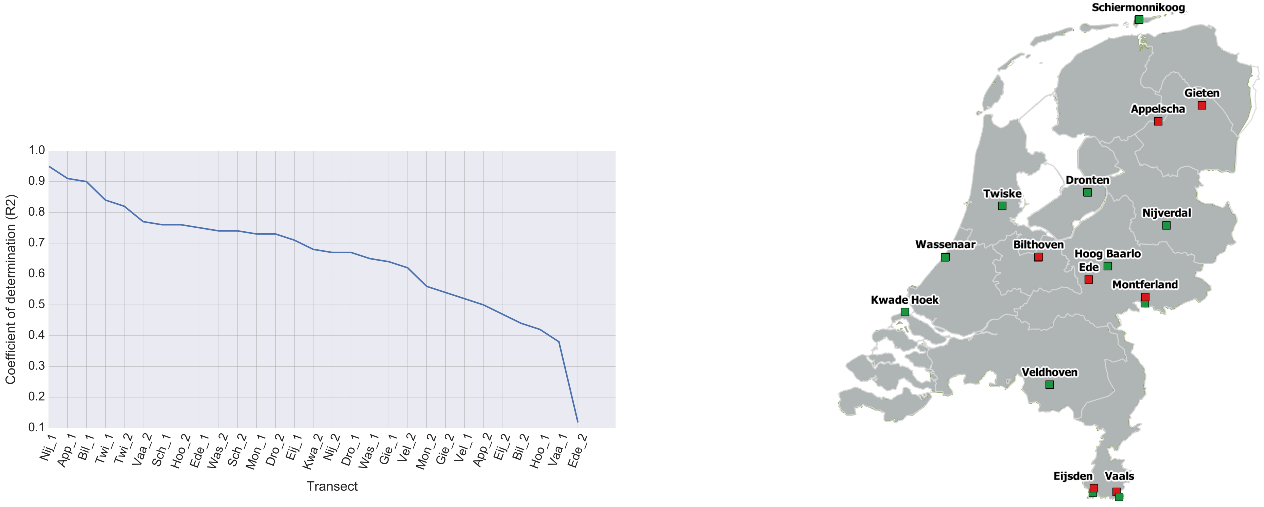

Figure 2.4: Performance of the selected RF model per transect sorted by the R^2 score (left) with its geographic location (right). The left image shows that the model is able to fit half of the sites with a moderate-to-strong R^2 (i.e. $0.7 < R^2 < 1$), whereas the performance in the remaining sites is weak-to-moderate (i.e. $0.3 < R^2 < 0.7$). This means that the transects with a high R^2 score, respond better to weather variables than the rest of the transects, which may be driven by variables not included in the current model (e.g. wildlife) due to inavailability. The right figure shows the geographic location of the transects: green squares represent transects with moderate-to-strong fitting, whereas red squares represents transects with weak-to-moderate fitting.

	1		2		3		4		5		6		7		14		30		90		365	
Rank	**F**	**I**	**F**	**I**	**F**	**I**	**F**	**I**	**F**	**I**	**F**	**I**	**F**	**I**	**F**	**I**	**F**	**I**	**F**	**I**	**F**	**I**
1	EV	15	EV	15	EV	15	EV	15	TX	16	TX	14	EV	14	TX	14	TX	15	RH	17	RH	16
2	TX	13	TX	14	TX	14	TX	14	EV	13	EV	13	TX	13	TN	13	PR	14	EV	15	TX	15
3	TN	11	TN	12	PR	13	PR	11	TN	12	P	13	TN	12	EV	12	RH	12	PR	14	PR	13
4	RH	11	RH	11	TN	12	TN	11	PR	11	RH	12	RH	12	PR	12	TN	12	SD	9	EV	13
5	PR	10	PR	9	RH	10	RH	10	RH	10	TN	9	PR	11	RH	11	EV	11	TN	9	TN	12

Table 2.6: Ranking of the top five features for the selected RF model across all temporal scales. Each feature is accompanied by its importance, which has been calculated as the mean of 10 runs. Features involving atmospheric water levels (i.e. evapotranspiration and relative humidity) are found to be relevant to model tick activity in all temporal scales. These results are consistent with the ones provided by the general model. Interestingly, evapotranspiration is marked as the most relevant feature in very short-term time scales, whereas relative humidity is a better predictor for long time scales.

2.5.2 Feature importance across multiple time scales

The model structure selected in the previous section is used here to find out the best temporal scale to model AQT. To do so, we train one RF model for each of the 11 time scales described in Section 2.3.2 and we execute the model with a subset of features of the input dataset: we keep all the non-weather features (a total of 24 features) and we add the weather features corresponding to that particular time scale (7 features). Thus, we run the modelling process 11 times with 31 features, providing at each iteration the feature importance. In this way, it is possible to get new insights about whether the importance of the features to model AQT change over increasing temporal windows, which might guide the choice of a particular time scale to model AQT optimally.

Table 2.6 shows the importance of the features at multiple time scales. Each column of the table shows the top five most important features (out of 31) for each of the selected time scales. To ease the description of results at multiple time scales, we restrict the ranking of features to the most relevant top five. This table shows that the most explanatory features for all time scales are weather-based ones and that non-weather features (i.e. vegetation, land cover, tick habitat, mast years) do not significantly contribute to model tick dynamics. Evapotranspiration, relative humidity and the maximum temperature appear to be the most important features. EV better performed in the short-term experiments (i.e. temporal aggregation from 1 day to 4 days before the sampling date), whereas RH is the best one in the long-term experiments (i.e. seasonal and annual temporal aggregation). TX appears to be the best predictor in the remaining experiments.

2.5.3 Mapping tick dynamics

The RF model selected in Section 2.5.1 was used in a country-wide exercise to produce three map products: the mean and standard deviation of AQT for the year 2014, and the AQT on a date expected to be close to the peak activity of ticks in nymphal stage. We selected the year 2014 because it is the last of the AQT time-series.

Since the flagging sites are located in forested areas, we identified forested pixels in the land cover map and extracted their locations. Then, to this selection of pixels, we applied the process of feature engineering described in Section 2.3.2 for each day of the selected year, thus obtaining 365 country level datasets. The model was retrained with the 86 features available at the country level (i.e. remove tick habitat and mast year features) and tested with the newly created datasets for forested pixels. The predictions yielded by the model were transformed into raster format to obtain three products: first, we obtained the annual mean and standard deviation of AQT based on the daily computed values, which identifies at the country level regions with higher or lower tick activity; second, we obtained the temporal profile of the pixels containing the sites to visualize the daily seasonality of AQT; third, we mapped the AQT for a particular day of the year, which is expected to be close to the peak of tick populations in the nymphal stage.

Figure 2.5 illustrates the annual mean (left) and standard deviation (right) of predicted AQT. A visual inspection of the mean map shows that there are more AQT in the eastern half of the country (i.e. orange to red regions), especially within the provinces of Overijssel and Drenthe. The standard deviation map depicts the spatial variability of the predictions: regions in light green and yellow show areas where the predictions oscillated significantly above or below the mean AQT, whereas regions in dark blue show locations where the prediction is stable throughout the year. Figure 2.6 shows the daily temporal evolution of AQT for the grid cells where the flagging sites are located. This figure also shows three additional elements: the long-term monthly average for all sites obtained from the boxplots, the long-term monthly average for the site, and the 2014 monthly average for the site. Note that there are 15 sub plots, because each grid cell overlaps the two transects. This allows to visually identify sites whose predicted AQT is (dis)similar to the averages. Figure 2.7 shows the predicted AQT for June 1st, which we expect to be close to this peak population of nymphs. As seen, the highest predictions of AQT are predicted in the east half of the country, but there is another spot of high activity in the southern province of Noord-Brabant.

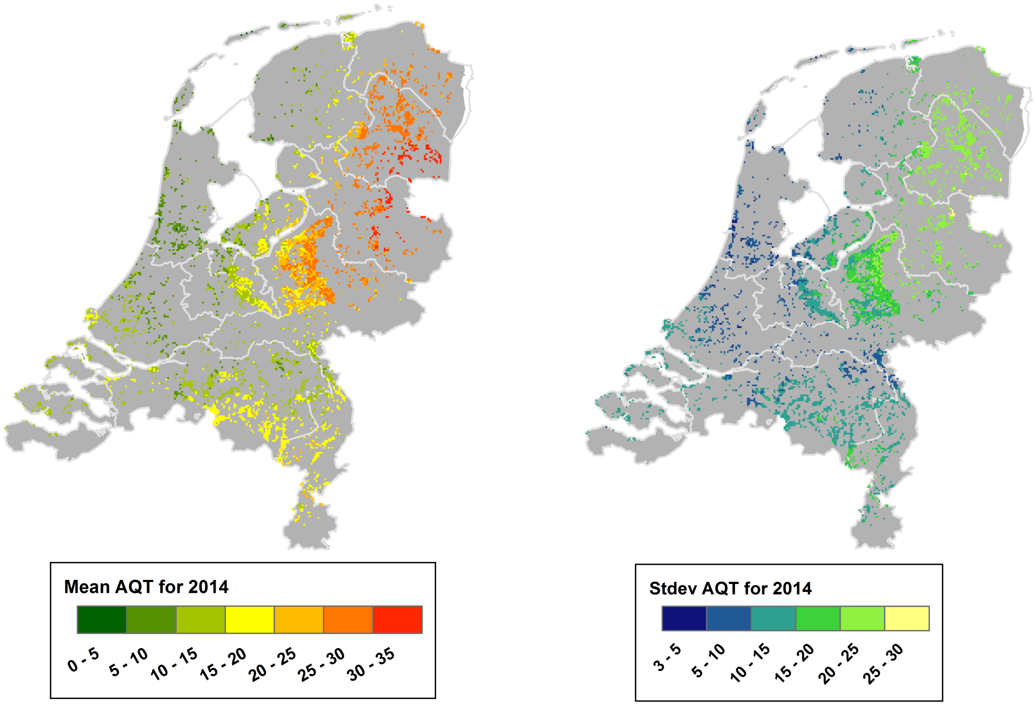

Figure 2.5: Predicted mean (left) and std. dev. (right) of AQT for 2014. The map of mean AQT ranges between 0 (green) and 32 (red) and shows how the highest values of tick activity during 2014 are concentrated within the provinces of Overijssel and Drenthe. The map of the std. dev. of AQT ranges between 3 (dark blue) and 26 (yellow) and shows the deviation from the mean of all forested pixels in the country. The east half of the country presents higher variations, indicating that tick activity in this areas may change significantly. Predictions along the coastal provinces (i.e. Noord-Holland, Zuid-Holland and Zeeland) do not seem to deviate significantly from the predicted mean.

2.6 Discussion

Ensemble learning algorithms such as RF can model non-linear relationships in complex natural processes. This data-driven algorithm provides an indication of the relative feature importance of the predictors involved in the analysis. This is particularly useful to model processes in which the main drivers are unknown. RF has a robust and stable behavior when handling potentially noisy data, a condition often occurring in volunteered datasets, like our AQT time-series. However, since RF is not a time-aware method, it performs sub optimally when modelling seasonal phenomena. In this work, we provide a methodological innovation to introduce time-awareness in RF by transforming our target AQT signal into a monthly bounded one, thus helping the trees in the ensemble to distinguish time.

The study on the importance of the features show that water-related features (i.e. evapotranspiration and relative humidity) are better predictors of the tick activity than temperature or vegetation. This suggests that the tick activity may be driven by atmospheric water levels, which are crucial for tick survival. A closer look of the model performances regarding statistical metrics, indicate that the model can fit the AQT signal for half of the transects, but the fitting decays for the other half. A hypothesis that may explain this difference is that the tick activity may be driven by different variable depending on the geographic location: the sites with a higher R2 score might be strongly influenced by atmospheric conditions, whereas the sites with a lower R^2 score might be driven by variables currently not included in the model (e.g. wildlife). In addition, the analysis of the feature importance at multiple temporal scales is consistent with the results of the general model, because water-related features are spotted as the most prominent features across all temporal scales.

The two major hurdles encountered during this experiment are related with the weather uniformity in the country and the low spatial resolution of the available environmental datasets. First, the low elevation and the small size of the Netherlands make the country very uniform in terms of weather and vegetation variables (e.g. reduced north-south temperature gradient, high and persistent "greenness"). This means that vegetation indices, often included in previous studies in the field, are not discriminative enough to model AQT. Second, the available weather and vegetation datasets have a spatial resolution which is too coarse to model tick dynamics, a phenomenon which was found to be very local as suggested in (Estrada-Peña et al., 2012). This might have an impact when characterizing the AQT with the environmental datasets: different locations with similar weather conditions yield an uncorrelated number of AQT, masking the relationship between weather and ticks and increasing the errors of the model. To mitigate these effects and decrease the average error of the models, we recommend using weather datasets at a finer resolution or involve more volunteers in this long-term citizen science project to get more data.

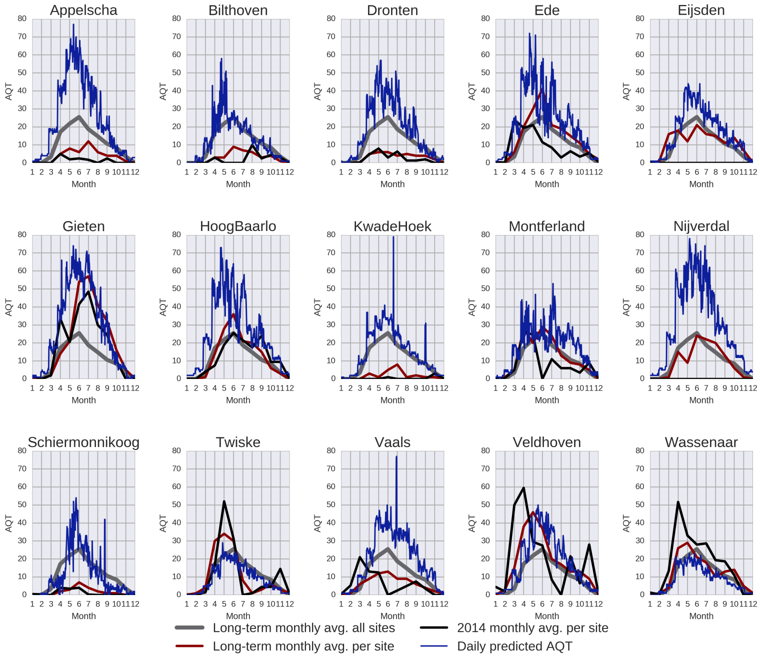

Figure 2.6: Daily predicted AQT for the locations of the flagging sites. Each subplot contains four curves: the long-term monthly average (grey) across all sites, the long-term monthly average for the flagging site (dark red), the 2014 monthly average for the flagging site (black), and the daily predicted AQT for 2014 yielded by RF (blue). Note that in 2014, the sampling stopped in Nijverdal and Eijsden.

2.7 Conclusion

Citizen science initiatives allow monitoring of environmental phenomena via crowdsourcing, and produce geospatial data collections that can support scientific analysis. The question at the beginning of this work was whether the collective effort carried out by a group of volunteers, would translate into predictive models estimating tick activity in the Netherlands. Results show that combining volunteered AQT data with environmental variables and modelling them with a time-aware version of RF, can capture most of the spatial and temporal variation in the number of active questing ticks in the country.

The combined analysis of volunteered AQT and environmental variables consistently spotted that water-based features, especially evapotranspiration, play a crucial role in predicting AQT. In this sense, further studies in the field of tick ecology should consider adding, besides the classical temperature and vegetation indices, water-based features. Aside identifying the most

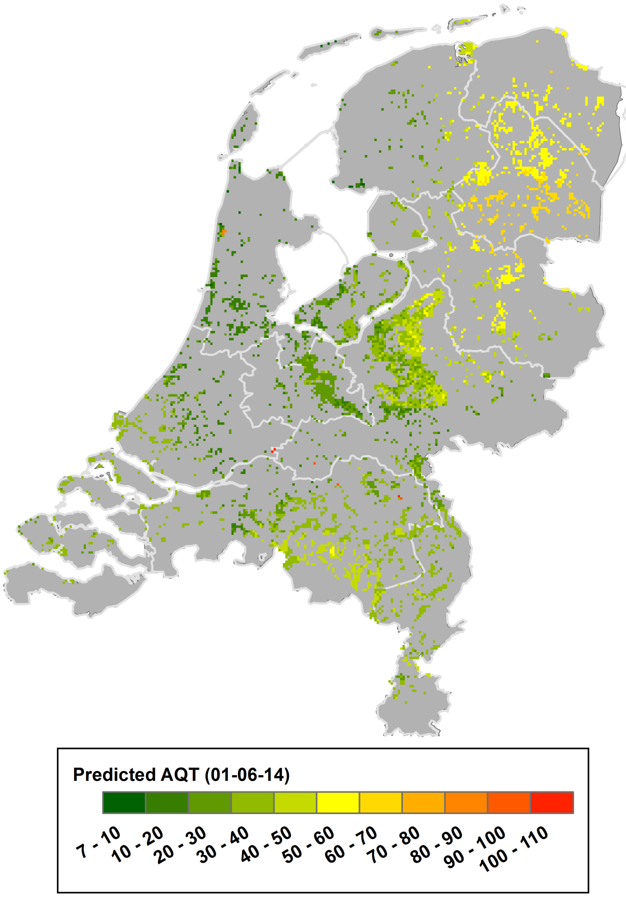

Figure 2.7: Predicted AQT at the country level for June 1st of 2014. The map ranges from low values of tick activity (dark green) to high values of tick activity (red). The peak populations of ticks in the nymphal stage reaches its maximum between May and June depending on the weather conditions of the year. Thus, we expect this date to be a close depiction of the tick activity at its maximum. The highest AQT are predicted in the east half of the country, in particular within the province of Drenthe, whereas coastal regions present lower levels of tick activity.

important variables to model tick dynamics, this study has produced a model that, scaling up from volunteered observations, can map daily tick activity at the country level. The use of this model may open the way to study spatial patterns and seasonal trends at the national level, not only tick activity, but also of other non-linear natural phenomena, such as phenological events or species distributions.

With these new insights, we envision different applications in the field of tick related ecological research, nature management and public health. In ecological research, our tick activity model allows the identification of tick hotspots and of the sampling sites where the model fit was good or bad, which can be used to better select new monitoring sites. With the model we can better analyze the impact of extreme weather events and climate change on tick dynamics and population development. In nature management, these maps can help owners of green areas to be more aware of the variation of tick dynamics in the areas they are responsible for, which can lead to a better planning in space and time of different forestry management activities. In tick hotspots with many visitors, nature managers could consider to more frequently mow the grass directly next to walking and cycling trails or picnic areas to try to reduce tick populations. In public health, this model can be used to better inform people that visit natural areas on current tick activity levels. In combination with weather forecasts also detailed forecasts for the coming days can be given. The proposed model predicting tick activity will replace the very basic tick activity forecast currently implemented in the Dutch citizen science website Tekenradar (Tick radar, www.tekenradar.nl).The spatially detailed tick activity forecasts are expected to raise awareness among the general public and many different stakeholders involved in the problem of Lyme disease. Having more detailed information hopefully translates in an increase of protective and preventive measures when visiting forested areas. Overall, we expect that a better understanding of tick dynamics may contribute to design interventions to reduce the incidence of Lyme disease.

3 Identifying environmental and human factors associated with tick bites using volunteered reports and frequent pattern mining

3.1 Introduction

Tick populations and tick-borne infections like Lyme borreliosis are steadily increasing since the mid-1990s. This concurrent increase has been observed in different European countries (Heyman et al., 2010; Jaenson et al., 2009), in the United States (Subak, 2003; Tuite et al., 2013) and in Canada (Ogden et al., 2014). In the Netherlands, more than 20,000 people per year develop Lyme borreliosis (Hofhuis et al., 2015b), and its disease burden is substantial, especially in patients that develop long-term persisting symptoms.

Global environmental change is pushing the distribution of ticks northwards, modifying its spatio-temporal dynamics and increasing the abundance of ticks in nature (Medlock et al., 2013). Previous research efforts in the field of mathematics, biology and environmental modelling have found a tight relationship between wildlife (Ostfeld et al., 2006), environment conditions (Swart et al., 2014; Tack et al., 2013), weather (Medlock et al., 2013) and tick populations. However, the increase of tick populations does not necessarily translate in more tick bites, because we need to account for outdoor human recreational activities, which are a crucial factor in Lyme disease monitoring. Outdoor recreational activities (e.g. visit to natural areas, parks or gardens) increase the exposure of humans to ticks and also the chances of getting a tick bite while performing outdoor activities. Thanks to advances in ICT it is now possible to collect geospatially-enabled tick bite reports and start citizen science projects with information directly provided by people that

This chapter is based on: (Garcia-Martí et al., 2016)
Available at: https://doi.org/10.1111/tgis.12211

have been bitten by a tick. This geospatially-enabled data is also known as Volunteered Geographic Information or VGI (Goodchild, 2007a).

Despite the well-known problems associated to VGI, such as quality and accuracy (Goodchild and Li, 2012; Mehdipoor et al., 2015) a remarkable number of volunteer projects have been created or re-invigorated in the last decade. Comparatively however, the number of efforts incorporating volunteer information to a scientific workflow is still limited (Mehdipoor et al., 2015; Rosemartin et al., 2015). In 2006, Wageningen University started collecting volunteer tick bites through the educational phenology platform Natuurkalender (NK; 'nature's calendar', www.natuurkalender.nl), gathering nearly 10,000 volunteered tick bites in six years. This pioneering project attracted the attention of the Dutch National Institute for Public Health and the Environment (RIVM) and in 2012, the platform Tekenradar (TR; 'tick radar', www.tekenradar.nl) was launched together with Wageningen University. TR is a web platform especially conceived to inform citizens about the risk and prevention of tick bites and at the same time a citizen science platform to collect volunteer tick bites and erythema migrans [1] observations. In a bit more than three years, TR has collected nearly 25,000 volunteer observations. Both projects, to the best of our knowledge, are the first citizen science project in the world that specifically focuses on ticks and tick-borne diseases. At present there is an insufficient understanding of the factors that determine the tick bite risk. Previous research efforts based on volunteered tick bite reports have identified risky landscapes, activities and vulnerable age groups (Mulder et al., 2013). However, no studies have simultaneously analyzed the importance of weather data, remotely-sensed vegetation indices and volunteered data to identify environmental and human factors associated to tick bites and to map them in the geographic space. In this research we will tackle these issues by identifying frequent environmental patterns associated to tick bite reports. This research opens the way for creating a national tick bite risk map and for designing interventions that will decrease the number of tick bites and reduce the incidence of Lyme disease.

3.2 Data and methods

3.2.1 Volunteered tick bite reports

This study is based on the collection of tick bite reports collected by the NK and the TR initiatives. The NK dataset contains 9,256 observations registered between 2006 and 2012 and the TR dataset contains 24,584 observations registered from March 2012 to November 2014. This relatively large number of tick bites reports was achieved because both initiatives received substantial [2] in the Netherlands (van Vliet et al., 2014). The TR attracted

[1] Red rash around the tick bite, indicative of infection

[2] https://www.rivm.nl/documenten/tekenradarnl-webplatform-over-tekenbeten-en-ziekte-van-lyme

more observations than NK mainly due to the fact that TR also provided the possibility to send in ticks for analysis and because it included a daily and dynamic spatially specific tick activity forecast that attracted lots of media and visitors.

The tick bite reports were first filtered to remove records without a geographic coordinate (i.e. volunteer did not specify a location) and to remove records outside of the boundaries of the Netherlands. After that, a total of 28,865 valid observations were found. Figure 3.1 shows the geographical distribution of all the reports for the period 2006-2014. There is a clear spatial clustering along the coast and in the center part of the country, which are likely driven by human activity patterns during leisure time in combination with tick densities in nature; *de Hoge Veluwe* National Park, the *Utrechtse Heuvelrug* forest, the recreational areas of the West Frisian Islands and along the coast, are typical tick habitats and have a high human recreational pressure.

The filtered collection of tick bite reports was used to extract the volunteered-related data reported along with the tick bite report: date, year of birth, address, type of environment where the tick bite occurred (e.g., forest, garden, park) and activity the person was carrying out when he/she got bitten by a tick (e.g., walking, picking fruits, gardening). See Section for more details about these categories. The address field is free-text and volunteers report their home address with more or less detail. Therefore, some records contain a postal code with 6 digits (which identifies a range of houses or buildings in the same street) and some others have a coarser resolution, by only indicating the neighborhood in a city or the city itself. To create a more standardized data set, we used geocoding services to extract the postal code, street address, municipality subdivision, municipality and province whenever possible. More specifically, we used the geocoding service from the Dutch National Spatial Data Infrastructure (PDOK[3]). This service exposes a REST interface that receives a string of data and returns a XML file with the structured address belonging to this string.

Finally, we re-projected the original coordinates of the TR dataset, which were in decimal degrees (EPSG:4326, WGS84), to the official coordinate system of the Netherlands (EPSG:28992, RD_New). The reason for this change is that NK reports were already in RD_New and also because GIS software is usually faster with projected coordinates.

3.2.2 Environmental data

Tick-related literature indicates that temperature, precipitation, soil moisture, air humidity and vegetation type determine both tick presence and tick dynamics in a given area. In the following sections we describe the environmental datasets used to represent these variables.

[3] https://www.pdok.nl

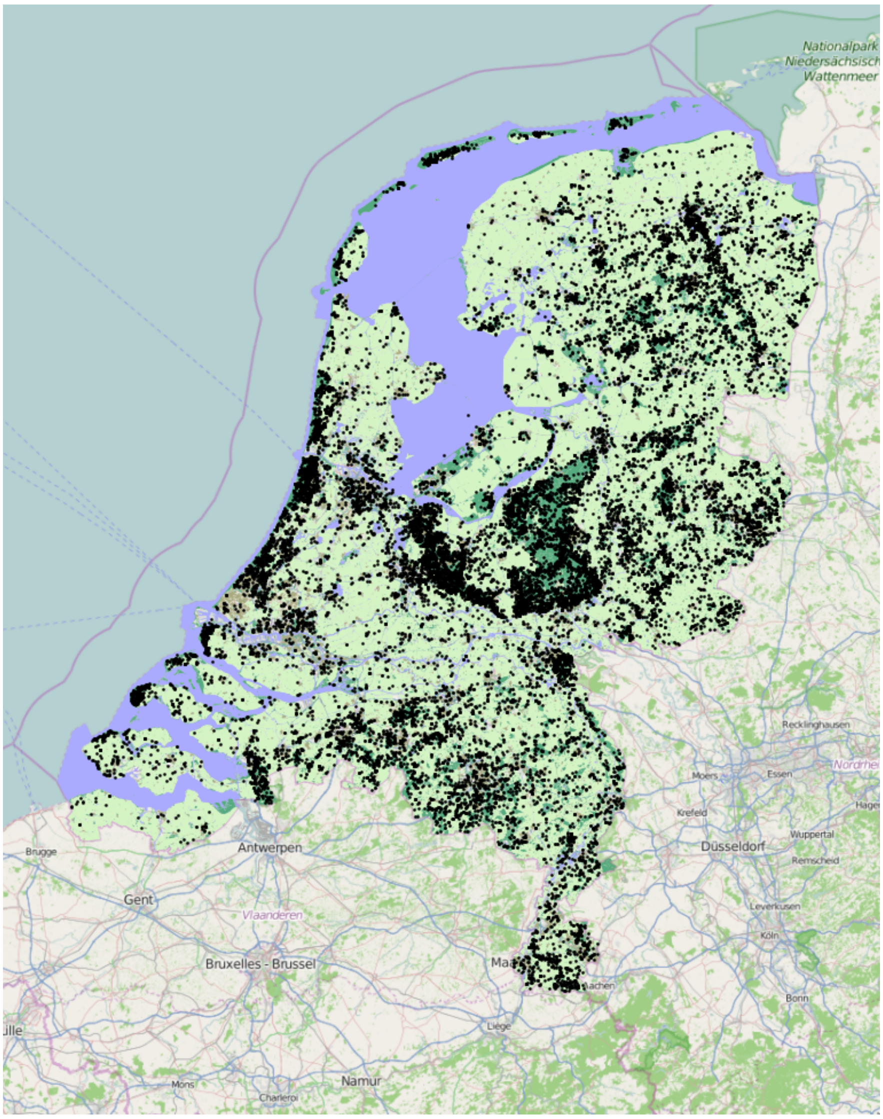

Figure 3.1: Geographic projection of the volunteer tick bites observations for the period 2006-2014. Each black dot represents a tick bite report.

3.2.2.1 Weather data

Weather data was obtained from the Royal Netherlands Meteorological Institute (KNMI [4]). Concretely, we used daily gridded (1 km) temperature and precipitation data for the period 2006-2014. These weather layers are an interpolation of 34 automatic weather stations scattered across the country. Temperature determines the start of the questing season, tick population development rate and the survival probability through winter (Ogden et al., 2006; Randolph et al., 2008). Precipitation is necessary during the hot summer season to sustain tick populations (Bennet et al., 2006). Besides that, precipitation during spring (especially in May) may contribute to increase the prevalence of tick bites and Lyme disease during summer (Jore et al., 2014). Lyme disease cases in the northeastern United States were positively correlated with precipitation during May and June of the same years, but not with the precipitation of the previous ones (McCabe and Bunnell, 2011).

3.2.2.2 Vegetation indicators

Ticks are particularly sensitive to the thickness of forest canopy and soil moisture at the litter level (Medlock et al., 2013). Earth observation satellites enable monitoring these environmental conditions over large areas. This is why tick populations have been modeled by combining weather variables and satellite-derived vegetation indices (Barrios et al., 2012; Estrada-Peña et al., 2012). In this research, we used three vegetation indices: the Normalized Difference Vegetation Index (NDVI), the Enhanced Vegetation Index (EVI), and the Normalized Difference Water Index (NDWI).

NDVI has traditionally been used to measure vegetation greenness and density. Previous studies have shown that the fluctuations in NDVI correlate well with the number of nymph and adult ticks at different moments of the year (Estrada-Peña, 2001; Randolph, 2000). More recent studies show that novel vegetation indices like the EVI are better indicators of tick abundance (Estrada-Peña et al., 2011) and that indices that measure the water content of vegetation (e.g. the NDWI) might outperform NDVI and EVI for tick population modelling (Barrios González, 2013). All three vegetation indices were obtained for the period 2006-2014 from the Google Earth Engine [5] [6] (GEE) platform. GEE is a free image processing cloud platform for environmental analysis, which aggregates products coming from different Earth observation sensors, such as the Moderate-Resolution Imaging Spectroradiometer (MODIS). MODIS provides daily global imagery at 250, 500 and 1000 meters of spatial resolution. However, due to the persistent cloud coverage over the Netherlands we used MODIS composite products. In particular, we used the MCD43A4 product, which provides the NDVI, EVI and NDWI indices as derived from daily surface reflectance at 500m spatial resolution and using data of the previous 16 days. This product is released every 8 days so there is a 50% temporal overlap between each composite.

[4] https://data.knmi.nl/portal/KNMI-DataCentre.html

[5] https://ee-api.appspot.com

[6] https://earthengine.google.org

3.2.2.3 Land use and soil data

Previous works have shown the influence of land use and landscape composition on the spatial distribution of Lyme disease (Lambin et al., 2010). Land use is related to human exposure because it determines human presence in a particular location, e.g. recreation, agriculture, urban area (Linard et al., 2007). The chances of humans being bitten by ticks depend on the exposure and interaction of people with different land uses, especially forested areas (Vanwambeke et al., 2010). Soil type favors the growing of certain types of trees that, in turn, are more or less capable of sustaining wildlife. For an instance, previous works have shown that coniferous stands tend to present lower tick densities when compared to deciduous forests (Tack et al., 2012).

In this study, land use information was obtained from a map produced by the Dutch Central Bureau for Statistics (CBS). This map (BBG2008) contains 14 different types of land use, however, here we grouped classes referring to transportation and built-up areas and several agricultural land uses to create a map with the following categories: agriculture, built-up, dry terrain, forest, recreation, transportation, water and wet terrain. We also used the national soil type map produced by Alterra [7], a research institute of Wageningen University. This map contains 10 different types of soil, which were also reduced to 8 classes: clay, built-up, gravel, loam, peat, sand, water. We also included a Physical Geography layer that contains the 10 main physiographic units: clay area, dunes, fenlands, Dutch hills, inlets, intertidal zone, river lands, sandy soils, North Sea and others.

3.2.3 Mining frequent patterns

The data mining workflow designed to identify frequent patterns in the tick bite reports consisted of three main steps: 1) generation of features from the environmental datasets and volunteered tick bite reports; 2) classification of the features with the Jenks Natural Breaks algorithm in order to discretize the input data and 3) extraction of the most frequent patterns using the AprioriClose algorithm.

3.2.3.1 Feature engineering

Feature engineering is a typical machine learning process to obtain new variables that incorporate the knowledge of a particular domain to create prediction and classification models. In our case, we obtained features from the original data sources mentioned in Sections 3.2.1 and 3.2.2. The aim of these features is to characterize each tick bite report using the environmental and volunteered datasets. In this work we are using 39 features (Table 3.2) classified in six different types: Temperature (T), Precipitation (P), Vegetation (V), Volunteer (R), Soil (S) and Distances (D). In addition to the type, each of the features can be continuous or discrete, depending if they are representing a variable aggregated in a particular time window or expressing a point

ID	Feature Name	Short Description	Type
1	Accspr	Acc. of spring mean temperature	T
2	Accsum	Acc. of summer mean temperature	T
3	Accaut	Acc. of autumn mean temperature	T
4	accwin	Acc. of winter mean temperature	T
5	Accd	Acc. of temperature from 1st Jan. until the tick bite date	T
6	d20	No. of days in a year with temperature over 20°C	T
7	d25	No. of days in a year with temperature over 25°C	T
8	d30	No. of days in a year with temperature over 30°C	T
9	precspr	Acc. of spring precipitation	P
10	precsum	Acc. of summer precipitation	P
11	precaut	Acc. of autumn precipitation	P
12	precwin	Acc. of winter precipitation	P
13	norainspr	No. of days in spring without precipitation	P
14	norainsum	No. of days in summer without precipitation	P
15	norainaut	No. of days in autumn without precipitation	P
16	norainwin	No. of days in winter without precipitation	P
17	ndvispr	Acc. of NDVI in spring	V
18	ndvisum	Acc. of NDVI in summer	V
19	ndviaut	Acc. of NDVI in autum	V
20	ndwiwin	Acc. of NDVI in winter	V
21	evispr	Acc. of EVI in spring	V
22	evisum	Acc. of EVI in summer	V
23	eviaut	Acc. of EVI in autumn	V
24	eviwin	Acc. of EVI in winter	V
25	ndwispr	Acc. of NDWI in spring	V
26	ndwisum	Acc. of NDWI in summer	V
27	ndwiaut	Acc. of NDWI in autumn	V
28	ndwiwin	Acc. of NDWI in winter	V
29	weekday	Day of the week when the tick bite occurred	R
30	Woy	Week of the year when the tick bite occurred	R
31	Env	Type of environment where the tick bite occurred	R
32	Actv	Type of activity carried out when the tick bite occurred	R
33	agegroup	Age group of the volunteer	R
34	landuse	Type of land use in the location of the tick bite	S
35	Soil	Type of soil in the location of the tick bite	S
36	Fisio	Physiographic unit in the location of the tick bite	S
37	dforests	Distance to the closest forest	D
38	drecreation	Distance to the closest recreation area	D
39	dbuiltup	Distance to the closest built-up area	D

Table 3.2: List of features used in this work. Column type stands for: T (Temperature), P (Precipitation), V (Vegetation), R (Volunteer's reports), S (Soil), D (Distance)

value, respectively. The following subsections detail the features used in this experiment.

Features derived from the volunteered tick bites reports Features directly derived from the volunteered reports (indices 29-33, Type R) and from the location of the tick bites (indices 37-39, Type D) are shown in Table 3.2. First we extracted the type of environment and activity the volunteer reported and we encoded both of them in eight different categories. The environment-related categories which the volunteer can report online, comprise the following fixed categories: dunes, forest, garden, heath lands, meadow, urban park, wetland and others. Note that this feature is subject to the user knowing where the tick bite occurred and about his/her perception/knowledge about natural spaces. The activities the volunteer can report online consist of the following categories: gardening, pic-nic, picking fruits, playing, walking, walking with the dog, others and not reported. The age of the volunteer was classified into three categories: children (0-17), adults (18-60) and elderly (>60). This process of aggregating data to a coarser level of detail, eases the process of finding patterns in data. We also extracted the day of the week when each tick bite occurred and we calculated the distance to the closest forest patch, recreational area and urban area with the aim of identifying risky areas.

Features derived from environmental data The weather and satellite-derived vegetation indices were aggregated to a seasonal scale (i.e. spring, summer, autumn and winter), with the intention of capturing intra annual clusters in tick bites reports. Using this temporal scale allowed us to obtain two different types of climatic and vegetation features: continuous accumulative features and discrete features. The accumulative features consist in obtaining the seasonal accumulation of the variable by adding the daily values for each tick bite location using the weather (temperature and precipitation) and vegetation (NDVI, EVI and NDWI) datasets. This accumulation captures the seasonal variability in each tick bite location and provides hints about if this place is "greener" or "warmer" than other places in the Netherlands. In this way, we obtained 20 accumulative features (indices 1-5, 9-12 and 17-28 in Table 3.2, Type T, P and V respectively). We also calculated the temperature accumulation between 1st of January until the date of the tick bite, with the aim of capturing the temporal structure associated to each tick bite report. For instance, tick bites occurring in spring have a smaller temperature accumulation than those from autumn.

Discrete features are counters of days where a particular condition was matched (indices 6-8, 13-16 in Table 3.2, Type T and P respectively). Human activities in nature might be the main driver for tick bites occurrence, therefore, being able to identify if good weather conditions are directly linked with tick bites, could provide solid hints on how to model this phenomenon in the future. For each tick bite location in the dataset, we used the daily

[7] http://www.geodata.alterra.nl/Grondsoorten.htm

weather data to count the number of days per season above three temperature thresholds (i.e. 20, 25 and 30 degrees Celsius) and we counted the number of days per season with no rain registered. Finally, we also extracted the type of land use, soil type and physiographic unit (indices 34-36, Type S in Table 3.2). These three features contain discrete data in the form of class label categories (e.g. sandy soil, agricultural lands or built-up area).

3.2.3.2 Classification of features

The feature engineering process (Section 3.2.3.1) resulted in a table with 39 features (Table 3.2) where each row is unique and, therefore, it is difficult to find frequent patterns. To overcome this problem, we used the Jenks Natural Breaks (Jenks, 1967) algorithm to convert all non-categorical features into discrete classes, because categorical features (e.g. land use, soil type, activity carried out by the volunteer) are already classified.

The Jenks Natural Breaks (JNB) algorithm finds optimal class thresholds by minimizing the intra cluster variance and maximizing the distance between classes in continuous features. The chosen Python implementation of this algorithm [8] needs an initial parameter to find the number of breaks in data. The optimal number of classes was found by iteratively testing all values between 2 and 10 and calculating the goodness of variance fit (GFV) of the resulting classification. The GFV measures how well a statistical linear model fits data by yielding a value between 0 (worst fit) and 1 (best fit). We selected as optimal the first classification that yielded a GFV equal or larger than 0.8. Results were visualized using a stacked bar graph (Figure 3.2) that shows the classes found for each of the features.

3.2.3.3 Frequent patterns in tick bite locations

Frequent environmental conditions associated to tick bites were mined using the closed form of Apriori algorithm. This version applies a closure operation to features, by limiting the search space to frequent closed item sets which is usually smaller than the frequent itemsets. This means less operations need to be done, improving the performance of the original algorithm. The Apriori algorithm (Agrawal and Srikant, 1994) is one of the most popular algorithms to discover frequent association rules in large datasets. A frequent association rule is defined as an ordered item set that appears in data in, at least, *minsup* percent (0%-100%) of the data. The notion of this minimum support, *minsup*, is an important concept in the Apriori algorithm. Support is the number of repetitions an item set has been seen in input data. For each tick bite observation (item set) appearing in data, a new node is created and propagated down the tree until a leaf or root position. Each node contains a "hit counter" with the number of repetitions of the item set that is updated as long as new tick bite observations are read. The minimum support defined by the user implies that only nodes containing a number of repetitions bigger than this threshold are kept for the patterns extraction.

[8] https://github.com/perrygeo/jenks

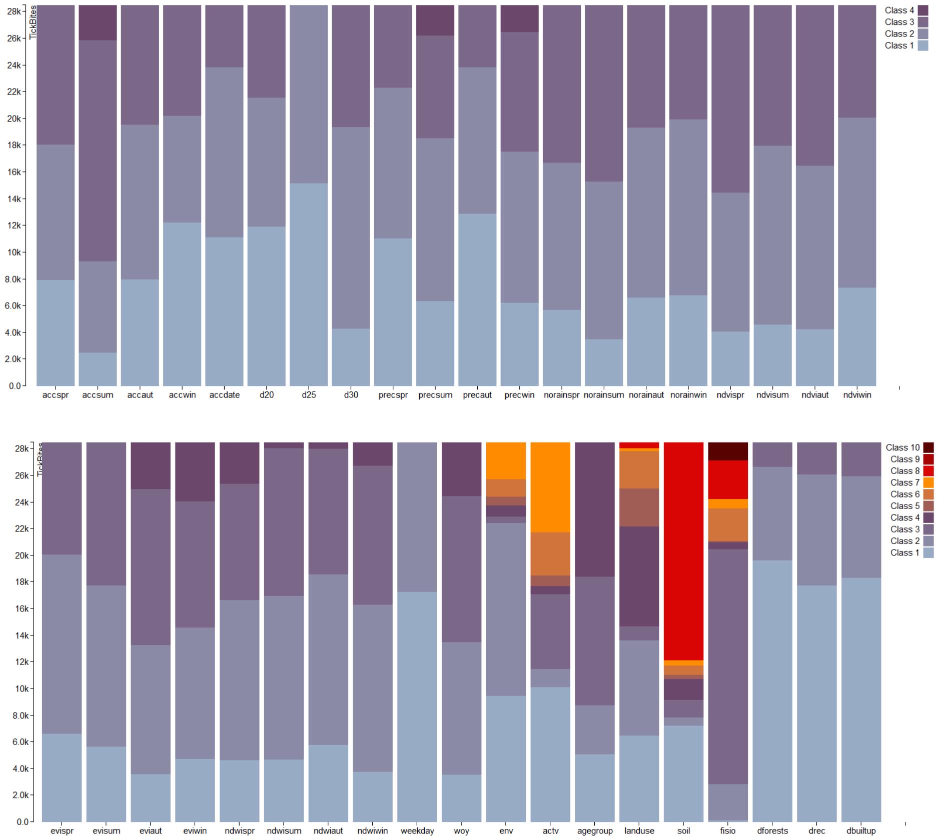

Figure 3.2: Number of tick bite reports per classified feature. For each feature, the numbers represent the upper and lower thresholds of the classes identified by the JNB algorithm. Temperature thresholds are expressed in degrees Celsius, precipitation in millimeters, vegetation indices in seasonal accumulations, discrete features are expressed in either categories or in number of days and distances are in meters. For a complete description of the features see Table 3.2.

Thus, if this parameter is set to low values, it means that the threshold to include a pattern is also low, and the patterns obtained are longer in number of items. Contrarily, if the parameter is set to high values, it means that the user is only interested in keeping patterns with a very high frequency, and this is prone to produce shorter patterns.

One of the major drawbacks of Apriori is its slow performance for moderately big input transactions. The number of candidates generated grows exponentially and thus its efficiency is low. Therefore, we have used in this work an enhanced version of the original algorithm that limits the search space to item sets that accomplish a closure property, improving its general performance. AprioriClose (Pasquier et al., 1999), works by only generating candidates that are not included in any superset having exactly the same support. Thus, the output of this algorithm guarantees that only the maximal superset values are generated, reducing the number of patterns obtained.

The chosen platform for the frequent pattern mining operation is SPMF [9] (Fournier-Viger et al., 2014), which is a Java-based open-source platform specialized in pattern finding. The closure operation in AprioriClose algorithm implies that an operation with two frequent patterns, yields a pattern that is also a frequent pattern. This rule reduces the search space and therefore, the number of operations performed by the algorithm, improving its performance compared to the original algorithm.

AprioriClose requires a matrix of integers, therefore, the discretized table with JNB was encoded accordingly for the processing. For this purpose, we encoded each observation in 39 items of 6 digits as follows: the first pair of digits was fixed to 99, to avoid filling with zeros the header of the number; the second pair of digits indicated which of the 39 features is in use and the last pair of digits indicated the class within the feature. In this work we used a minimum support of *minsup* 25%, because constraining the algorithm more by increasing minsup only produced patterns of length 1 or 2, which are not informative.

To study the effect of data imbalance, we performed the frequent pattern operation in both collections separately with two goals: 1) identify if both collections identify consistently the relevant features and 2) analyze if there is any transference of information from NK to TR that can set up time-persistent patterns. Both points are required to know if the data imbalance favorable to TR is actually driving the tick bites phenomenon. The same process of feature engineering described in Section 3.2.3.1 was applied to NK and TR separately to obtain the corresponding table of features. After this, we applied the same Jenks Natural Breaks classification used with the tick bites reports and mined the most frequent patterns using AprioriClose with the same input parameter (*minsup*=25%).To ease the visual exploration of patterns, we have used three types of graphical elements: 1) heat maps in Python to summarize the patterns produced in each experiment;

[9] http://www.philippe-fournier-viger.com/spmf/index.php

2) interactive ring maps in Javascript to give a general overview of the relevance of features in patterns; 3) maps to display two selected patterns in the geographical space. Heat maps is a compact representation of patterns, useful to explore the frequency of appearance of the features and its classes. Ring maps provide a quick overview of the type of the patterns produced by AprioriClose. Ring maps depict patterns in a series of concentric rings, where each ring represents a feature identified in the pattern. Each ring is divided in segments that will contain the type of a specific feature. We have assigned to each type of feature a different color: Red represents temperature features (Type T), blue is precipitation (Type P), green are the vegetation indices (Type V), pink are the volunteer features (Type R), yellow are distance features (Type D) and brown represents soil type and land use (Type S). The size of the segments is variable, depending on the number of patterns found using a particular feature. Thus, interpreting ring maps requires moving from the inner ring to the outer ring, only moving to a new segment if it is adjacent to the previous one, to understand the combinations of features found in data. Maps are helpful to explore the spatio-temporal distribution of the tick bites reports associated to a particular pattern. For this purpose, we have selected two time-persistent patterns (i.e. occurring every year in the time series) and projected them in the geographical space.

3.2.4 Frequent environmental patterns in the Netherlands

The patterns mined from the tick bites reports collection are patterns expected to be capturing environmental conditions associated to tick bites, however, they might also contain typical conditions found in the Netherlands. To be able to remove such patterns that are just typical or average Dutch conditions we mined patterns in a pseudo-random locations dataset and compared the patterns yielded by the two collections. The underlying idea is that patterns occurring in, both, the tick bite reports and the random locations are likely a reflection of the average situation in the Netherlands and thus, are not discriminating and should be removed from further analysis.

The pseudo-random locations were generated by choosing a random point within a 10 kilometer buffer around each tick bite location. For the calculation of the features as described in Section 3.2.3.1, this random point got assigned the date of the corresponding tick bite report. We thus generated the “validation” random dataset keeping a certain spatio-temporal structure resembling the tick bites distribution; this is stricter than generating random points in space and time, as these might fall on land types that clearly are non-tick habitats and/or in time intervals that fall outside of the typical activity period – e.g. hard freeze periods. It is important to highlight that volunteered-based features (Type R, Table 3.2) were not included in this validation experiment, due to the impossibility of knowing them for these particular locations.

The features obtained from the pseudo-random locations were first discretized using the JNB classification obtained for the tick bite reports and then we mined the most frequent patterns using AprioriClose with a minsup of 25%. Like for the tick bites, results are shown using heat maps and ring maps.

3.3 Results

3.3.1 Classification of features

The application of the JNB algorithm to the non-categorical features defined in Section 3.2.3.1 yielded both the optimal number of classes for each feature and their upper and lower break values. Using these break values we created a discretized version of the features. Figure 3.2 shows the number of tick bite reports per feature and per class. Features are presented in the same order shown in Table 3.2 (i.e. temperature, precipitation, vegetation, volunteer, soil and distances) and, for readability reasons, seasonal temperature accumulations were divided by 90 (number of standard days in a season) to calculate averages, as these are easier to interpret.

Figure 3.2 also shows JNB discretized most of the non-categorical features into 3 classes, although some of them were discretized into 2 (i.e. `d25`) or 4 classes (i.e. NDWI indices). Additionally, it shows that accumulative features tend to have a similar number of tick bites reports per class whereas discrete features have differing amounts of tick bites per class. For instance, temperature and precipitation accumulations in spring, autumn and winter contain a similar number of tick bites per class. Therefore, it is not possible to study the effect of a warm/cold or rainy/dry seasons over the number of tick bites reports using this kind of features. However, discrete features seem to be more discriminative. For instance, the features that count the number of days with a a threshold (i.e. `d20`, `d25`, `d30` in Table 3.2), show classes that contain a significant portion of the tick bites reports. Here we see the expected positive correlation between warm summers and the number of tick bites. Moreover, features that count the numbers of non-rainy days (i.e. `norainspr`, `norainsum`, `norainaut`, `norainwin` in Table 3.2) show that the drier the season, the more tick bites.

Vegetation features show that most of the tick bites occurred either close to the peak of greenness of deciduous and mixed forests or in evergreen forests, because the NDVI and EVI values are relatively high. The NDWI-based features indicate that most tick bites occur in medium-to-high moisture levels, which is in line with the fact that ticks need moisture conditions to survive (Barrios et al., 2012). In addition, the fact that people tend to do recreational activities in nature during the summer break (which is in an advanced greenness state), could also explain that the majority of tick bites reports are produced in similar vegetation and moisture conditions.

The feature landuse shows that most of the reports come from forests, built-up areas (including gardens and parks) and agricultural lands. Landuse also shows that tick bites are distributed evenly across these three classes, meaning that not only forests, the natural habitat for ticks, pose a risk for humans. The features soil and physio show that most of the tick bites occur in sandy soils. Distance features, dforests, drecreational and dbuiltup, show that the majority of tick bites occur quite close to these types of land uses. Concretely, most volunteers report their tick bites from locations that are at a maximum distance of 450 m of a forest, 525 m of a recreational area or 575 m of a built-up area. In fact, the number of tick bites produced in built-up areas and the information derived from the distance-based features, suggest that most of the tick bites occur in peri-urban residential areas, where different types of landscape are mixed.

Regarding the volunteer features, the week of the year in which the tick bite was reported (woy) shows that most tick bites occur between weeks 19 and 26 (i.e. May to mid-June) and weeks 26 to 34 (i.e. mid-June to mid-August). Remarkably, the number of tick bite reports for these two periods is similar. The feature dayofweek shows that most of the observations are reported on Saturday and Sunday, when people have more time for outdoor leisure activities, as confirmed by the features actv and env, which show that most volunteers were gardening, playing or walking in forests. The feature agegroup shows that most of the reports refer to children (below 18 years of age) and elderly people (older than 60 years of age). These results might be used by decision-makers to target further Lyme-disease prevention campaigns.

3.3.2 Mining frequent patterns

This section presents the results of applying the AprioriClose algorithm to the features associated to the complete tick bites collection (NK+TR). After that, and to evaluate and validate these patterns, we present the results of separately mining the patterns associated to the NK and TR datasets to study the impact of data imbalance and mining the patterns associated to random locations to determine common environmental conditions in the study area.

3.3.2.1 Frequent patterns associated to all the tick bites reports

The application of AprioriClose to the JNB-classified features associated to tick bites locations resulted in a total of 429 patterns with lengths ranging from one to four items. Notice that patterns with one and two items are deemed uninteresting. The list of patterns yielded show that there is an apparent tie between continuous (i.e. accumulations and distances) and discrete (i.e. counters and categories) features. Out of the 62 patterns with three and four items, 31 contain two or three continuous features and 31 contain two or three discrete features. However, out of the 31 with continuous features, only four contain temperature accumulations and the rest mainly contain distance features. Regarding the discrete features, the combinations of the number of warm days with non-rainy days appear 8 times in patterns. This means that spring and summer seasons with few rain and warm weather, which are more favorable conditions for outdoor activities, are related to a higher number of tick bites regardless of temperature accumulations. In this case, distance features also appear in a high number of patterns, a behavior in line with the big number of tick bites produced at a maximum distance of 500 meters (Figure 3.2) away of a forest, recreational or built-up area. Figure 3.3(a) shows a heat map summarizing the 62 patterns produced with all tick bites locations. Horizontal axis represents the 39 features used in this work (Table 3.2) and the vertical axis the classes that JNB identified (Figure 3.2). The grayscale assigns a frequency to the features and classes found frequent after mining patterns with AprioriClose. Note that feature soil has been moved to a different position to make the chart more compact.

#	Feature	Class	Feature	Class	Feature	Class	Frequency
1	**accwin**	0C – 3C	**d25**	14 – 41 days	**d30**	2 – 6 days	**9669**
2	**accwin**	0C – 3C	**d25**	14 – 41 days	**norainsum**	49 – 69 days	**8357**
3	**accwin**	0C – 3C	**d25**	14 – 41 days	**dforests**	0 – 463 meters	**8026**
4	**accwin**	5C – 8C	**d20**	37 – 55 days	**d25**	0 – 14 days	**8277**
5	**d30**	2 – 6 days	**drec**	0 – 525 meters	**dbuiltup**	0 – 571 meters	**7332**
6	**weekday**	Weekends	**soil**	Sandy soil	**dforest**	0 – 463 meters	**8755**
7	**soil**	Sandy soil	**fisio**	Sandy soil	**dforests**	0 – 463 meters	**11823**
8	**fisio**	Sandy soil	**drec**	0 – 525 meters	**dbuiltup**	0 – 571 meters	**7407**
9	**dforests**	0 – 463 meters	**drec**	0 – 525 meters	**dbuiltup**	0 – 571 meters	**8161**

Table 3.4: Patterns of length 3 found in NK, TR and NK+TR. For a complete description of the features see Table 1. The frequency refers to the number of times each pattern appears in NK+TR.

Figure 3.4 provides a general overview of the patterns found by AprioriClose using ring maps. There is a clear difference between both images: while patterns with three items present more heterogeneity in the features they contain, patterns with four items are clearly dominated by accumulative temperature features. It is important to highlight that Figure 3.4 show that distance, temperature and soil type features are present in most of the combinations and precipitation features appear a reasonable number of times. Figure 3.4 (left) shows two patterns marked with a white dot (patterns 5 and 6, Table 3.4). Each feature in the pattern corresponds to one of the segments starting from the inner ring and moving to the outer ring.

3.3.2.2 Frequent patterns associated to NK and TR reports

The application of AprioriClose to the features associated to the NK reports produced 453 patterns, ranging from one to four items. The same operation applied to TR features resulted in 762 patterns, ranging from one to seven items. The difference in the length of the patterns stems from the imbalance in the number of volunteers sampling tick bites in NK and TR collections: TR collection has near four times more tick bites reports than NK, so the variability it contains is higher and this subsequently turns into longer patterns. Figure 3.3(a-c) graphically points out the data imbalance: the resemblance between NK+TR (Figure 3.3a) and TR (Figure 3.3b) heat maps is remarkable, indicating that TR collection is strongly influencing the pattern mining for NK+TR experiment described in Section 3.3.2.1.

Following this line, we have depicted the patterns extracted from NK and TR in figures 3.5 and 3.6, respectively. Figure 3.5 shows that the most relevant features in NK collection contain temperature, precipitation and distance features. Figure 3.6 reveals the importance of temperature and distance features in TR and, to a less extent, precipitation. As seen, both data collections tend to converge in the identification of relevant features to monitor tick bites phenomenon. However, if we compare these two figures with Figure 3.4, we can see that the ring maps produced for the patterns of all tick bites is losing the relevance of precipitation features and increasing the relevance of soil features. This means that data imbalance has a certain effect on the patterns produced, because it is still identifying relevant conditions, but there is some information loss on the process of mining frequent patterns from NK+TR collections.

Persistent patterns found in NK and TR collections are shown in Table 3.4. These 9 patterns have occurred along all years in the time series and this can provide hints about the main drivers of tick bites phenomenon: the proximity to built-up area, forest and/or recreational area, and temperature thresholds have a strong presence, contrarily to temperature accumulations. These findings are in line with results described in Section 3.3.1 and reveal recurrent reported conditions that might be used to identify tick bite risk conditions for humans.

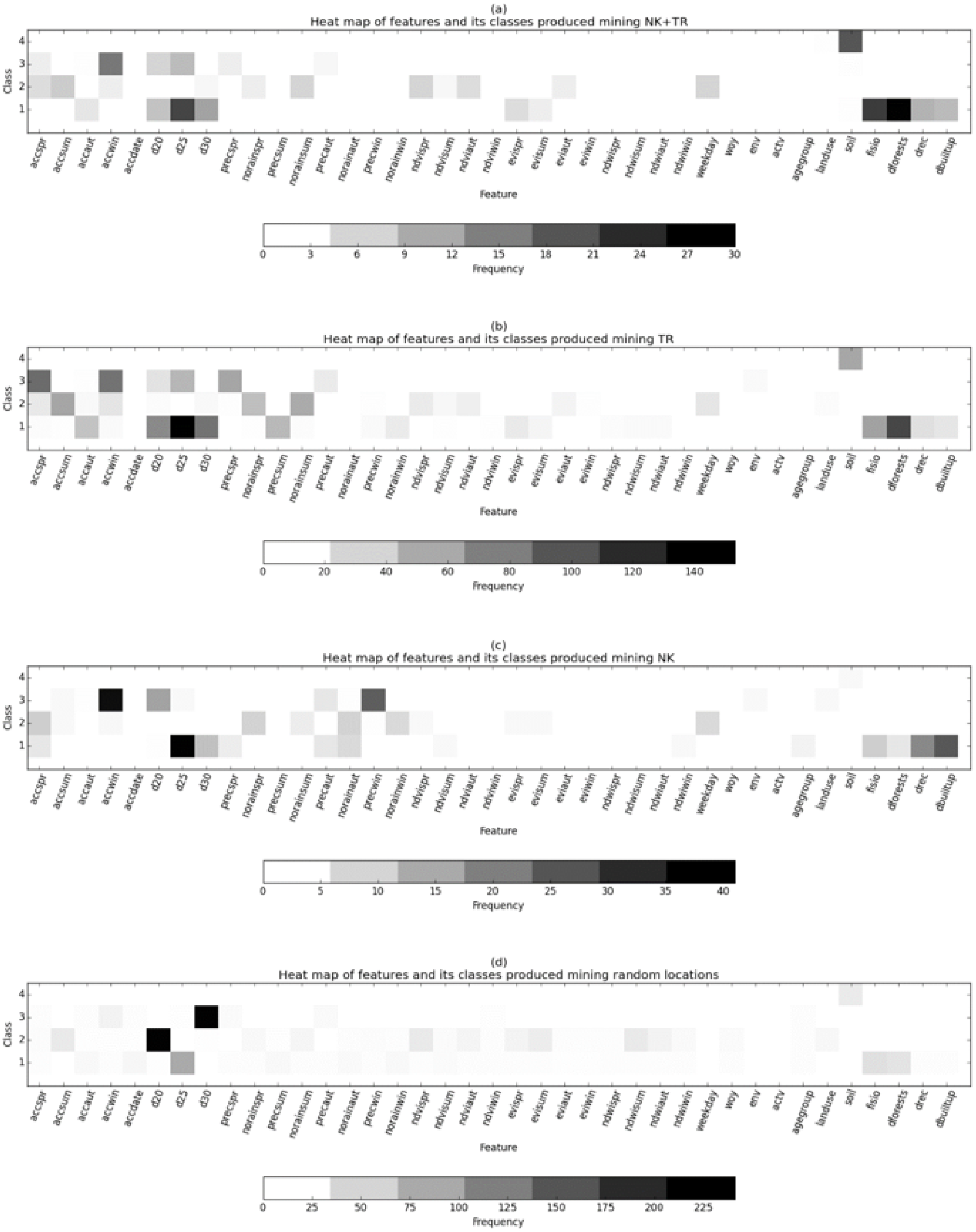

Figure 3.3: Heat maps showing the frequency of features and classes for NK +TR (a), TR (b), NK (c) and random locations (d) for patterns with three and four items. The horizontal axis lists all the features created for this work (Table 3.2) and the vertical axis indicates the class that has been identified in the frequent patterns (Figure 3.2)

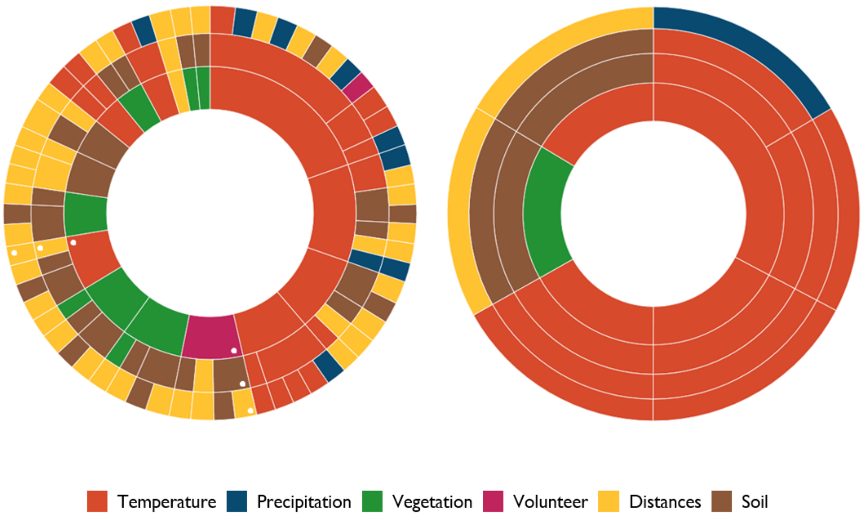

Figure 3.4: Ring maps for the patterns of length 3 (left) and 4 (right) that have a minimum support of 25% in the complete tick bites collection (NK+TR). Each ring represents a feature identified in each pattern, the width of the segment the number of times a feature has appeared in the patterns and segments marked with a white dot illustrate how to interpret ring maps, by moving from the inner to the outer rings.

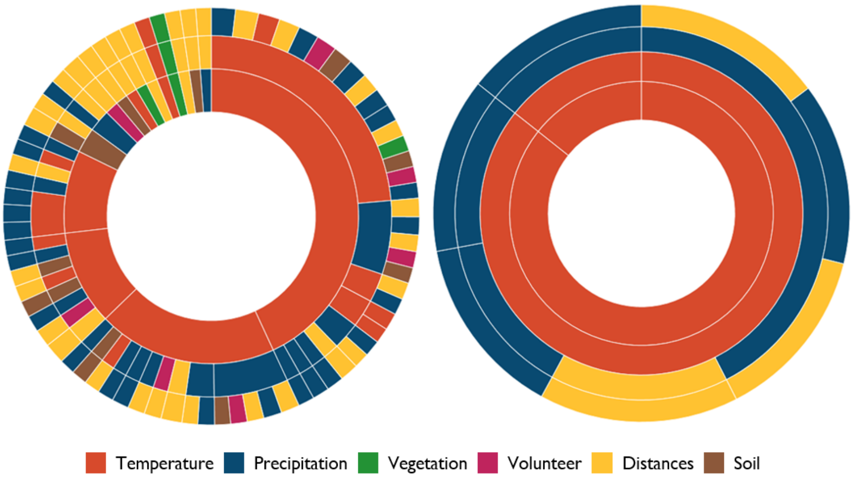

Figure 3.5: Ring maps for the patterns of length 3 (left) and 4 (right) that have a minimum support of 25% in the NK collection.

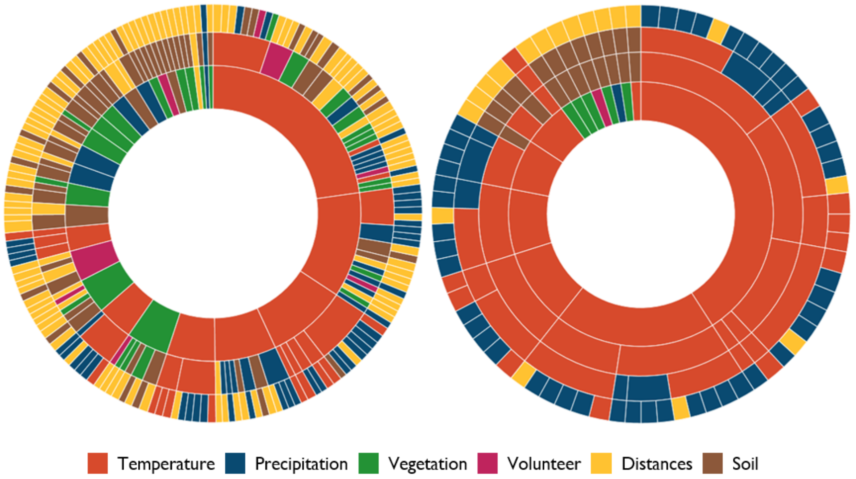

Figure 3.6: Ring maps for the patterns of length 3 (left) and 4 (right) that have a minimum support of 25% in the TR collection

3.3.2.3 Frequent patterns associated to random locations

The application of AprioriClose to the features associated to the pseudo-random locations produced 744 patterns ranging from one to five features. The comparison of these patterns with the patterns associated to tick bites yielded zero coincidences. This means that the frequent patterns associated to the tick bite reports truly represent conditions linked to tick bites and not the common environmental conditions in our study area. Figure 3.3(d) shows the summary of the patterns obtained for the random locations. As seen, the difference between this plot and the three above is substantial in the features identified as frequent and also in its frequency, explaining the zero coincidences between this data collection and tick bites collection. Figure 3.7 shows the ring maps for patterns with three and four items for the random dataset. As seen, temperature features are clearly dominant in both maps, however, an interesting characteristic in both ring maps is the relatively high amount of patterns yielded using vegetation indices, in contrast with the patterns depicted in Figures 3.4, 3.5 and 3.6. This means the common patterns in the Netherlands are associated to the "greenness" levels of the country.

3.3.3 Spatio-temporal visualization

Figures 3.8 and 3.9 illustrate the spatio-temporal distribution of tick bites associated to two of the persistent patterns (i.e. frequent patterns found in the NK+TR, NK and TR). We chose these two patterns because they contain several of the most frequent features identified by AprioriClose. More precisely, these figures illustrate the patterns: `d30`, `drec`, `dbuiltup` (Pattern 5, Table 3.4) and `weekday`, `soil`, `dforests` (Pattern 6, **Table 3.4**).

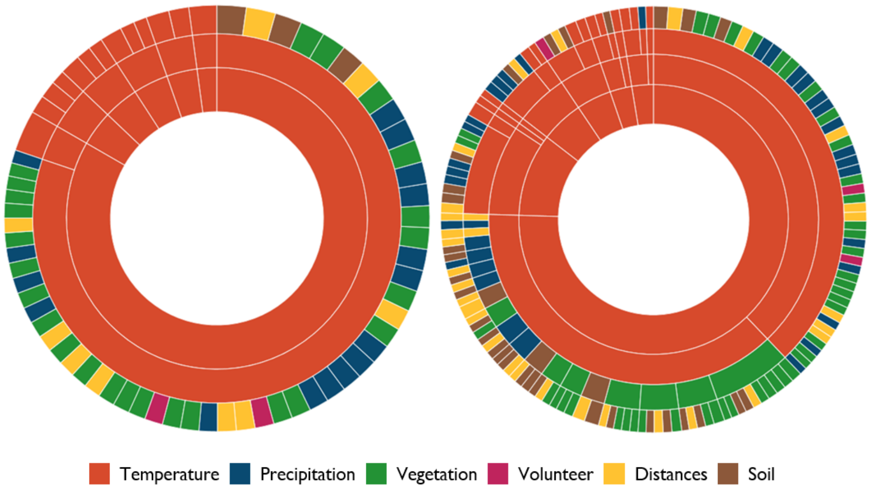

Figure 3.7: Ring maps for the patterns of length 3 (left) and 4 (right) that have a minimum support of 25% in the pseudo-random locations collection

At the level of classes in the feature, the first pattern shows tick bites that occurred at a maximum distance of 525 m of a recreational area, 571 m of a built-up area and in a year with 14 - 41 days with a temperature above 25 degrees Celsius, and the second pattern shows tick bites that occurred during weekends, in sandy soil and at a maximum distance of 463 m of a forest.

The inclusion of distance features makes the tick bites reports to be clustered around any of the selected land uses (i.e. forest, built-up, recreational area). For each year, this clustering effect reveals locations with a high recreational pressure in the conditions specified by the patterns. Thus, popular places for recreation (i.e. areas along the coast and forested areas in the center of the country) consistently tend to accumulate observations, due to the higher human recreational pressure. The imbalance in the number of tick bite reports is appreciated in both figures where we can see that there are more tick bites from 2012 (the year when TR was launched) onwards. However, it is important to note that 2014, a year particularly rich in tick bite reports, does not contribute with many tick bites to Pattern 5 (**Table 3.4**) represented in Figure 3.9. In fact, we observe that 2006 contains more tick bites reports than 2014. This may indicate that other features (e.g. number of non-rainy days) were the major drivers of tick bites for 2014.

Table 3.6 shows the percentage of tick bites reports associated to each pattern per year and the total percentage of tick bites reports per year represented by the combination of these two patterns. As seen, Pattern 6 consistently explains around a 30% of the tick bites reports per year. This is produced because this pattern includes a soil feature which is something temporally

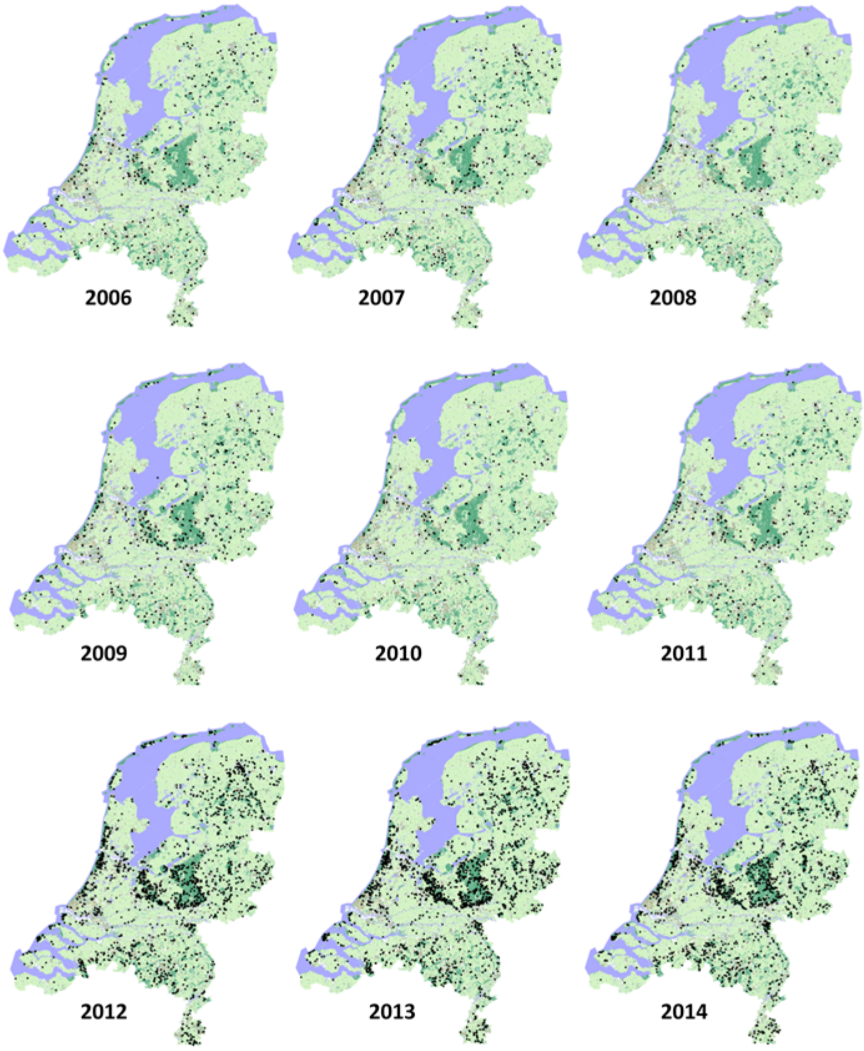

Figure 3.8: Spatio-temporal projection of pattern: weekday, soil, dforests per year. This pattern shows tick bites that occurred at a maximum distance of 525 m of a recreational area, 571 m of a built-up area and in a year with 14 - 41 days with a temperature above 25 degrees Celsius

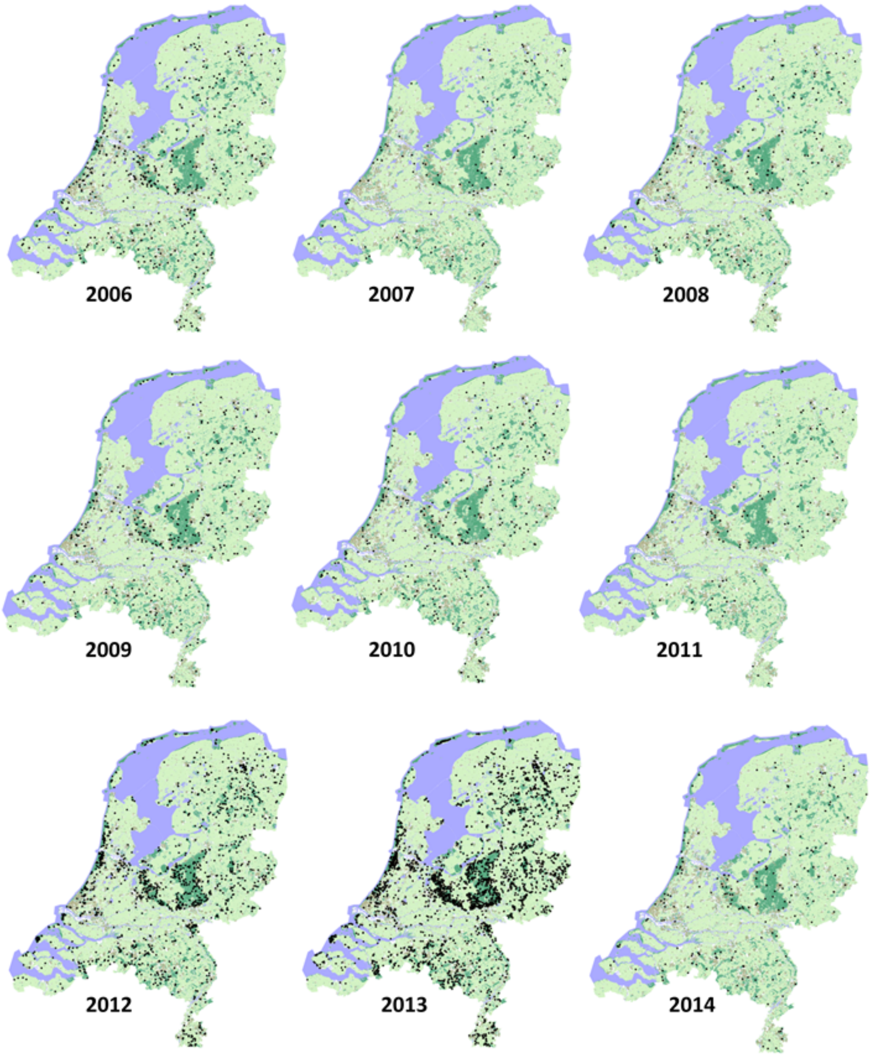

Figure 3.9: Spatio-temporal projection of pattern: d30, drec, dbuiltup per year.This pattern shows tick bites that occurred during weekends, in sandy soil and at a maximum distance of 463 m of a forest.

Year	Pattern 5	Pattern 6	Total
2006	52%	32%	53%
2007	9%	29%	33%
2008	26%	28%	39%
2009	21%	32%	40%
2010	37%	30%	45%
2011	10%	34%	38%
2012	23%	31%	40%
2013	44%	29%	46%
2014	3%	30%	32%

Table 3.6: Percentage of tick bites that occur under the environmental and/or human factors depicted by the selected persistent patterns (patterns 5 and 6, Table 3.4). Pattern 5 shows tick bites occurring at a maximum distance of 525 m of a recreational area, 571 m of a build-up area and in a year with 14 – 41 days contains with a temperature above 25 degrees Celsius, and Pattern 6 shows tick bites that occurred during weekends, in sandy soil and at a maximum distance of 463 m of a forest The total percentage was calculated using the number of unique tick bites in the selected patterns.

constant in a country. Pattern 5 shows a more variable representativity of the tick bites per year, due to the temperature feature included: years with a lower percentage of tick bites reports (e.g. 2007, 2011 and 2014) may be linked to a year that was colder or there were wetter conditions that prevented people to go to nature.

3.4 Discussion

Citizen science projects allow the collective monitoring of a wide range of (environmental) phenomena at detailed spatio-temporal scales. Therefore, these projects have the potential to accelerate scientific discovery. However, it remains challenging for researchers to work with volunteered data because of its heterogeneous character and its ever-questioned quality. In our case, we had to deal with an imbalanced inter-annual distribution of tick bites reports as the NK dataset contains appreciably less reports than the TR dataset. This means that TR reports may be over represented in the results. A possible way to mitigate this problem is to sample the TR reports so as to have a similar number of reports per year. However, this leads to data (information) loss and we took the decision of analyzing the complete tick bite collection. The impact of data imbalance was studied by mining frequent patterns in the NK and TR datasets separately. This analysis showed that there are NK and TR specific patterns that do not appear when mining the complete tick bite collection. However, a similar set of relevant features and classes were mined from both datasets. In fact, this analysis also showed that there are persistent patterns in both collections. Tekenradar.nl is an ongoing project that collects about 8,000 tick bites reports each year. In this amount remains stable, data imbalance will not play a major role in future analysis.

Contrarily to tick modelling studies, this study revealed that days above a certain temperature threshold or precipitation threshold and the proximity to a forest, recreational or build up areas, are more relevant to model tick bites than the accumulation of environmental (e.g. weather) features. We believe that this is because these features are closely related to human behavior (e.g. outdoor activities). Our results also show that there is not a strong link between seasonally accumulated environmental factors and tick bites. This suggests that shorter time frames (e.g. weeks, months) could be tested as environmental features are known to explain tick distributions and activity (Estrada-Peña et al., 2012; Ostfeld et al., 2006). In this line, the incorporation of landscape fragmentation indices in further studies could prove useful since previous research has found a link between landscape, tick densities and Lyme disease incidence (Lambin et al., 2010; Li et al., 2012). In the current study, around 65% of the tick bites reports are produced in locations that are less than 575 m of a forest, built-up area or recreational area and 25% of the tick bites are produced in agricultural lands. The fact that a few land use types concentrate most of the tick bites, and the remarkably high number of patterns with features on proximity to forest, built-up or recreational area, suggests that landscape fragmentation could indeed be useful to explain tick bite occurrences.

The analytical engine of this study is the AprioriClose algorithm, which is an unsupervised method to explore large and complex datasets. This unsupervised character makes it hard to perform a traditional validation of the results. Therefore, to determine the validity of these patterns we generated a random locations dataset in a 10 km buffer around each tick bite report so as to keep the spatio-temporal distribution of the original reports. This method can be seen as a strict validation of the patterns because it prevents the random points from falling far from both the tick habitat and the tick season. This allows confirming whether the mined patterns are indeed strongly associated to tick bites, conditional on the area under investigation being suitable for ticks and without taking into account the volunteered-based features. We found that most of the patterns mined in random location stem from vegetation and temperature features and that there was no overlap between the patterns produced when mining tick bite locations and random locations. This means the collective "sampling" effort carried out by the thousands of volunteers that support the NK and TR initiatives, captures well the occurrence of tick bites in space and time. In turn, this means that the tick bite dataset used in this study has a high scientific value.

From a geographic perspective, mapping two of the persistent patterns revealed that the corresponding tick bites are scattered across the country and tend to be clustered in forested areas. The minimum support threshold applied to the AprioriClose algorithm makes sure that each pattern represents at least 25% of the tick bites. Therefore, the two chosen patterns could represent up to 50% of all the tick bites. With the six features identified with the selected patterns we are capable of explaining an average of 40% of the tick bites phenomenon per year. The reason not to reach this 50% is due to the overlapping between the tick bites reports associated to each pattern. Adding a new persistent pattern to the initial selection may increase the representativity of the tick bites phenomenon.

3.5 Conclusion

Tick bites are the result of two interrelated geographic phenomena: 1) tick abundance and activity, and, 2) human presence. The volunteered tick bite datasets used in this work report the intersection of these two phenomena in a particular location and time. Here we used frequent pattern mining to characterize the environmental and human factors associated to "where" and "when" tick bites occur. In particular, the individual and combined analysis of the NK and TR datasets revealed the following main findings: 1) there are temporally persistent patterns in the tick bite datasets; 2) there is a positive correlation between the number of warm days or non-rainy days and the number of tick bites; 3) proximity to forest, recreational or built-up area appears a remarkably high number of times in the patterns yielded and 4) in tick-suitable areas, vegetation features do not seem to be sufficient to explain the tick bites phenomenon. Thus, we conclude that human-related factors are more important to model tick bites than using environmental variables that characterize the environmental conditions of a given location for a particular time frame.

Regarding the methods, to the best of our knowledge, this is the first study that applies frequent pattern mining method to a volunteered dataset that pertains to the public health domain. This work proved that frequent pattern mining is an efficient form of extracting information from crowdsourced geo-information. This is important as grass-root campaigns involving citizens (may) overcome the limitations of traditional statistical surveys and observational sensor networks by providing a high amount of data that allows the study of geographical phenomena at unique spatio-temporal resolutions.

In the public health domain, the identification of environmental and human factors associated to tick bites does not only open the way for modelling and mapping tick bite risk but it also demonstrates the scientific value of volunteered geographical information for public health applications. Further work will focus on the development of an improved forecasting model for tick bite risk. This new model could be linked to the current Tekenradar website to provide specific spatio-temporal information that, hopefully, contributes to the reduction of the number of tick bites and of Lyme borreliosis in the Netherlands.

Using volunteered observations to map human exposure to ticks

4

4.1 Introduction

Forests are complex dynamic systems that provide a wide array of ecosystem services to society, such as groundwater protection, and wood and fiber production (Hengeveld et al., 2017). Also, forests provide recreational services (e.g. sports, leisure activities), which may have positive (e.g. stress reduction) and negative (e.g. increased exposure to pathogens) impacts on human health (Sandifer et al., 2015). The transmission of pathogens causing tick-borne diseases to human hosts poses an important threat to public health (Ehrmann et al., 2017). In this regard, the European Centre for Disease Prevention and Control (ECDC) surveys five different tick-borne diseases (i.e. Lyme borreliosis, tick-borne encephalitis, relapsing fever, Crimean-Congo hemorrhagic fever, and Mediterranean spotted fever, from ECDC, last accessed July 5th, 2018), and reports the prevalence of these diseases across the European continent.

The incidence of Lyme borreliosis (LB), the most prevalent tick-borne disease in Europe, has steadily increased during the period 1990-2010 in, at least, nine European countries (Medlock et al., 2013). In recent years, however, sub European sentinel networks of general practitioners have identified the first signs of stabilization (Altpeter et al., 2013; Bleyenheuft et al., 2015; Vandenesch et al., 2014). In the Netherlands alone, the number of LB cases has continuously increased since the mid-1990s (Hofhuis et al., 2015b; Hofhuis A et al., 2006), but a recent study shows that this trend is also stabilizing (Hofhuis et al., 2016). Yet, over 20,000 citizens are diagnosed each year with LB, a situation that prompted researchers from Wageningen University and the Dutch Institute for Public Health and the Environment (RIVM), to start two crowdsourced projects to collect data on tick bites. Since 2006, the platforms Natuurkalender (NK; “nature’s calendar”; www.natuurkalender.nl) and Tekenradar (TR; “tick radar”; www.tekenradar.nl), have collected nearly 50,000 geo-located tick bite reports. To the best of our knowledge, these

This chapter is based on: (Garcia-Marti et al., 2018)
Available at: https://doi.org/10.1038/s41598-018-33900-2

platforms constitute the first citizen science projects that specifically focus on ticks and tick-borne diseases. Citizen science projects present an interesting characteristic when compared to classical forms of data acquisition (e.g. ground surveys or sensor networks): the ubiquity of the crowd allows a fine-grained sampling in space and time, which makes it possible to monitor elusive public health threats, such as tick bites.

The volunteered tick bite reports collected by the NK and TR projects depict the risk of getting a tick bite. This risk is the result of the interaction of two components: hazard (e.g. tick activity) and human exposure (e.g. recreational intensity in a location) (Braks et al., 2016), which tends to occur outdoors (e.g. forests, urban parks). There are significant efforts in literature to quantify the hazard component (De Keukeleire et al., 2015; Garcia-Martí et al., 2017; Swart et al., 2014), but finding proxies of exposure is a harder task, due to the unavailability of human recreational metrics at the national scale. Quantifying recreational pressure in nature is of interest in fields as diverse as public health, forestry management or environmental science. In public health, knowing the intensity of recreational activities might help delimiting locations that serve as an interface between natural elements and humans (De Keukeleire et al., 2015; Hall et al., 2017; Hansford et al., 2017; Mulder et al., 2013). This can be useful to better design tick bite prevention campaigns in forested areas. However, the complexity of forest ownership in the Netherlands might pose hurdles to this task. The sixth Dutch National Forest Inventory (Schelhaas et al., 2014) indicates that forests occupy 373,480 hectares (roughly, 11% of the territory). Public organizations (e.g. Staatbosbeheer: www.staatsbosbeheer.nl, municipalities) own 260,900 hectares of forest, whereas private partners (e.g. Natuurmonumenten: www.natuurmonumenten.nl, citizens) own 112,600 hectares of forest. Dutch law requires private owners with a property larger than 5 hectares to be registered at the Dutch Industrial Board for Forest and Nature (i.e. Bosschap www.bosschap.nl). Currently, 1,431 citizens own 61,000 hectares of forest larger than 5 hectares, and an unknown number of citizens own 51,600 hectares of smaller forest patches (Hoogstra-Klein, 2016; Schelhaas et al., 2014). Thus, educating a relatively small group of public and private foresters to reduce tick bites in their domain, might lead to a substantial decrease in the number of LB cases per year. For instance, forest owners with parcels presenting a medium or high human exposure, could implement preventive measures of tick habitat manipulation, such as grass mowing, removing leaf litter, or covering heavily visited locations with dry substrates (e.g. wood chips, gravel) (Hubálek et al., 2006).

In this work, we present a novel method to quantify country-wide exposure to ticks in forested areas. Our method is based on volunteered tick bite reports, and on a hazard model developed in our previous work (Garcia-Martí et al., 2017). This hazard model was also based on volunteered data about tick activity in forested areas. The resulting human exposure is presented as a categorical map that shows the recreational intensity across Dutch forested areas. Such a human exposure map should facilitate the cooperation between public health specialists and forest managers to jointly tackle public awareness about tick bite prevention, especially in locations with a high intensity of human exposure. In addition, this first-of-its-kind exposure map helps to understand and locate sources of potential landscape disturbance (by visitors) and could support better ecological management practices.

4.2 Modelling human exposure to tick bites

Obtaining measures of human exposure to ticks is a challenging task due to the unavailability of nation-wide datasets representing human activities outdoors. However, human exposure is tightly related to risk and hazard, so it can be calculated from these two variables. In this section we describe the theoretical background, operationalisation and the data processing workflow that allows the calculation of human exposure and graphical mapping of this variable.

4.2.1 Theoretical background

In the field of risk assessment, risk (R) is often modelled as a function of hazard (H) and exposure (E). The relationship between the three variables can be conceptualized as $R = H \times E$ (Braks et al., 2016). In the paragraphs below, we discuss in more detail how we can conceptualize R, H, and E in terms of probabilities, and how those probabilities connect to our data. The exposure below pertains to a single location (i.e. grid cell in the rasterized map). In Section 4.2.2 we operationalize the theory by linking it to the NK and TR studies.

We can define a normalized measure of exposure E as a probability: $E = P(visit)$. Following the same logic, we can define a normalized measure for H as the conditional probability of a tick bite given a visit: $H = P(bite|visit)$.Then, we can define R as the unconditional probability of getting a tick bite: $R = P(bite)$. From the law of total probability, and the obvious fact that a tick bit requires a visit, we have: $P(bite) = P(bite|visit)P(visit) + P(bite|novisit)P(novisit) = P(bite|visit)P(visit)$, from which we recover $R = H \times E$.

4.2.2 Operationalisation

Let v be a variable representing the unknown total number of visits to each grid cell, and let n be the total number of records in the NK and TR studies, then an estimate of $E = P(visit)$ is v/n. The probability $H = P(bite|visit)$ can be estimated as the total number of reported tick bites b at the location, divided by the number of visits b/v. Note that H may be zero when $b = 0$, and is undefined when $v = 0$. The unconditional probability: $R = P(bite)$ can be estimated by dividing the total number of tick bites by the number of records in the NK and TR studies b/n.

Using these estimates for the probabilities, we find that v can be calculated when the number of person-days in the study (i.e. the number of tick bite reports), and a measure of the hazard of the active ticks are known:

$$E = \frac{R}{H} = \frac{b}{nH} \tag{4.1}$$

Hence, it is possible to obtain a measure of human exposure to ticks by dividing the number of tick bites in a location by a measure of the hazard for that location. For a proxy of H, we use a previously developed hazard model (Garcia-Martí et al., 2017). Note that the value of n is immaterial as it is a constant over the Netherlands, and the output of the hazard model is not H but assumed to be proportional to H. The above is conditional on H being not equal to zero, which is satisfied since a property of our hazard model is that it always yields strictly positive hazard.

4.2.3 Data processing

The calculation of human exposure to ticks req uires calculating risk and hazard. The risk of getting a tick bite can be estimated from the volunteered tick bite reports, and hazard can be derived from a tick activity model developed in our previous work (Garcia-Martí et al., 2017). The workflow designed to obtain human exposure to ticks consists of four steps: 1) mapping the risk of getting a tick bite; 2) estimating tick hazard; 3) calculating and mapping human exposure; 4) validating the results. Note that this work has been developed using different Python libraries: numpy (Oliphant, 2006) has been used to handle the different arrays, jenks (Perry, 2014) package implements Jenks Natural Breaks (JNB) algorithm (Jenks, 1967) to classify the exposure layer, GDAL (GDAL Development Team, 2018) and cartopy (Met Office UK, 2010) were used to process geospatial data and prepare the maps, and matplotlib (Hunter, 2007), and seaborn (Waskom et al., 2017) to prepare the plots.

The transformation of tick bite reports into a risk layer requires selecting a spatial aggregation unit. Here we choose a regular grid with cells of 1km^2 as spatial unit because the existing hazard model works for grid cells with this spatial resolution. Thus, the tick bite reports were aggregated to grid cells of 1km^2. Risk is defined as the cumulative sum of tick bite reports over the period 2006-2016 occurring in each grid cell (Figure 4.1A), if the cumulative sum is greater than zero. The number of tick bites per cell ranges between 1 (in blue) and 353 (in red). The visual inspection of Figure 4.1A shows that coastal (e.g. from Haarlem to Middelburg), and forested areas (e.g. Veluwe national park, Utrechtse Heuvelrug) in the center of the country present a high concentration of high risk locations. These are well-known locations for outdoor recreation. Smaller regions of high tick bite risk can be found in the rest of the country (e.g. provinces of Drenthe and Groningen).

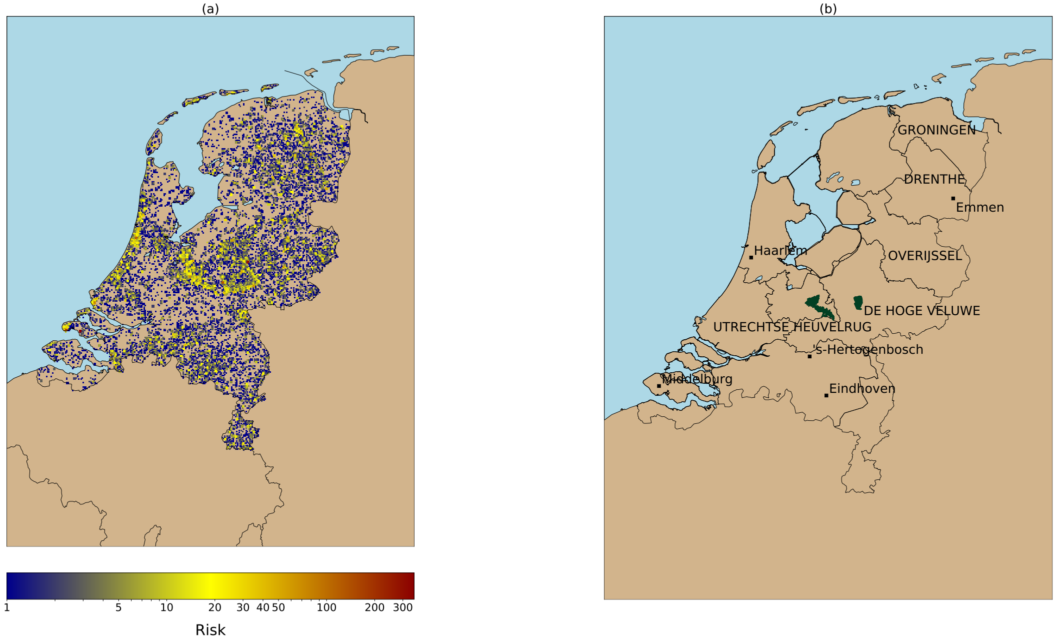

Figure 4.1: Risk of tick bites (2006-2016) as collected by the NK and TR volunteered projects. The cumulative sum of tick bites reports per 1km grid cell are used as a proxy of tick bite risk. The image reveals that tick bites are produced throughout the country. However, the reports tend to be clustered around forests (e.g. Veluwe national park, center of the country), or recreational areas (e.g. coastal areas). B: locations mentioned along the text. The provinces and the national parks are labeled with capital letters, whereas cities are labeled in lower case.

Case	Risk	Hazard	Exposure	Interpretation	Representation
1	R > 0	H > 0	E > 0	Std. case	Fig. 4.3, class. E
2	R > 0	H = undef.	E = undef.	TB in non-forested loc.	Fig. 4.6, light green
3	R = 0	H > 0	E = 0	Forest low recr. or H	Fig. 4.6, yellow
4	R = 0	H = undef.	E = undef.	No data	Fig. 4.6, grey

Table 4.2: The four possible cases that can occur when dividing risk by hazard.

Hazard is defined as the tick activity provided by a data-driven model(Garcia-Martí et al., 2017) that predicts daily tick activity in vegetated areas suitable for ticks (i.e. forests and natural grasslands). This model was built using nine years (2006-2014) of volunteered tick activity data (acquired in the Netherlands by using cloth dragging) and a large suite of environmental variables. Volunteers sampled 15 vegetated locations on a monthly basis, and counted the number of ticks per life stage (i.e. larvae, nymph, and adult). Our model, which was only calibrated for nymphs as they pose the highest hazard to humans, uses 101 biotic and abiotic environmental predictors. These predictors include data about the habitat conditions for ticks (e.g. litter, moss), mast years for three tree species, weather (e.g. temperature, evapotranspiration, relative humidity), satellite-derived vegetation indices (e.g. NDVI), and land cover. To account for the effect that short- and long-term weather conditions have on tick activity, we aggregated the weather data to 11 temporal resolutions (i.e. 1-7, 14, 30, 90, 365 days). The model yields daily tick activity for 4,132km^2 for forests and grasslands (hereinafter 'forests') , which enables further studies in the fields of ecological research, nature management and public health. Limitations of this model include the lack of data about wildlife, due to unavailability of this type of data at the national level and for the entire study period. Hazard predictions are included in this work by running our data-driven model for each day of each year included in the study period to compute the maximum annual tick activity, and by averaging these annual values to obtain a robust proxy of tick hazard at 1km (Figure 4.1). A visual inspection of Figure 4.2 shows that the area of highest tick activity is located in the northeastern half of the country (e.g. provinces of Overijssel, and Drenthe). Forested areas in the center and south of the country present an average tick activity and coastal areas present a low tick activity.

As explained in section 4.2.2, human exposure can be calculated by dividing the risk and the hazard components. Note that there are locations in which these measures are not available. This means that there are locations in which no tick bites are reported, or locations outside forests with no measurement on tick activity available that are excluded from the exposure calculation. Depending on the values of these variables, four cases will be found (c.f. Table 4.2): 1) risk and hazard are positive (i.e. R > 0 and H > 0); 2) risk is positive and hazard is undefined (i.e. R > 0 and H = undefined); 3) risk

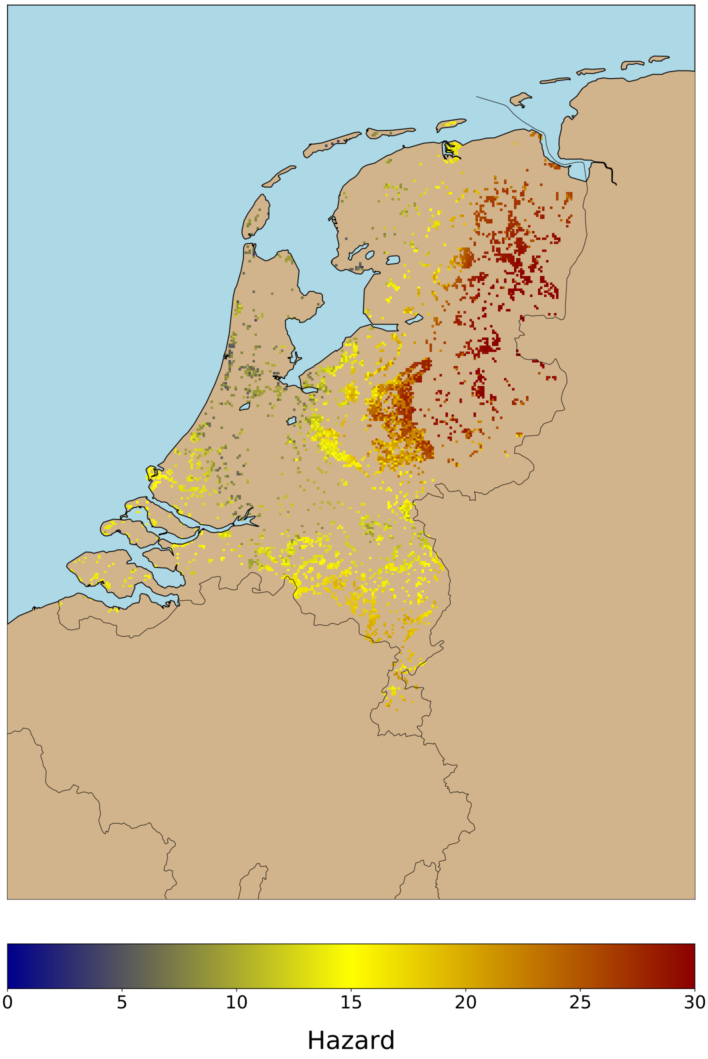

Figure 4.2: Hazard (e.g. tick activity) per 1km grid cell. We used the model developed in (Garcia-Martí et al., 2017) to predict daily tick activity for the period 2006-2016. Then, we calculated the maximum mean tick activity for the period to devise this map. The numbers in the legend indicate the average number of active questing ticks per grid cell. The locations of the highest hazard are within the provinces of Groningen, Drenthe, and Overijssel, whereas the lowest hazard levels are located along the coastal areas.

is zero and hazard is positive (i.e. R = 0 and H > 0); 4) risk is zero and hazard is undefined (i.e. R = 0 and H = undefined). Cases 2 and 4 lead to a mathematically undefined operation, hence the exposure is undefined too. Case 1 represents locations in which there is tick activity and human exposure. Case 2 can be used to characterize locations where tick bites are reported outside forests (e.g. urban and peri-urban areas). Case 3 depicts forested locations with a low recreational intensity or a low hazard (i.e. no tick bites reported), and case 4 shows areas in which there is no risk and hazard data available.

To ease the interpretation of the results, the resulting exposure was classified using the JNB algorithm. This algorithm minimizes the intra cluster variance and maximizes the distance between clusters. The optimal number of classes is iteratively found testing all values between 2 and 10 and calculating the goodness of variance fit of the resulting classification. Using the classes yielded by this algorithm, we created a categorical human exposure to ticks map.

To assess the validity of our exposure results, we compared them with a publicly available map depicting the attractiveness of the landscape (i.e. Belevingswaarde van het landschap, last accessed July 5th, 2018) (Crommentuijn et al., 2007; Roos-Klein Lankhorst et al., 2005). The attractiveness map relies on 3 positive variables (i.e. naturalness, terrain elevation, historical value), and 3 negative variables (i.e. visibility of the horizon, urbanization, noise pollution) to classify the attractiveness of each grid cell. In short, this map shows how much citizens find a landscape attractive, expressed as six categories in the range 6-8. The less attractive locations have an attractiveness lower or equal to 7, and the most attractive ones an attractiveness higher than 7. For our validity assessment we first extracted the value of landscape attractiveness for each of the forested locations with calculated exposure. Then, we counted the number of grid cells in our exposure map that belong to each of the six attractiveness classes. Finally, we normalized these counts by the total number of grid cells belonging to each attractiveness class to obtain the percentages of tick bites that occurred in each exposure class.

4.3 Results

Figure 4.3 shows the classified tick exposure map, obtained by dividing the risk of getting a tick bite risk (Figure 4.1A) by the hazard (Figure 4.2). The application of the JNB algorithm resulted in the identification of three exposure classes: low, medium, and high. A visual inspection of Fig. 4.3 shows that there is a high amount of grid cells belonging to the medium exposure class. Those cells are especially concentrated along the forest edges of the Utrechtse Heuvelrug forest and of the Veluwe national park (center of the country). The class high exposure corresponds to highly popular places for outdoor activities, such as the coastal areas from Haarlem to Middelburg, or with a lower intensity, areas close to Hertogenbosch, Eindhoven, and around the small forest patches between Groningen and Emmen (north of

	Landscape attractiveness	
Exposure	Low	High
High	29 (6%)	193 (6%)
Medium	41 (9%)	514 (15%)
Low	201 (43%)	1,987 (58%)
Zero	196 (42%)	726 (21%)

Table 4.4: Forested areas per human exposure and simplified landscape attractiveness classes (i.e. low corresponds to scores $\leq$ 7, whereas high to score s>7). Areas as expressed in km2 and as percentage over the total forested areas. There are 2,965 km2 of forests in which citizens are exposed to ticks, and 992 km2 with an exposure equal to zero (i.e. no tick bites recorded, although there might be a certain tick activity in those areas)

the country). The class low exposure indicates locations that are less visited, and yet visitors could get bitten by ticks.

Figure 4.4 explore the relationship between E, H, and R. The boxplot in Figure 4.4a show that the risk of getting a tick bite has a skewed distribution (i.e. long-tailed distribution) spanning up to 353 tick bites per grid cell (not shown due to visual cluttering of the box plots), regardless of the exposure class. The medians of the boxplots (plot A) correspond to 3, 8, and 23 tick bites per km^2 for the low, medium, and high exposure classes. The height of each box indicates the variety of risky conditions in which tick bites ensue. Low and medium exposure classes present a narrow range of risky conditions, whereas the high exposure class occurs in a wider range of conditions. The boxplots in Figure 4.4b show that the hazards have a unimodal distribution, regardless on the exposure class. The medians indicate that tick bites occur in locations with similar levels of hazard, and the fairly uniform height of the boxes show that the range of risky conditions in which tick bites occur is alike. Note that Figure 4.4a shows that the risk increases as the exposure increases, whereas Figure 4.4b shows how hazard is almost constant as long as the exposure increases.

Figure 4.5 shows the relationship between human exposure and the attractiveness of the landscape. Each cell represents a number of grid cells belonging to both categories. To ease the interpretation of results, note that we included the locations in which the exposure or the hazard are low (i.e. no tick bites registered in that forested location), and also that we applied a per-column normalization to turn the raw numbers to percentages. The visual inspection of the attractiveness layer and Figure 4.6 reveals that the first three columns show the exposure in forest patches which are less attractive for citizens (i.e. $\leq$ 7), whereas the last three columns correspond to attractive forested and rural locations (i.e. $>$ 7). As seen, the distribution of exposure grid cells within the first group is similar among the different attractiveness classes. In the second group, there are two columns showing

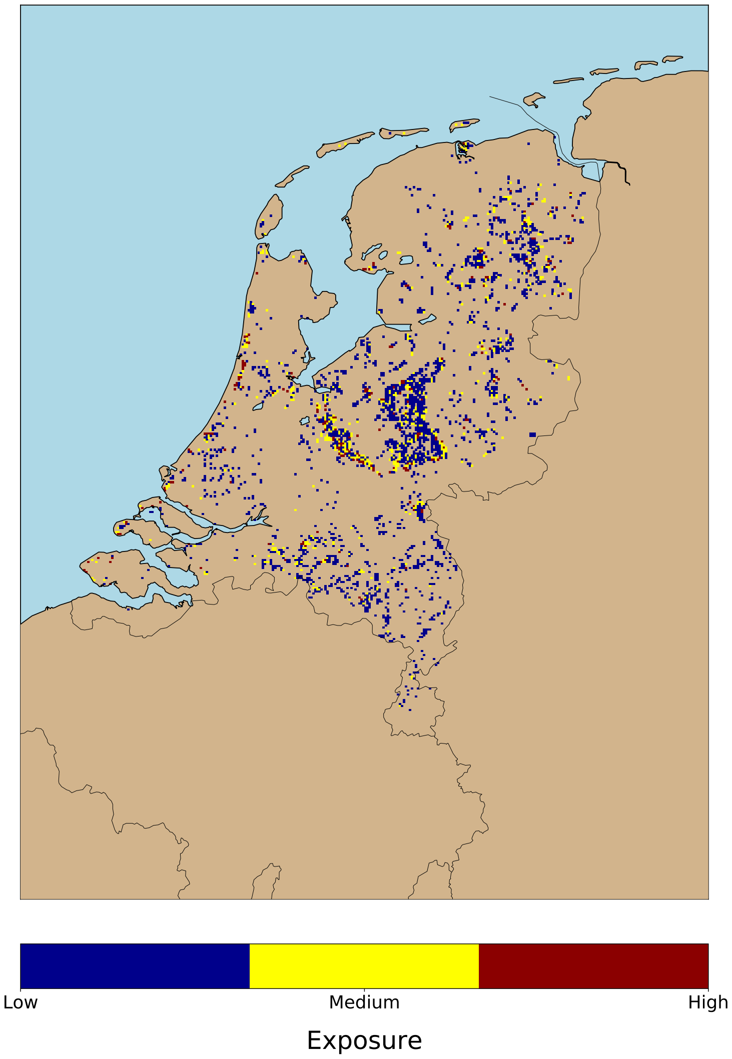

Figure 4.3: Human exposure to tick bites as a result of combining the maps in the previous two figures. Background color refers to non-forested locations or locations without tick bite reports. The exposure is classified in three categories. Well-known forest edges (e.g. Veluwe national park, Utrechtse Heuvelrug forest) and popular outdoor recreational areas (e.g. coastal areas from Haarlem to Middelburg) are classified as places with a medium and high human exposure. The remaining low exposure class depicts locations less intensely visited by people. Both results suggest that human exposure to tick bites is driven by two types of users (i.e. recreational, residential) as spotted in previous works (De Keukeleire et al., 2015; Zeman and Benes, 2014), that may require different treatment in the design of public health campaigns targeting a decrease on tick bites occurrence.

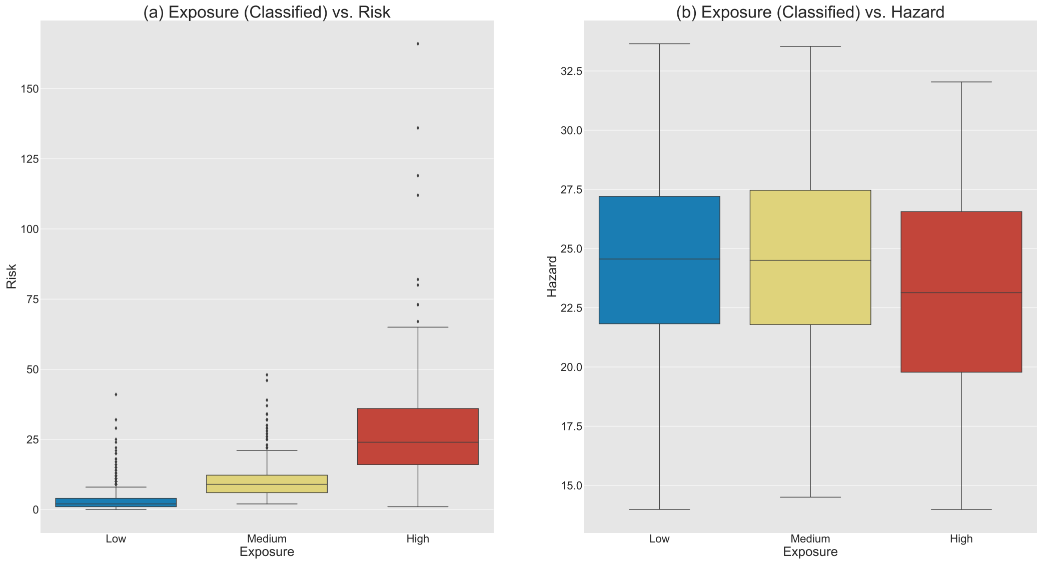

Figure 4.4: Boxplots showing the relationship between exposure and risk (a), and the relationship between exposure and hazard (b). For both figures, the X-axis shows the exposure class, and the Y-axis shows the number of tick bites per grid cells (a), and the tick activity per grid cell (b), respectively. Risk is a skewed distribution (i.e. long-tailed), thus presents low averages per boxplot, and a high number of outliers (reaching the maximum of 353 tick bite reports/cell), whereas hazard is a Gaussian-like distribution and so the averages per boxplot occupy the central part of the distribution. Plot A shows how the risk increases as the exposure increases, and plot B shows how the hazard remains (almost) constant as long as the exposure increases. This means that the risk of getting tick bites is mainly driven by exposure factors, regardless of the amount of hazard (e.g. tick activity) in a location.

interesting patterns. The fifth column shows that 65% and 26% of the grid cells in the attractiveness class 7.5 – 8 have a zero and low exposure, respectively. The last column shows that 17% and 61 of the grid cells in the > 8 class have a zero and low exposure, respectively. This indicates that citizens prefer visiting certain forested locations. The absolute numbers in Figure 4.6 show that the majority of the human exposure is concentrated along the column with the maximum attractiveness. In Table 4.4 we provide a summary on the area in which humans are exposed to ticks. As seen, citizens are exposed to ticks in 271 km^2 of forests in unattractive locations, and 2,694 km^2 of forests in attractive locations. Note that there are 922 km^2 of forest in which the exposure is zero. Thus, this study shows that citizens are more exposed to ticks in locations that are very appealing to the general public.

Figure 4.6 shows the four possible cases (Table 4.2) that can occur when dividing the risk by the hazard layer. To avoid visual cluttering, the three exposure classes from Figure 4.3 have been merged into a single class. Figure 4.6 also shows the locations with tick bites outside forests, forests with a low recreational intensity, and the locations with zero tick bites reported. The visualization of cases 2 and 3 (Table 4.2) reveals two new insights. First, we can see that the occurrence of tick bites is a pervasive phenomenon that goes beyond the forest edges, because there are tick bites reported out of this scope, and across the country. Thus, vegetated landscape types (e.g. residential areas close to forest, urban parks with a dense tree coverage, natural coastal dunes with dense shrubs), might be optimal locations in which ticks and humans are in close contact. Second, there is a number of small forest patches without tick bite reports, which indicates that citizens do not intensively visit these locations.

4.4 Discussion

Our results show that clusters of high human exposure are concentrated along forest edges and popular places for recreation. The analysis reveals that the exposure categories in Figure 4.3 mostly occur in forested locations that are very attractive locations, as seen in Figure 4.5 and Table 4.4. This can be related to previous literature, since transitional vegetation (i.e. ecotones) has been identified as a risky place to acquire tick-borne pathogens: humans tend to carry out outdoor activities along the forest edge, rather than going inside (Lambin et al., 2010; Linard et al., 2007). Moreover, forest edges are suitable locations for mammalian species to forage, and present higher abundances of ticks (Brownstein et al., 2005; Madder and Baeten, 2012). The Netherlands is heavily urbanized (EEA, 2011), which means that multiple land uses are intertwined in a small area unit, thus bringing humans and ticks in peri-urban and residential areas in close contact. All the above suggests that the exposure to tick bites is driven by two types of activities, namely recreational, and residential, as suggested in previous works (De Keukeleire et al., 2015; Zeman and Benes, 2014). Therefore, we suggest defining different LB prevention campaigns and public health policies for

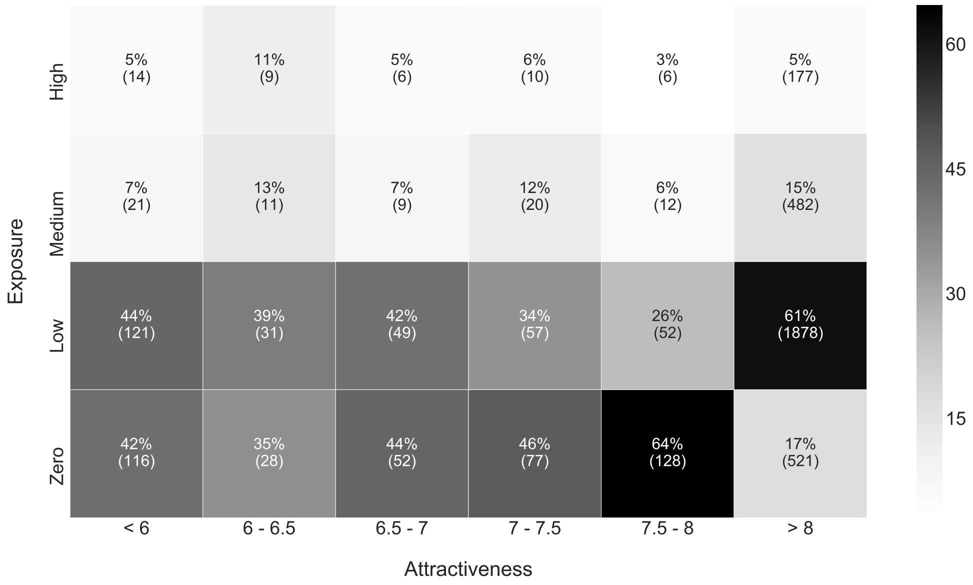

Figure 4.5: Heat map showing the relationship between the exposure classes and the attractiveness classes. The X-axis represents the six classes available in the attractiveness map, and the Y-axis the three classes of the exposure map. Thus, each cell in the heat map represents the number of grid cells belonging to both categories. Note that we applied a per-column normalization of the raw numbers to percentages to ease the interpretation of results, but both values are shown. The first three columns correspond to forest patches that are less attractive for citizens, whereas the last three columns correspond to attractive forested and rural locations. Thus, the first group of columns show a more urban exposure to ticks, whereas the second group of columns show human exposure to ticks in forested locations. The last two columns show an interesting pattern. The fifth column shows that 65% and 26% of the grid cells in the attractiveness class 7.5 – 8 have a zero and low exposure, respectively. The last column shows that 17% and 61% of the grid cells in the maximum attractiveness class have a zero and low exposure, respectively. This means that within forested locations, citizens have a preference for visiting a subset of them. Absolute numbers show that the majority of the exposure grid cells are concentrated along the column with the maximum attractiveness. This indicates that human recreational intensity is mainly concentrated in very appealing locations.

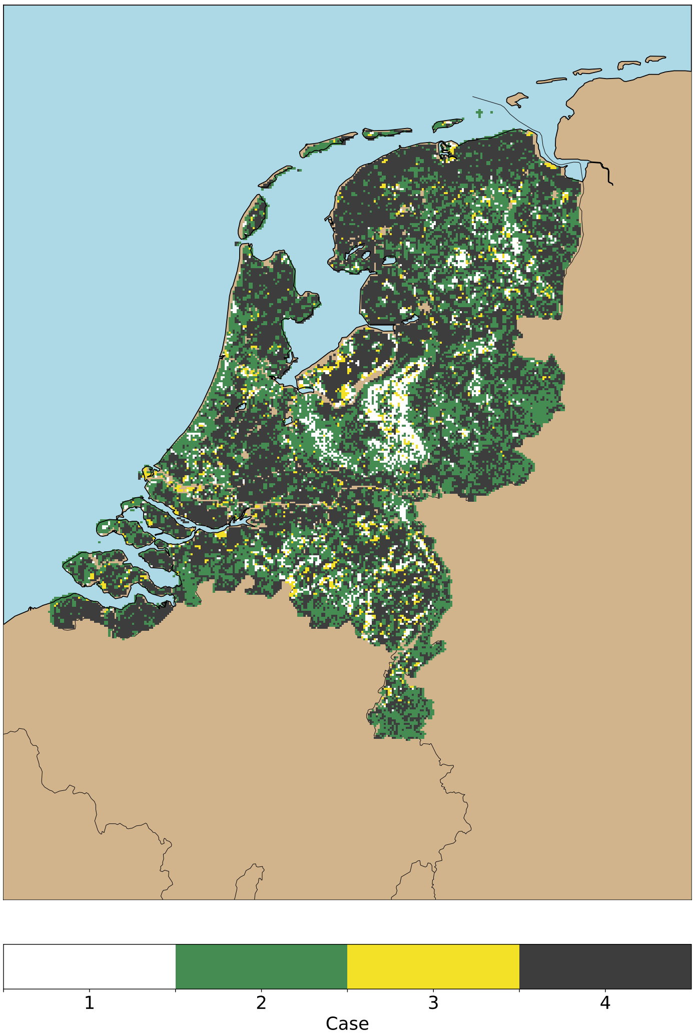

Figure 4.6: Visual representation of the four cases described in Table 1. To avoid visual cluttering, the classes in the first case (i.e. R > 0 and H > 0) have been condensed in one category (white). The remaining cases, namely, tick bites reported outside forests (i.e. R > 0 and H = undefined), forests with a low recreational intensity (i.e. R = 0 and H > 0), and locations with zero tick bites reported (i.e. low exposure or low hazard) R = 0 and H = undefined), are shown in the image in light green, yellow, and grey respectively.

each activity. For an instance, activities such as gardening (Mulder et al., 2013) can be linked to the residential exposure, whereas other activities such as scouting (De Keukeleire et al., 2015), or outdoor sport (Hall et al., 2017) competitions could be linked to the recreational exposure. Note that a limitation of this work is that we are unable to provide a measure of occupational risk, but farmers (Keukeleire et al., 2016), veterinarians (Szekeres et al., 2015), landscapers (Li et al., 2018), or forest workers (de Groot, 2016) are known to have an elevated risk of LB infection. Unfortunately, information on whether the tick bite was acquired during a work-related activity was only available for the TR data and is therefore not incorporated in the model. However, these collectives tend to present a higher seroprevalence for LB (Keukeleire et al., 2016; Szekeres et al., 2015).

The boxplots in Figure 4.4 show that the risk of getting a tick bite increases as the exposure increases, whereas the hazard remains constant as long as the exposure increases, which is indicative that the risk of getting a tick bite is driven by the human exposure in a location more than to the existing hazard. This makes sense, because humans might be unaware of the hazard, and do not consider this threat when organizing outdoor activities. For example, the recreational coastal areas between Haarlem and Middelburg present a relatively low hazard (Figure 4.2), however, the risk of getting a tick bite (Figure 4.1) is very high, because the human exposure around the area is high as well. This is also supported by Figure 4.5 and Table 4.4, as they show that citizens mainly get exposed to ticks in areas which are very attractive, and therefore, it is likely that they have high numbers of recreational visitors.

These new insights show that hazard maps alone are insufficient to identify locations with a high risk for LB, motivating the creation of human exposure maps for public health specialists and forest managers. In this line, maps like the one presented in Figure 4.3 and 4.6 may help to facilitate the cooperation between public health specialists and foresters to implement prevention campaigns. For instance, these maps can be used to classify patches of forests that require active management and those that can suffice with only public awareness campaigns. In this work we encountered three main hurdles. First, the difficulty of validating the exposure results, since exposure heavily depends on the quality of the hazard model, and on the representativeness of the tick bite reports. The hazard model can capture general tick dynamics. However, hazard predictions might be uncertain in locations in which atmospheric conditions are not the main driver of tick activity. At the same time, the risk map contains an unknown factor of citizen's reporting errors (e.g. positional inaccuracy at the time of adding the tick bite report to the NK and TR platforms, or citizens that over- or under-report tick bites). To overcome these issues and to validate the exposure products ICT data (e.g. mobile phones locations), geolocated data streams from social networks (e.g. Twitter) could be incorporated in the analysis. However, the use of this data is limited by privacy laws. Second, the exposure remains unknown in locations where there is no data from hazard or risk. This means that locations in which the hazard model is unable to predict the tick activity (e.g.

non-forested areas), or locations where there are no tick bites registered, it is not possible to apply eq. 1, thus we cannot estimate human exposure to ticks. Third, there is a substantial number of locations in which there are no tick bites reported during the study period. We do not think that the risk is actually inexistent in these locations, but with the current data collections we are unable to disentangle if the zero tick bite count is due to a low human exposure or a low hazard. As a consequence, we choose to exclude these locations from the exposure calculation, because it is hard to assess whether or not it is a true zero, and subsequently we do not know if it is a robust indicator for low risk of tick bite.

4.5 Conclusion

In this paper we present a first-of-its-kind map of human exposure to ticks in forested areas created from volunteered data. This map will hopefully contribute to mitigate the number of tick bites and hence of LB cases because exposure information might encourage forests managers and public health specialists to implement preventive measures. For instance, this map could be used to design targeted informative campaigns in recreational locations. Moreover, ecologists might use exposure information to locate hot spots of human disturbance, which could support better nature management practices. Future work should aim at quantifying uncertainties in the risk, hazard and exposure information, studying the main drivers of tick bites at each location (i.e. high hazard, vs. high exposure areas), and analyzing exposure in residential areas.

5 Modelling tick bite risk by combining random forests and count data regression models

5.1 Background

Over the last couple of decades, urban areas have dramatically expanded (EEA, 2006)]. In Europe, the development of low density residential areas in the periphery of cities has become the norm for urban growth (EEA, 2006). This phenomenon, known as urban sprawl, has a plethora of negative effects over the local climate (e.g. urban heat islands), the modification of the landscape (e.g. fragmentation), and the alteration of ecosystems (e.g. biodiversity loss) (EEA, 2011). In addition, urban sprawl also brings urban settlers in closer contact with nature and the countryside (Tack, 2013). As a response, several bird (e.g. thrushes) and mammal species (e.g. rodents, foxes, raccoons) have adapted their ethology to be able to live at the interface between forests and urban regions (e.g. more food, less predators) (Uspensky, 2017). This also means that the parasites and pathogens that several wildlife species carry get closer to residential areas and that, for instance, the hazard for tick-borne diseases increases (Allan et al., 2003). In parallel to this, the progressive adoption of healthier lifestyles encourages citizens to spend more time outdoors carrying out leisure (Mulder et al., 2013) or sportive activities (Hall et al., 2017). This behavior leads to a higher exposure to tick-borne diseases (Sandifer et al., 2015).

Socio-economic changes and the subsequent response of nature means that citizens are more vulnerable to tick-borne diseases today than in the past. Hence, the transmission of pathogens causing tick-borne diseases is an important public health threat (Ehrmann et al., 2017). In fact, recent research demonstrates that ticks have trespassed the limits of forests and natural grasslands to start inhabiting green spaces within metropolitan areas. Urban parks in Zurich (Oechslin et al., 2017), Milan (Olivieri et al., 2017), Kiev

This chapter is based on: (Garcia-Marti et al., 2019)
Available at: https://doi.org/10.1101/642728

(Didyk et al., 2017), Warsaw (Kowalec et al., 2017) or Lisbon (Santos et al., 2018), and suburban forests in Paris (Paul et al., 2016), Budapest (Szekeres et al., 2016, 2018) or Wroclaw (Kiewra et al., 2017), present tick populations, and researchers were able to identify pathogens capable of causing Lyme borreliosis (LB) or tick-borne encephalitis (TBE) in humans. Since parks and suburban forests are potentially visited by thousands to millions of citizens every year, it is necessary to fully comprehend and model the risk of getting a tick bite to prevent this major public health threat.

The risk of getting a tick bite is the result of the interaction between its exposure and hazard components. Traditionally, researchers have tried to represent the risk of LB by quantifying the hazard component alone (Eisen et al., 2010; Gassner et al., 2016; LoGiudice et al., 2003), but in the last years researchers have worked on integrating hazard and exposure metrics to model tick bite risk (De Keukeleire et al., 2015; Zeimes et al., 2014) because hazard maps alone are insufficient to identify locations with a high risk for LB (Garcia-Marti et al., 2018).

The location of citizens is key to model the level of risk they are exposed to, but acquiring this type of information requires a partnership of researchers and public-health specialists to create (inter-)national networks of surveillance and citizen observatories. In the Netherlands, Wageningen University & Research (WUR) and the Dutch Institute for Public Health and the Environment (RIVM) started two citizen science projects to collect data on ticks and tick bites. These projects, which started in 2006 and 2012, have attracted enough media attention over the years to engage citizens at contributing tick bite reports. This engagement has resulted in over 50,000 volunteered tick bite reports in the Netherlands. This unique dataset enables new approaches to monitor and model elusive public health threats, such as tick bites. However, volunteered data is often unstructured, contains positional inaccuracies and reporting bias, and observations have a variable quality (Mehdipoor et al., 2018a; Senaratne et al., 2017), conditions that might cause difficulties when including volunteered data in a scientific workflow.

A major challenge in our work was dealing with the highly skewed and zero-inflated distribution of the tick bite reports. These two types of disproportions are inherent traits of our data collection. Skewness refers to the asymmetry of the distribution, whereas zero-inflation refers to distributions in which zero-valued observations are frequent. In this work, our goal is creating a spatial tick bite risk model with national coverage. However, adding the spatial dimension implies the simultaneous modelling of a few locations reporting a high number of tick bites, and a substantial amount of locations in which zero tick bites are recorded.

Although these characteristics make it hard to use machine learning methods (Krawczyk, 2016), we pursue a solution based on machine learning because of its proven ability to handle non-linear and complex relationships. Classical count data statistical models are better equipped to handle skewed and zero-inflated datasets but they are unable to capture the inherent non-linearity in data. Thus, here we propose a solution integrating machine learning and classical statistic models. We combine the "segmentation capabilities" of the well-known Random Forest regressor (Breiman, 2001), which can split the data into homogenous groups using decision tree rules, with count data models of the Poisson family, which are better suited to model count data. Thus, our scientific contribution is two-fold: we present a tick bite risk model based on a wide array of hazard and exposure metrics, and we propose a methodological step forward by combining Random Forest and count data models to better model skewed and zero-inflated distributions.

5.2 Risk, exposure, and hazard

In the field of risk assessment, risk (R) is often modelled as a function of hazard (H) and exposure (E). The relationship between these three variables can be conceptualized as R = H × E (Braks et al., 2016). Thus, the calculation of tick bite risk should include variables representing both the H and E components, which likely have different underlying factors.

In the case of ticks and LB, the first spirochetes were identified in the early 1980s (Burgdorfer et al., 1982), and it took several years for the first studies to point out at human exposure to ticks as the source of the disease. Back then, various studies (e.g. (Falco and Fish, 1989; Magnarelli et al., 1995) had already identified urban recreational parks as risky locations for LB, thus recommending prevention campaigns at parks and to inform citizens living nearby a green space. LB emerges from a complex ecological system driven by a wide array of factors (e.g. wildlife, climate, vegetation) (Ostfeld, 2012). For over 20 years scientists have studied the interactions between these factors to devise robust models of tick hazard. Multiple efforts can be found in literature since the late 1990s to quantify and map this component of tick bite risk. However, in our recent research (Garcia-Marti et al., 2018) we found out that the E component may be more relevant to determine tick bite risk. The quantification of the E component is a challenging task, due to the unavailability of human exposure metrics at the national scale. Thus, in this work we devoted special effort and creative thinking at developing novel human exposure indicators, which are combined with our tick hazard model (Garcia-Martí et al., 2017) to predict tick bite risk.

5.2.1 Tick hazard

The H component of tick bite risk has been widely studied since the late 1990s. Scientists have thoroughly worked to understand the impact that wildlife (Ostfeld et al., 2006; Randolph and Storey, 1999), mast years (Buonaccorsi et al., 2003; Kelly et al., 2008), vegetation type (Tack et al., 2012), and weather variables (Berger et al., 2014b) have on tick populations. The pursuit of reliable models on tick hazard has prompted researchers to model this component of risk from multiple perspectives. Thus, we can find studies dedicated to tick habitat suitability (Estrada-Peña et al., 2015), tick presence (Swart et al., 2014), tick activity (Bennet et al., 2006), or tick dynamics (Garcia-Martí et al., 2017), with a varying number of biotic and abiotic parameters, and applied from local to continental spatial scales. In this work we use tick activity as a proxy for tick hazard. Tick activity represents the number of ticks that are questing for blood meals, which are the ones biting humans. Tick activity is extracted from a data-driven model that predicts daily tick activity in forests and natural grasslands (Garcia-Martí et al., 2017). The map in Figure 5.1 shows the predicted tick activity of this model, which is the average number of questing ticks per grid cell for the entire study period (2006-2014).

5.2.2 Human exposure

Human exposure to ticks is intrinsically linked to human behavior outdoors and to diverse socio-economic factors. For instance, (Zeman and Benes, 2014) discuss the peri-urbanization process in the Czech Republic, which prompted wealthy settlers to move away from large metropolitan areas into peri-urban areas to be in closer contact with nature and open spaces. This, in turn, triggered an increase in the number of tick-borne infections that was not directly related with any identifiable expansion on the tick habitat range.

Similarly, the societal adoption of healthier lifestyles encourages citizens to spend more time outdoors carrying out physical or sportive activities. For instance, in (Hall et al., 2017) the authors used a mass-participation cross-country marathon competition in Ireland to survey a large number of citizens and assess their exposure to ticks. Also, (Padgett and Bonilla, 2011) identify common picnic spots in a national park in the USA, as locations posing a risk of human exposure to ticks. Children participating in scouting or summer camp activities are found to be vulnerable to tick bites in a study in Belgium (De Keukeleire et al., 2015). All these examples are associated to the so-called "recreational exposure", however, there are also studies that pinpoint activities in the peri domestic environment as risky for tick bites. Previous works considering the "residential exposure" include a study in the Netherlands finding a high risk of tick bites in gardens (Mulder et al., 2013) and two studies in the USA (Hahn et al., 2017) and Czech Republic (Zeman et al., 2015) indicating that properties in the peri-urban environment with a large interface between a forest and the garden or backyard pose a high risk for inhabitants of acquiring pathogens.

A thorough study for TBE in Stockholm demonstrates a use case in which exposure and hazard variables are combined to obtain tick bite risk indicators (Zeimes et al., 2014). The authors create metrics for human exposure based in accessibility and scenic beauty, whereas for the hazard they include variables of wildlife, forest, and land cover. Indeed, accessibility measured as the distance to an access road or trail is an important variable to account for when modelling tick bite risk. In (Li et al., 2016) the authors assess the willingness of residents to travel for woodland leisure, because it varies as a

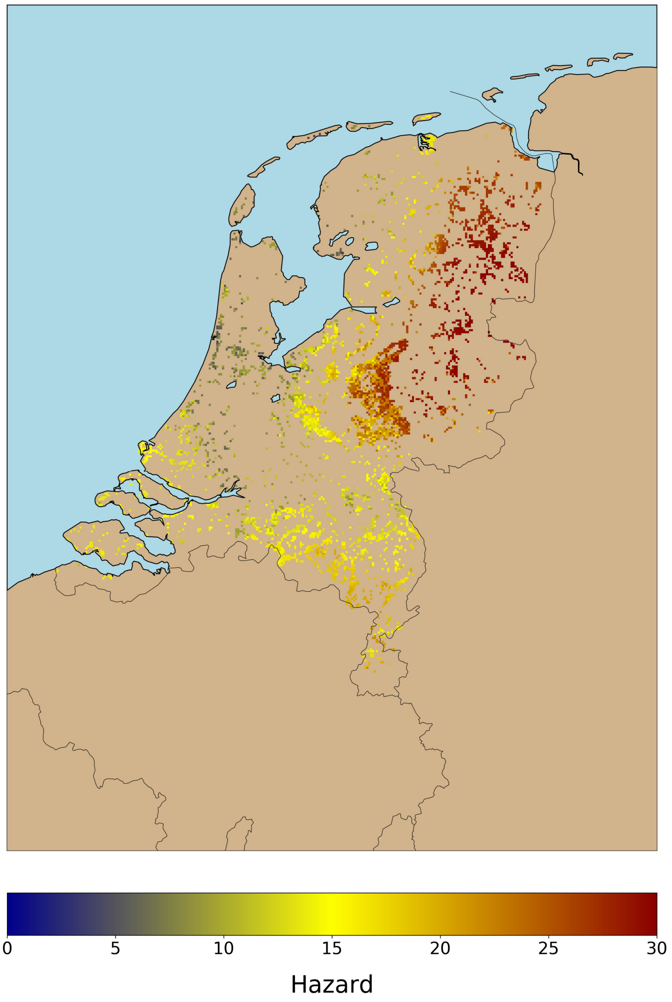

Figure 5.1: Tick activity per 1km grid cell. This tick activity estimation is provided by the data-driven model in (Garcia-Martí et al., 2017), which is capable of predicting daily tick activity. We run this model for each day during the period (2006-2014) and we calculated a robust long-term metric of hazard, showing the maximum mean tick activity for the entire period. As seen, hazard is minimum along the coastline and maximum in the northeast of the country.

function of whether citizens have to walk, cycle, or drive to the leisure place. However, accessibility is not the only factor to account for human exposure. There are locations in nature that are attractive for citizens due to presence of recreational areas or amenities (Lambin et al., 2010), or because they have intrinsic value such as high scenic beauty or a good preservation (Nielsen et al., 2012).

5.3 Methods

5.3.1 Tick bite risk monitored by volunteers

This work is based on the collection of volunteered tick bite reports from the Natuurkalender (NK; "nature's calendar"; www.natuurkalender.nl) and the Tekenradar (TR; "tick radar"; www.tekenradar.nl) citizen science initiatives promoted by WUR and the RIVM. During the six years (2006-2012) that NK registered tick bite reports, this platform gathered 9,256 user contributions, whereas the TR initiative collected 46,655 reports for the period 2012-2016. This means that in total there are 55,911 reports available. However, some of these reports lack geographic coordinates or are placed outside the boundaries of the Netherlands. Hence, a total of 46,838 valid reports were found. Here we approximate the risk of getting a tick bite in a given area by the cumulative sum of tick bites reports in that area for the whole study period (i.e. 2006-2016).

Prior to the modelling phase a spatial aggregation operation was used to transform the individual tick bite reports into a tick bite risk proxy. We choose a regular grid with cells of 1km^2 for the aggregation because that is the resolution of the existing hazard model described in Section 5.2.1. This aggregation groups together observations that are close in the geographic space (Figure 5.2a). However, it also creates a grid with right-skewed and zero-inflated (Figure 5.3) grid cell values. More precisely, the grid has a total of 36,866 cells. 9,985 cells have at least one tick bite report and the remaining 26,881 cells have zero tick bite reports. This means that for each grid cell with at least 1 tick bite, we have 3 in which no tick bites are recorded. Skewedness and zero-inflation are common real-world problems, especially when modelling count data, (Hadiji et al., 2015). Thus the analysis of the tick bite reports requires a modelling approach capable of handling these characteristics (Krawczyk, 2016).

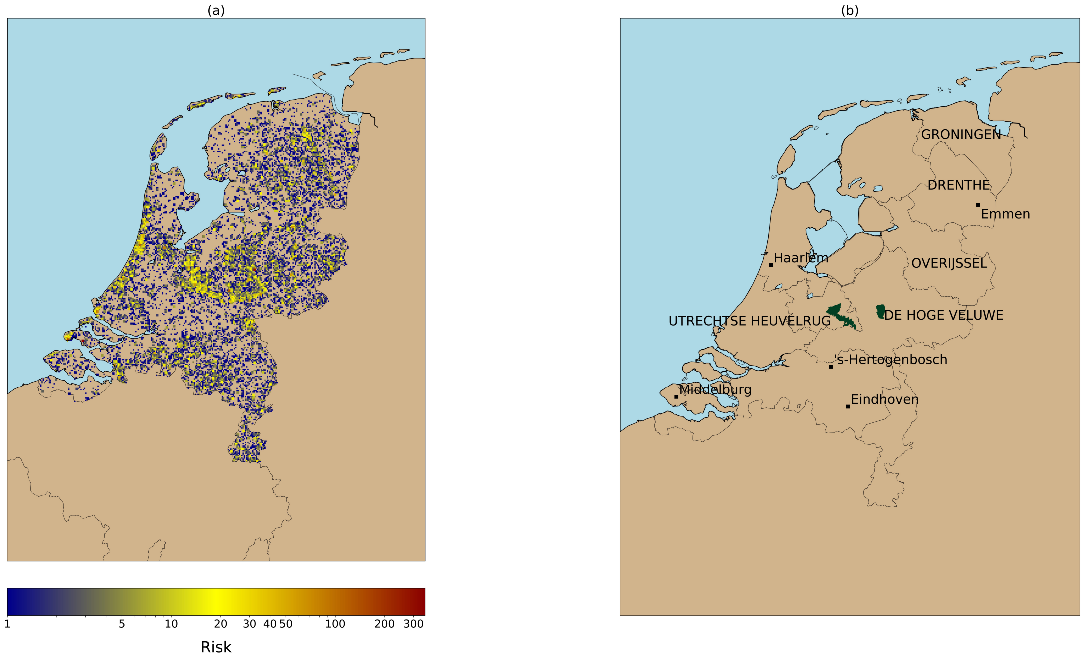

Figure 5.2: (a) Tick bite risk (2006-2016) as monitored by the NK and TR initiatives. We use as a proxy of tick bite risk the cumulative sum of tick bite reports per 1km grid cell. This image shows that tick bites are produced throughout the country, but the reports tend to be clustered around forests (e.g. Veluwe national park), or recreational areas (e.g. coastal areas). (b) Geographic locations. Provinces and national parks are labeled with capital letters, cities are labeled in lowercase.

5.3.2 Ensemble learning from skewed and zero-inflated datasets

Random Forest (RF) (Breiman, 2001) is an ensemble learning method that can be used both for classifications an regression problems. The ensemble is formed by decision trees, whose individual predictions typically have a high variance, but when combined, they produce a robust and highly stable estimator (Louppe et al., 2013). RF combines bagging (Breiman, 1996) and the random subspace method (Ho, 1998). Bagging allows RF to see multiple variations of the input data and the random subspace method introduces randomness in the features presented to each tree during the learning phase. These two mechanisms are responsible of the diversity of the ensemble. RF predicts unseen samples by averaging the predictions of the trees in the ensemble.

RF and other canonical machine learning algorithms work under the assumption of having a similar number of samples per class or range. If this is not the case, the application of a canonical RF tend to produce results biased to the majority class or most common values (Japkowicz and Stephen, 2002; Krawczyk, 2016). Learning from a imbalanced (classification) or skewed (regression) dataset, is a non-trivial problem that started to receive attention in the early 2000s (He and Garcia, 2009). According to (Krawczyk, 2016) there are three categories of methods to learn from imbalanced or skewed data: 1) data-level methods; 2) algorithm-level methods; 3) hybrid methods. Data-level methods aim at balancing the dependent variable by applying over/under sampling techniques. Algorithm-level methods require the modification of the method in use to (partially) remove the bias towards the majority class or most common range of values. The hybrid methods combine the balancing of the dependent variable and a modification of the method in use. We propose an algorithm-level method to mitigate the effects of the data imbalance over the predictive power of RF.

As explained in Section 5.3.1, the aggregation of the tick bite reports to create a 1km raster layer of tick bite risk created right-skewed and zero-inflated dataset (Figure 5.3). The sum of tick bites reported in each grid cell can be viewed as a discrete random variable that only takes non-negative values. This means that these reports can be modelled using well-known discrete probabilistic models for count data (hereinafter: count data models), such as Poisson (POI) and negative binomial (NB). Because of the large proportion of zero tick bites per grid cell, we also tested the zero-inflated versions of these models: the zero-inflated Poisson (ZIP) (Lambert, 1992), and the zero-inflated negative binomial (ZINB) (Greene, 1994) models. The difference between the original and the zero-inflated models is that in the latter type of models data is assumed to be derived from a two-stage process: 1) a Bernoulli trial deciding whether the event occurs or not (with probability p, the zero-inflation factor; 2) in case the event occurs, the counts will happen according to some rate Λ. Note that this second process can also generate zeros. This two-stage process is convenient for the problem we are modelling. First, we check the presence of ticks (and humans) and if present, we check

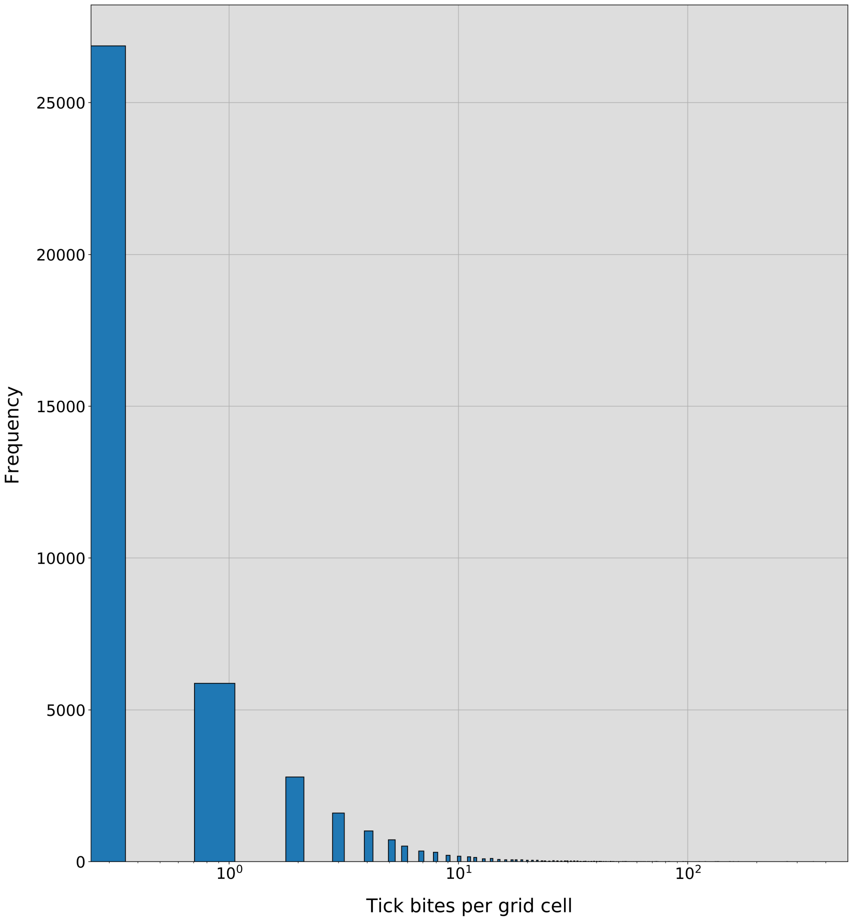

Figure 5.3: Histogram of the tick bites per grid cell. As seen, after the process of spatial aggregation described in Section 5.3.1, a skewed distribution with zero-inflation is created. Note that the X-axis is represented in log-scale to ease the visualization of the distribution. The number of grid cells containing more than 30 – 40 tick bite reports is almost negligible, but the distribution spans until a maximum of 353 tick bites per cell.

the "rate of transmission", conceptually composed of visiting rates and biting rates.

Zero-inflated models have been used to predict TBE in a set-up in which the majority of the available samples had a zero (Stefanoff et al., 2018) . However, this approach is limiting because count data models do not generally work well in set ups where there are complex non-linear interactions between the predictors and the response variable. In our work, the use of RF allows the inclusion of a wide array of predictors and the identification of the main ones to segment the problem into more homogeneous cases, which can afterwards be modelled using count data models. We propose a modelling approach that combines weak (i.e. decision trees) and strong (i.e. models for count data) estimators to improve the canonical form of RF. Figure 5.4 sketches our modelling approach where ensemble trees are only grown until their terminal leaves hold a minimal number of relatively homogeneous samples. These samples are subsequently analyzed with the four selected count data models (i.e. POI, NB, ZIP, ZINB). During the testing phase, each of the test samples will be propagated down each tree in the ensemble and will yield four predictions (one per count data model). The final prediction of the ensemble will be calculated by averaging the predictions of each model type, just like a canonical RF operates.

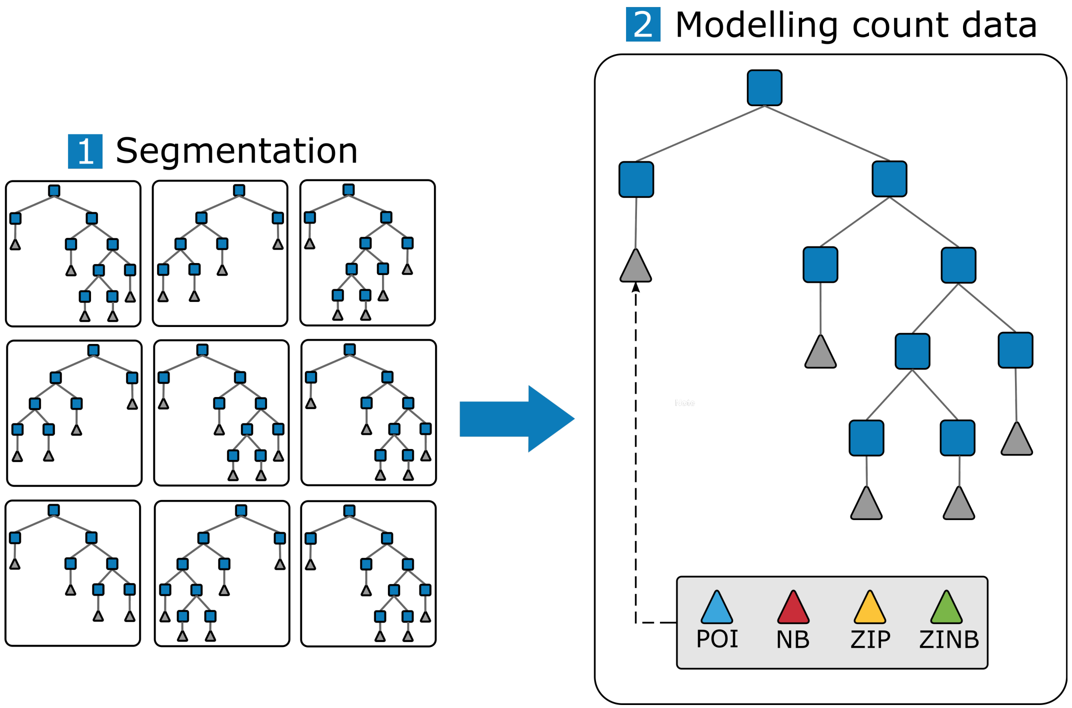

Figure 5.4: Proposed approach to couple RF and count data models (Poisson: POI, negative binomial: NB, zero-inflated Poisson: ZIP, and zero-inflated negative binomial: ZINB). First, the ensemble of decision trees is used to segment the samples into groups with similar characteristics. These trees are shallow trees, so that each of the leaf nodes contains a few hundred of samples. Second, we plug the selected count data models to each of the leaf nodes in the ensemble. The predictions yielded by the count data models are subsequently averaged to obtain the final prediction for each RF and count data model combination.

5.3.3 Modelling tick bite risk

The data and modelling approach described in the previous sections were used to model tick bite risk in the Netherlands. First, we explain the process of feature engineering applied to enrich each of the tick bite reports with hazard and exposure variables. Then, we describe our modelling experiments. Note that our work was developed using various Python libraries: numpy (Oliphant, 2006) to handle the multidimensional arrays, statsmodels (Seabold and Perktold, 2010) to fit the count data models, GDAL (GDAL Development Team, 2018) and cartopy (Met Office UK, 2010) to process geospatial data and prepare the visualizations through map layers, matplotlib (Hunter, 2007) to prepare the rest of the figures, and SkillMetrics [1] library and scipy (Oliphant, 2007) to obtain the statistical metrics used to evaluate the model.

5.3.3.1 Feature engineering

In this study, we extend the ideas regarding human exposure described in (Zeimes et al., 2014). To do so, we use a substantial amount of official Dutch geospatial data, and of other user-contributed geo-sources, to derive accessibility and attractiveness metrics. Because of the aggregation of the tick bites to a uniform raster layer, the exposure metrics where calculated as the geographic distance between the center of each grid cell and a set of selected real-world features in which we expect the co-ocurrence of humans and ticks. As explained in Section 5.2.1 hazard metrics are extracted for forests and natural grasslands using the model developed by (Garcia-Martí et al., 2017). Table 5.2 presents the 19 exposure metrics and the 2 hazard metrics that were used to model tick bite risk in the Netherlands. The following paragraphs contain more details about each of the metrics.

[1] https://github.com/PeterRochford/SkillMetrics

Id	Name	Data type	Description	Source	Provider	Type
1	dbuiltup	int	Dist. built-up area	BBG2008	Statistics NL (CBS)	AC
2	dforest	int	Dist. forest patch	BBG2008	Statistics NL (CBS)	AC
3	drecreation	int	Dist. recreational area	BBG2008	Statistics NL (CBS)	AC
4	dbrr	int	Dist. regional bike route	OpenStreetMap	OSM Foundation	AC
5	dwrl	int	Dist. local walking path	OpenStreetMap	OSM Foundation	AC
6	dwrr	int	Dist. regional walking route	OpenStreetMap	OSM Foundation	AC
7	dwrn	int	Dist. national walking route	OpenStreetMap	OSM Foundation	AC
8	dcamping	int	Dist. camping location	TOP10NL	Cadaster NL	AT
9	dcaravan	int	Dist. caravan parking location	TOP10NL	Cadaster NL	AT
10	dcross	int	Dist. bike cross circuit	TOP10NL	Cadaster NL	AT
11	dgolf	int	Dist. golf course	TOP10NL	Cadaster NL	AT
12	dheem	int	Dist. wild garden	TOP10NL	Cadaster NL	AT
13	dhaven	int	Dist. non-commercial haven	TOP10NL	Cadaster NL	AT
14	dsafari	int	Dist. safari park	TOP10NL	Cadaster NL	AT
15	dwater	int	Dist. waterbody (pond or lake)	TOP10NL	Cadaster NL	AT
16	dbath	int	Dist. authorized bathing location	Swimming locations	National Registry	AT
17	attr	cat	Attr. of location	Lanscape attr.	WUR / National Registry	AT
18	LU	cat	Land use at location	BBG2008	Statistics NL (CBS)	AT
19	LC	cat	Land cover at location	LGN7	WUR	AT
20	maxmeanhaz	float	Max. mean H 2006-2014	Tick dynamics	(Garcia-Martí et al., 2017)	HZ
21	maxstdhaz	float	Max. std. dev. of H 2006-2014	Tick dynamics	(Garcia-Martí et al., 2017)	HZ

Table 5.2: Features used in this work

We derived accessibility metrics (Table 5.2, indices 1 – 7, Type AC) from the land-use map (BBG 2008) provided by Statistics Netherlands (CBS [2]), and from the transportation networks contributed by volunteers in OpenStreetMaps (OSM [3]). Using these data sources, we calculate the distance from each grid cell to a series of selected land uses and transportation networks. We compute the geographic distance (in meters) of each grid cell to the closest of selected BBG 2008 land use types, namely, forests, recreational areas and urban areas. We downloaded the latest snapshot of OSM for the Netherlands (last access, July 2018) and extracted the user-contributed cycling and walking networks. The former is available at local, regional and national scales whereas the latter is only available at the national scale. Note that the bike networks do not overlap so, for instance, the national routes do not include routes between small cities or forest patches, but longer routes connecting the edges of the country. We compute the geographic distance (in meters) between each grid cell and each of the selected cycling and walking networks.

We obtained attractiveness metrics (Table 5.2, indices 8 – 19, Type AT) by using data from the Dutch Cadaster, the Dutch National Registry, and WUR. From the Dutch Cadaster, we use the so-called functional polygons of their TOP10NL [4] product. These polygons demarcate areas with 296 types of functions (Full list available: http://geoplaza.vu.nl/data/dataset/top10nl, last accessed November 2nd, 2018). Here, we selected 8 functions related to outdoor activities where humans could meet ticks: campings, caravan parks, bike cross circuit, golf courses, wild gardens, non-commercial havens, and safari parks. We also extract from the TOP10NL the location of all lakes and ponds in the country, since they can serve as attractors of visitors to nature due to its scenic beauty or recreational use. We include a publicly available map (Dutch National Registry) categorizing the attractiveness of the Dutch landscape (i.e. *Belevingswaarde van het landschap* [5] , last accessed July 5th, 2018) (Crommentuijn et al., 2007; Roos-Klein Lankhorst et al., 2005). Finally, for each location, we extracted the land use and land cover categories from BBG 2008 and LGN7 database (produced by WUR), respectively.

The hazard metrics (Table 5.2, indices 20 – 21, Type HZ) are extracted from the model outlined in Section 5.2.1. This model, described in more detail in (Garcia-Martí et al., 2017), is based on nine years (2006-2014) of data collected by volunteers. These volunteers carried out a monthly sampling by cloth dragging of 15 vegetated locations in the Netherlands, counting the number of ticks per life stage (i.e. larvae, nymph, and adult). Our model was calibrated for the nymph life stage only, since nymphs pose the highest hazard to humans. The model also includes 101 biotic and abiotic environmental predictors. These predictors describe the tick habitat conditions (e.g. litter, moss), the occurrence of mast years for three tree

[2] https://www.cbs.nl/en-gb

[3] https://www.openstreetmap.org

[4] https://zakelijk.kadaster.nl/-/top10nl

[5] https://data.overheid.nl/data/dataset/49505-belevingswaarde-van-het-landschap

species, weather conditions (e.g. temperature, evapotranspiration, relative humidity), satellite-derived vegetation indices (e.g. NDVI), and land cover. To account for the effect that short- and long-term weather conditions have on tick activity, we aggregated the weather data to 11 temporal resolutions (i.e. 1–7, 14, 30, 90, 365 days). We run our data-driven model for each day in the period 2006-2014. Then we computed the average of the maximum tick activity of each year and its standard deviation to obtain robust and long-term proxies for tick hazard in forests and natural grasslands locations. This means that outside of these locations, the hazard is unknown. In this case, we are unable to use value imputation, since this would require imputing values for most of the country. Instead, outside of these locations, we used a symbolic value of minus one. This value is meant to separate locations for which we have and do not have hazard data. The selection of this value is backed up by recent research (Heylen et al., 2013), which shows that tick densities decrease along the forest-urban land use transition. Thus, the symbolic value of minus one helps at grouping together samples without hazard and samples with low hazard, which tend to occur outside forests.

5.3.3.2 Experiments

The spatial aggregation and feature engineering described in sections 5.3.1 and 5.3.3.1, resulted in a matrix with 36,866 rows and 21 columns. Each row represents a grid cell and each column the E or H features selected for this work. A series of experiments were designed to identify a tick bite risk model that can handle the skewness and zero-inflation present in this matrix. First, we randomly selected 60% of the data for training all the models and reserved the remaining 40% for testing them. Then, we defined a range of values for the two main RF parameters of our ensemble: 1) the number of tree estimators; and 2) the number of samples per terminal leaf node. We trained ensembles using 10, 20, and 50 trees and where each tree had at least 100, 200, 400, 600 and 800 samples per terminal leaf. The number of samples per leaf node determines the "level of development" of the trees in the ensemble. Thus, experiments with few samples per leaf node (e.g. 100 samples) create deep trees close to full development, whereas shallow trees are created when there are many samples per leaf node (e.g. 800 samples). In total, 15 RF ensembles were trained using the same split of training and test samples. Subsequently, these RF ensembles were crossed with the four discrete probability models for count data (i.e. POI, NB, ZIP, ZINB), which were fitted using a non-parametric approach (i.e. without having to specify any hyper parameter), using a Nelder-Meade optimization routine to obtain the maximum likelihood estimates of the parameters of the distributions.

Two issues could hamper the fitting of the count data models: excessive skewness or excessive zero-inflation. The selected count data models can deal with skewed distributions, but the segmentation carried out by RF might leave the leave nodes with a subset of samples highly skewed towards zero (i.e. 85% - 100% of zeros). We explored how often these circumstances occur for each tree in the ensemble and we found out that in average, the fitting does not converge in 5% - 9% of the leaf nodes in the ensemble. In those cases, we keep the default behavior of a canonical RF, which is returning the mean of the samples falling in that node. Finally, model performance is checked with the test dataset. For this, we track the itinerary of each of the test samples down the tree, and identify the leaf node in which it ends up. Then, we pass this sample to each of count data models fitted with data from that node to get the prediction of that tree. We do this for each tree in the ensemble, and we average these tree-based predictions to get the final (ensemble) prediction, following the default behavior of canonical RF models.

A modified Taylor diagram is used to graphically summarize the test results. This diagram shows three statistics in a single plot: the standard deviation, the root-mean-squared deviation (RMSD), and Pearson's correlation coefficient. Taylor diagrams represent the relationship between the three variables, which essentially lie on a 2D manifold and are projected onto a 2D flat geometry without loss of information. We use this diagram because an accuracy metric like RMSD alone is not informative in the case of heavily skewed data, where also measures of dispersion play an important role. Since our distribution is skewed and zero-inflated, we substituted Pearson's coefficient by Spearman and Kendall Tau ranked correlation coefficients. The ensemble and count data models combination that yield reasonable trade-off between RMSD, standard deviation and correlation coefficient are then used to create tick bite risk maps for the Netherlands. The map created when using a canonical RF is added to the list of best models to be able to evaluate the advantages of our approach. Finally, we intersect these maps with the human exposure layers showed in Figure 5.5.

5.4 Results

Figure 5.6 shows the ability of the four count data models and the canonical RF to fit the skewed distribution of tick bites in the different set-ups. Overdispersion is better fitted by POI and ZIP models compared to NB and ZINB, since the former models yield values between 0 and 30 tick bites per grid cell, whereas NB and ZINB are barely able to predict beyond 10 tick bites per cell. Interestingly, the zero-inflation seems to be better captured by NB and ZINB than POI and ZIP, as seen by the frequency of predicted zeros of these models is similar to the original distribution. RF performs similarly in all the prepared set-ups, and seems unable to predict over a wide range of values, most values typically being constrained to below 5 tick bites. In general terms, the NB, ZIP, and ZINB models seem to capture reasonably well the original distribution, but the POI model and the canonical RF do not perform well: the POI model yields predictions with a frequency considerably higher than the original values, whereas RF is unable to predict beyond few tick bites per grid cell. In addition, POI and RF are not able to capture the zero-inflation. As seen, the predicted distributions do not seem to considerably improve or deteriorate based on the increasing number

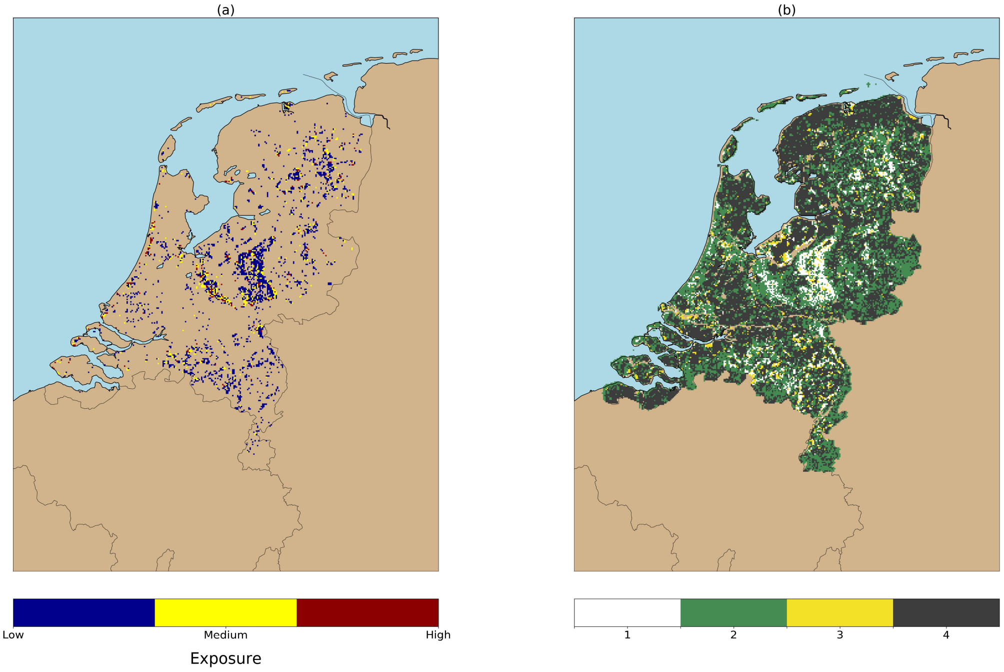

Figure 5.5: (a) Human exposure to tick bites classified in three categories. (b) Class 1 in this map correspond to the classes from (a), class 2 represents tick bites reported outside forests, class 3 represents forests with no tick bites recorded, and class 4 shows locations where no tick bites were reported during the study period. These results can be found in (Garcia-Marti et al., 2018), and we cross them with the tick bite risk maps obtained in this work to explore the risk per human exposure category (Figure 5.9).

of samples per leaf node (i.e. 100-600 samples), but the experiments with shallow trees (i.e. 800 samples) seem to have a negative impact in the ability of the models to predict zeros.

Figure 5.7 shows the performance of the ensemble in two modified Taylor diagrams. Each of the colored symbols represents an ensemble with a concrete number of tree estimators (T) and samples per leaf node (S). A visual inspection of the diagrams reveals that all ensembles yield predictions that are strongly correlated (i.e. correlation > 0.8 for Spearman's and > 0.7 for Kendall's Tau) with the tick bite data. The Taylor diagrams also show that the RMSD of these models is within a reasonable and stable range (i.e. 1 – 6) for all the experiments. However, in this work we are not only interested in models with a high correlation and low error, but also in those providing a realistic range of predictions, which is given by the standard deviation (stdev) represented by the dotted radial axes. The models present a variable skill to account for overdispersion (i.e. stdev 1 – 8). Considering these three statistical metrics, we think that the models better performing are located under the arc created by RMSD=2. Using the pink hexagon as a reference point, we can see that there are NB, ZIP, and ZINB models below this arc present a high Pearson/Kendall correlation (i.e. >0.9), a low RMSD (i.e. <2) and a fair range of stdev (i.e. 2 – 5). Out of these selected models, we can see that 2 ZIP and 1 ZINB models present a higher skill to model overdispersion (stdev > 4), whereas the small cluster of NB and ZINB models under the arc are better suited to predict zero-inflation. These diagrams also show that the optimal experiments correspond to 200-600 S and 20-50 T. To create the tick bite risk maps, we select the experiments with 200 samples per leaf node and 20 tree estimators since we believe they provide the best results. Figure 5.8 shows the tick bite risk maps produced by the four count data models (a-d), by the canonical RF (e), and a zoom-up of the maps obtained with the ZIP and ZINB models (f-g). The application of POI and ZIP models at the country level create maps that present a wide range of predicted tick bites. The NB and ZINB models yield maps that are visually less prominent, and the predictions of RF are mostly uniform throughout the territory and do not show any remarkable pattern. In Figure 5.6 and 5.7 we see that NB and ZINB present a higher skill to model zero-inflation, which means that they perform better than POI or ZIP at delineating regions with a low tick bite risk. In this sense, NB and ZIP mark large areas in the northwest (e.g. province of Friesland) and in the center west (e.g. region of the Groene Hart) of the country, either as very low or inexistent risk of tick bites. Figure 5.6 and 5.7 also show that ZIP has a better ability to predict over dispersed data, which is particularly suitable to identify less prominent risky locations, such as the patchy forest structures of the northeast of the country (e.g. provinces of Groningen and Drenthe) and the forests in the south (e.g. province of Noord-Brabant). The inspection of the zoomed ZIP and ZINB models (Figure 5.8; f-g) shows that the risk is maximum in popular recreational locations. The Veluwe national park, the Utrechtse Heuvelrug forests, and the recreational areas along the coast present the highest tick bite risk of the country.

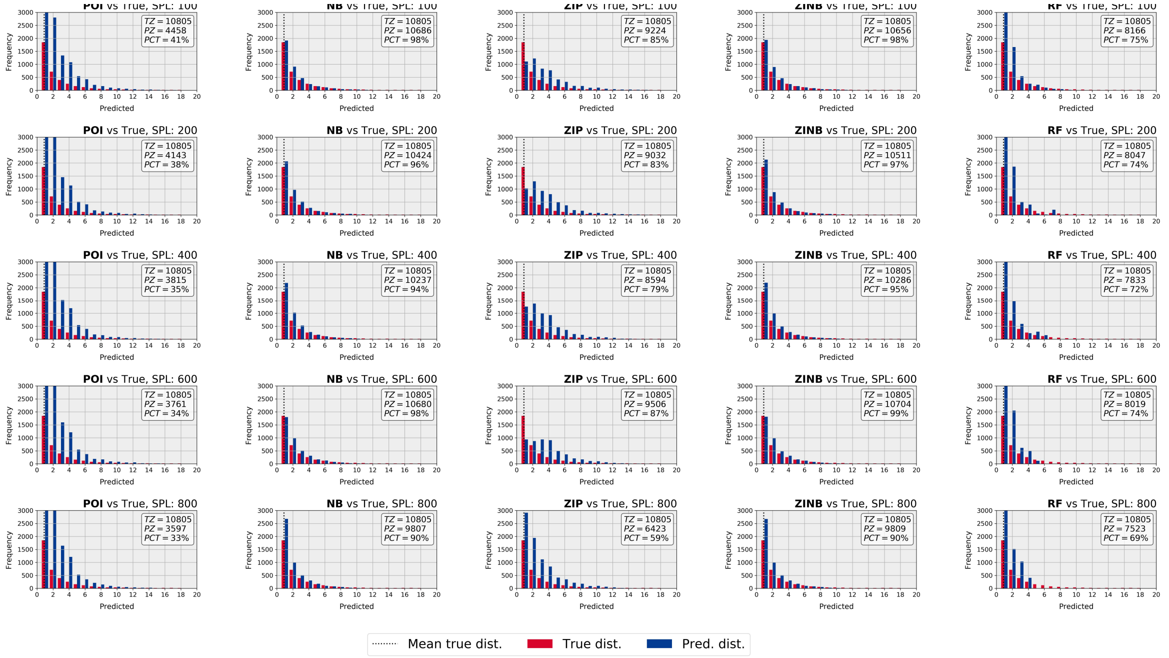

Figure 5.6: Histograms of original (red) and predicted (blue) distributions for an ensemble with 20 trees and an increasing number of samples per leaf node. Note that for visualization purposes, the axes have been limited and the zeros are summarized in the text box within each subplot, thus showing the number of true zeros, predicted zeros and the associated percentage. The first four columns correspond to the count data models, whereas the last column shows the performance of the canonical RF. As seen, POI and ZIP can predict for a wider range of values, whereas NB and ZINB are good at predicting zeros and the low part of the distribution.

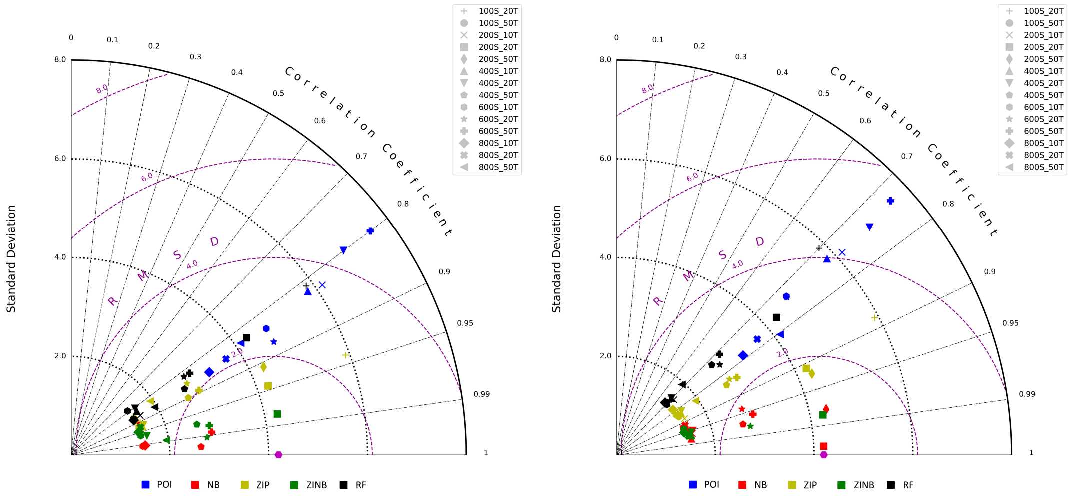

Figure 5.7: Modified Taylor diagrams showing the performance of the models based in three metrics: standard deviation, RMSD, and a correlation coefficient (left: Spearman's, right: Kendall Tau), which are represented by the Y, circular, and radial axes, respectively. Each of the colored symbols represents an ensemble with a concrete number of tree estimators (T) and samples per leaf node (S). The models better performing are located under the arc created by RMSD=2, since they present a high Pearson/Kendall coefficient, low RMSD, and a fair standard deviation. Out of these selected models, we can see that 2 ZIP and 1 ZINB models present a higher skill to model overdispersion (i.e. std. dev. > 4), whereas the small cluster of NB and ZINB models under the arc are better suited to predict zero-inflation. As seen, experiments with in the range of 200-600 samples per leaf node seem to perform optimally in both diagrams.

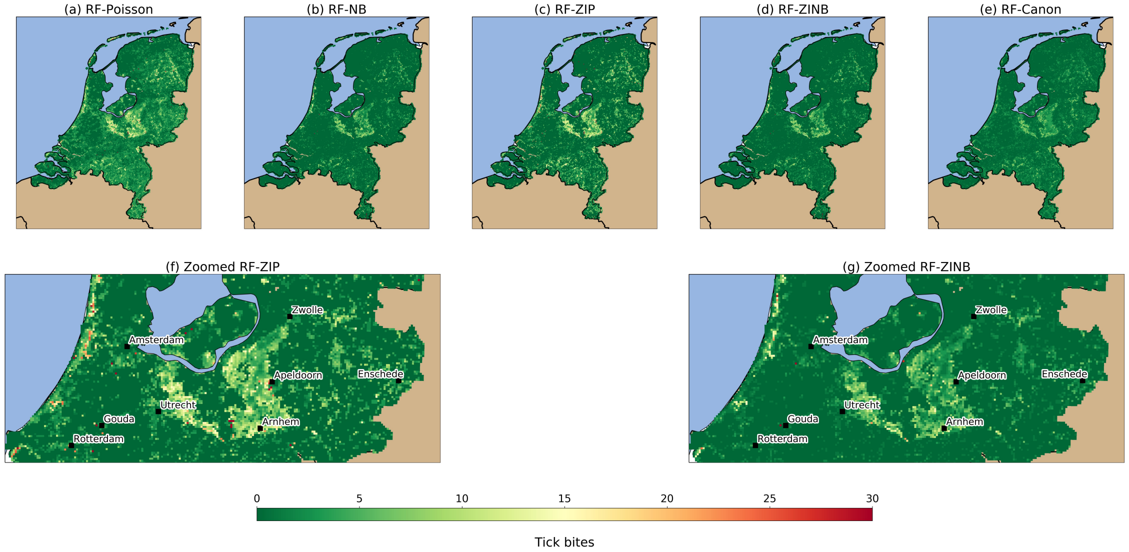

Figure 5.8: Tick bite risk maps produced by combining RF with each of the count data models. The upper row (a-e) shows the general overview of the models, whereas the bottom row shows a close-up for the ZIP (f) and ZINB (g) models. The NB (b) and ZINB (d) models are better suited to delineate regions with low or inexistent tick bite risk, thus they present sharper declines between different land uses. The ZIP (c) model is capable of predicting the risk of tick bite over a range of values, this is why its associated map presents smooth and gradual changes across the study area. POI (a) and RF (e) are over/under predicting, respectively, since the former finds risk in most locations of the country, whereas the latter yields and almost-homogeneous prediction. The visual inspection of the zoomed models (f, g) identify popular places for recreation intensely visited by citizens. The forested areas between Utrecht, Apeldoorn, and Arnhem, together with the recreational areas along the coast are regions where tick bite risk is particularly high.

Figure 5.9 depicts the risk of tick bite classified by the exposure levels found in Figure 5.5. We show a plot for each count model (a-d), for the canonical RF (e), and the original volunteered reports from NK and TR (f) classified by type of human exposure as well. The models have different skills to predict risk for each of the exposure classes in Figure 5.5. Considering the low, medium, and high exposure classes in Figure 5.5a, we see that ZIP better captures the overdispersion of data, since its interquartile range across classes (i.e. 0 – 10 TB/cell) is longer than NB and ZINB models (i.e. 0 – 6 TB/cell). Also, NB and ZINB present a higher skill at modelling the low exposure class, since it coincides with the original tick bite distribution (i.e. 0 – 4 TB/cell). ZIP provides more flexibility at predicting for the medium (i.e. 3 – 9 TB/cell) and high (i.e. 3 – 10 TB/cell) exposure classes, since these they span a range resembling the original distribution (i.e. 0 – 6 TB/cell and 0 – 14 TB/cell, respectively). Regarding the exposure classes in Figure 5.5b, we can see that ZIP seems to capture well the category of tick bite risk in non-intensively visited forests. The NB, ZIP, and ZINB models are not able to predict a range for the category of tick bites outside forests. The canonical RF shows a poor performance across the exposure classes, since it is only able to predict for a narrow margin of the original distribution. Based in the results provided in this section, we believe that, overall, the ZIP and ZINB models present stable predictions and the ability of modelling overdispersion and zero-inflation, respectively.

5.5 Discussion

In this work we illustrate that canonical RF models have difficulties capturing skewed distributions and we present our approach conceived to mitigate the effects of biasing the model towards the mean. To do so, we apply an algorithm-level modification of RF (Krawczyk, 2016), by combining weak estimators (i.e. decision trees) with robust estimators (i.e. count data models). By doing this, we keep two important characteristics of both types of estimators: a fast segmentation of the samples, and a realistic prediction of the tick bite risk. Thus, the integration of the segmentation capabilities of RF and the count data models creates a robust combined estimator.

Due to the skewed and zero-inflated nature of the tick bites per grid cell, our work does not aim at creating a model with the lowest performance metrics (i.e. low RMSD and standard deviation), but a model that finds the trade-off between the error and the capability of predicting tick bites over the reaslistic range of data values. For this, we tested various model configurations. The metrics represented in Figure 5.7 show the performance of the models based in three metrics: the standard deviation, the RMSD, and a correlation coefficient. Based on these three metrics, we think the that the ZIP and ZINB models are the ones performing better, since they present good correlation coefficients, reduced RMSD and are able to capture overdispersion or zero-inflation, respectively, which can open the door to multiple applications in ecological modelling and public health.

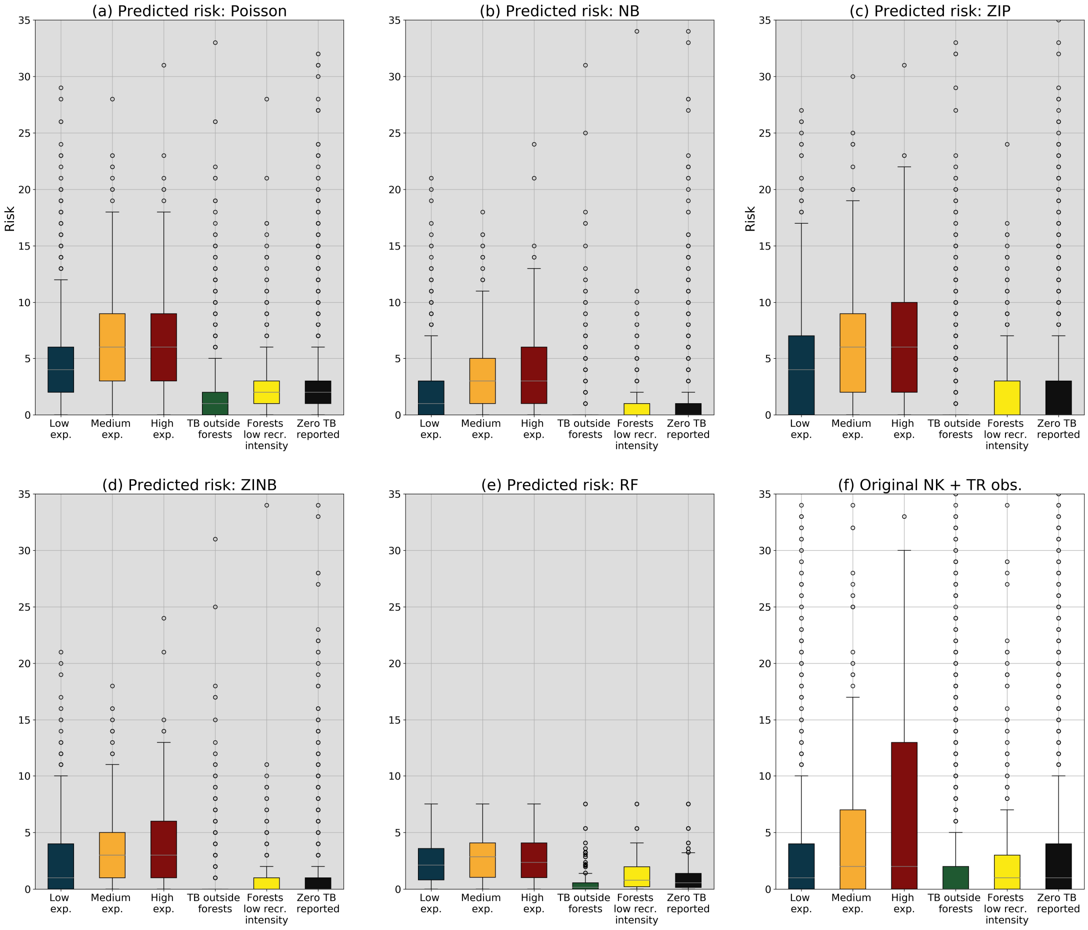

Figure 5.9: Tick bite risk classified based on the human exposure classes from Figure 5.5. The subplots show the predicted distributions per human exposure class for each of the count data models (a-d) and RF (e), and the type of human exposure using the original volunteered observations from NK and TR (f). The models present a different skill at predicting for each of the exposure classes. For example, ZIP and ZINB yield predictions for the low exposure class very similar to the original ones, whereas ZIP has analogous predictive capabilities for the forests with a low recreational intensity. Note, however, that although ZIP and ZINB can model the medium exposure class reasonably, none of the used models are able to capture the high skewness present in the high exposure class. The canonical RF is not able to predict the over dispersion of the original dataset.

The presented maps illustrate that the proposed approach can be used to estimate the tick bite risk in a location. The NB and ZINB models seem adequate for low-risk detection, since they perform better with zero-inflation in data, which subsequently enables the identification of low-risk regions. Then, the ZIP models are more suitable to fit over dispersed data, which enables the quantification of the risk within a wide range of values. Visually, this means that NB and ZINB maps identifies large regions with low-risk with sharp declines between different land uses, whereas the predictions yielded by ZIP show a richer range of predictions that can help at location risky locations in the country. The selected ZIP and ZINB models are able to identify locations of high risk in popular recreational places (e.g. Veluwe national park, coastal recreational areas), but they also have proved useful at detecting risky locations which are less intensively visited by citizens (e.g. patchy forests in the provinces of Noord-Braband, Drenthe, and Groningen). We believe that these maps can support several public health interventions intended to decrease the number of tick bites.

Using the categories from Figure 5.5 and the map layers in Figure 5.8, we inspected the risk of tick bites in function of human exposure inside and outside forests. In Figure 5.9 we see that some of the models are able to predict reasonably well for certain human exposure categories. For example, ZIP and ZINB yield predictions for the low exposure class very similar to the original ones, whereas ZIP has analogous predictive capabilities for the forests with a low recreational intensity. Note, however, that although ZIP and ZINB can model the medium exposure class reasonably, none of the used models are able to capture the high skewness present in the high exposure class. This limitation suggests that human exposure in highly visited locations might need additional features to better characterize the human activities outdoors. Considering all insights together, we think that these results suggest that a combination of RF and ZIP would be the most suitable one to estimate the tick bite risk in a location, whereas the combination of RF and ZINB would be adequate to detect locations with zero or very low risk.

In this work we encountered four hurdles. First, finding a proper validation metric for skewed distributions was challenging, because the most commonly used measures of model performance use statistical measures of location, not of dispersion, whereas in this case we are equally interested in predicting the dispersion. The (modified) Taylor diagram can help at evaluating the models because it can represent three statistical metrics in a single chart. Second, the TB collection is self-reported by each user of NK and TR. This means that this is a source of spatial inaccuracy based on the level of map literacy and spatial awareness of each user. With the current data collection, we are not able to quantify, nor correct, for this spatial inaccuracy. This means that at the time of the feature engineering we might be characterizing an observation which is incorrectly placed. We acknowledge the importance of citizen science campaigns, but we recommend that further data collection campaigns dedicate effort to find the positional accuracy of each observation. Third, there is a small fraction of the non-parametric count data model fittings that fail to converge due to excessive data imbalance for the optimization routines. Further work should aim at incorporating statistical knowledge to improve the fitting procedure, so that all models converge and contribute to the joint prediction of the ensemble. Fourth, the hazard model used in this work can produce a robust estimation of tick activity within forests, but not on other land uses. Thus, in this work the contribution of E and H could only be estimated in forested areas, whereas in the remaining land uses the model is entirely driven by E features. Further work should aim at combining different hazard metrics (e.g. tick suitability) to obtain a continuous picture of tick hazard throughout the country. This improved hazard metrics could help at disentangling which of the two factors (i.e. E or H) is dominant for each location, and thus would allow a deeper understanding of the factors of tick bite risk.

5.6 Conclusion

In this work we illustrate how canonical machine learning algorithms like RF may not perform well at modelling a skewed and zero-inflated distribution, and we present our algorithm-level solution to mitigate the bias towards the mean. Our approach consists in modifying the default behavior of RF by combining weak estimators (i.e. decision trees) with robust estimators (i.e. count data models). Thus, we connect four discrete probability models for count data (i.e. Poisson, negative binomial, zero-inflated Poisson, and zero-inflated negative-binomial) to each of the leaf nodes of RF. Subsequently, we enable RF to predict for skewed and zero-inflated distributions, which constitutes a methodological step forward in the machine learning field. We used this modified RF to model tick bite risk using volunteered reports collected by two Dutch citizen science projects. We extend the current state of the art on tick bite risk modelling by devising and integrating a wide array of hazard and exposure metrics. By doing this, we are able to create tick bite risk maps for the Netherlands, and to explore the risk based on human exposure. We hope that this double contribution can help other researchers across multiple fields at modelling skewed and zero-inflated distributions using machine learning methods. Finally, we believe that this work also demonstrates that the incorporation of volunteered data to a scientific workflow is not only possible, but recommended to model fine-grained phenomena that escape classic monitoring networks.

Synthesis 6

6.1 Introduction

This PhD research revolves around the investigation and development of innovative data-driven methods to advance the spatio-temporal modelling of tick dynamics and tick bite risk. Advancing the modelling of these two phenomena is important to better understand where and when tick bites can occur, so that preventive campaigns could be implemented in these locations to reduce the number of new LB infections. We apply data-driven methods to model the non-linear relationships between tick bite risk (R), human exposure to ticks (E) and tick hazard existent in nature (H), by means of integrating volunteered and environmental data, and considering the spatial and temporal dimensions.

Understanding these relationships required an in-depth modelling of the R, E, and H components, to enable the creation of an integrated analytical framework capable of predicting tick dynamics and tick bite risk at the country level. With this research, I am engaged with specialists in the fields of public health, forestry, or ecology, creating maps that can help at designing tick-borne mitigation campaigns or increase public awareness towards LB. In this chapter, I reflect on the work presented by discussing how the pieces of this dissertation connect to each other (Section 6.2). I provide a reasoned answer to the research objectives posed in the introduction chapter (Section 6.3), and I highlight the main contributions of this thesis (Section 6.4). Finally, I outline ideas for prospective research lines (Section 6.5).

6.2 Connecting the dots

The development of this PhD has been possible due to the contribution of anonymous volunteers to two long-term citizen science projects. The RIVM and WUR have coordinated for over a decade the collection of tick bites, and the monthly tick activity monitored by trained volunteers. The availability of these two geodata collections, has enabled complex analyses at a fine spatio-temporal resolution in the fields of ecology or public health. It also illustrates that the application of machine learning methods can provide new insights on well-consolidated fields with a remarkable body of associated literature. In this thesis we incrementally presented our approach to model in space and time the three components that determine the number of LB infections in humans. We used the intrinsic relationship between risk, hazard, and exposure to devise new contributions at modelling these components. The remaining of this subsection explains the relationship between chapters. Note that, in chronological order, Chapter 3 was developed first, followed by Chapter 2, 4, and 5.

The development of Chapter 3 prompted questions about the nature of the volunteered tick bites collection. Back then, questions such as *"what is the component that the tick bites collection is measuring?"*, *"are the volunteered reports containing a substantial amount of spatial inaccuracy?"*, or *"how to validate volunteered data collections?"*, fueled discussions among several specialists of different fields. After the development of this research, these questions might seem trivial, but at the time were crucial to understand whether the tick bites collection was a by-product of R, E, or H. This was important, since the remaining of the thesis would build at the top of these decisions. The experiments and discussions that we conducted concluded that the tick bite reports are a realization of R, thus meaning that there was an interaction of the E and H components (unknown at the time) producing the measured tick bite risk signal.

Realizing that the volunteered tick bites was a proxy for R, prompted this research to seek and estimate the E and H components. In Chapter 2 we presented our approach to calculate the H component stemming from the volunteered tick activity monitored by trained volunteers. The analytical framework that we proposed in this chapter was designed not only to produce an estimation of the tick activity, but also to elucidate questions such as *"what are the drivers of tick dynamics at multiple time-scales?"*, or *"how can we enable time-awareness in the selected modelling method?"*. Both were challenging tasks, since a deep study of the literature showed that the scientific community does not have reached yet a consensus on the recommended variables to model tick dynamics, neither about the appropriate spatio-temporal resolution(s) to do so. In addition, it is yet unclear how to make machine learning algorithms to understand the temporal dimension, which is needed to obtain realistic tick dynamics. The development of this workflow created a model capable of predicting daily tick activity at the country level, which is a realization of H component.

After creating the H component, we realized that it was essential to obtain the E, so we could create a framework to integrally model R. However, prior to modelling something as heterogeneous and ubiquitous as human exposure to ticks, we decided to make a pre-step to assess (and double check) whether the E component could be derived from the newly calculated H and the R. In this way, it would be possible to acquire some prior knowledge about the E that would help at deriving meaningful variables afterwards. In Chapter 4 we propose a methodology to obtain a novel spatial E component, temporally static, which completes the demonstration of the intrinsic relationship between the three components. This research also helped at clarifying a question present since the beginning of the research: *"is R actually triggered by the H or by E?"*. The visual inspection of this novel E component revealed locations at the interface of natural and urban areas, as long as popular places for recreation, with high levels of human exposure, which suggest that the risk of tick bite is strongly influenced by human exposure on a particular day and location.

Knowing the strength and location of the human exposure component was key to obtain meaningful variables characterizing this heterogeneous and ubiquitous component. In Chapter 5 we used this newly acquired knowledge to devise good indicators of human exposure outdoors. These indicators were targeted at characterizing the urban-natural interface and popular recreational locations. After this step, we had the R, H, and E components, which enabled to train a data-driven model to predict the risk of tick bite. This model enables researchers and specialists to respond questions such as *"what are the top three most risky locations for tick bite in the country?"*, a naïve question that entangles a great complexity to be answered.

6.3 Answers to research questions

RQ1: How to develop a data-driven approach combining volunteered and environmental geodata, which is capable of capturing tick dynamics and assess the major drivers of tick activity across time-scales? In Chapter 2, we present our spatio-temporal data-driven approach to model tick dynamics by integrating volunteered data on tick activity with environmental variables. An extensive study on the literature revealed that there are three gaps that scientists have not studied sufficiently when modelling tick dynamics. First, it is unclear what the recommended set of environmental predictors is. Second, it is unknown at what time scales the different predictors operate. Third, the use of linear methods is insufficient to model such a complex ecological problem as tick dynamics. In our study, we overcome these limitations by modelling tick dynamics with wide array of environmental predictors aggregated at multiple temporal scales. We apply a well-known machine learning method (i.e. RF) which is capable of mapping complex non-linear interactions between the predictors and the target variable. In addition, we modified this method to enable it to predict tick activity considering the time dimension, and we used the framework described above to predict daily tick dynamics. The main findings of this paper reveal that atmospheric water levels (e.g. evapotranspiration, relative humidity) are major drivers of tick activity across all temporal scales, and we proposed a model capable of identifying daily tick activity at the country level. The estimation of tick dynamics is a way of proxying the hazard that ticks represent to humans. Thus, this piece of work is reused during the development of chapters 4 and 5.

RQ2: How to use data mining methods to identify spatio-temporal patterns linked to tick bites and verify that these patterns, stemming from volunteered observations, are intrinsic to the phenomenon under study? In Chapter 3, we started the exploration of the volunteered tick bites by means of carrying out a thorough exploratory data analysis. Since there is little knowledge on how tick bites are produced, we performed this study with a double goal in mind: 1) to assess what are the environmental and human factors that might be associated with a higher risk of tick bites; 2) to assess whether the volunteered tick bites collection contains any information which is intrinsic to the tick bites phenomenon. We identified an array of environmental, human, and weather variables that we modelled with a classic machine learning algorithm for frequent pattern mining (i.e. Apriori). This algorithm automatically explores the co-ocurrence of features causing a tick bite, thus it is possible to assess and map what are the most frequent conditions triggering them. We found several persistent patterns throughout the study period that reveals that the number of warm days (or alternatively, rainy days) and the distance to forests, do have an impact on the occurrence of tick bites, since they prompt (or prevent) humans to carry outdoor activities. At the time of this analysis (2015) the previous two results suggested a new hypothesis regarding the tick bites phenomenon: tick bites could be strongly driven by human recreational patterns outdoors in peri-urban areas. Regarding the quality of the volunteered tick bites collection, we crafted an original validation method to ensure that the volunteered reports contained information associated to the tick bites phenomenon and was not a product of random sampling. This procedure confirmed that the volunteered data contributed by citizens, was indeed capturing the tick bites phenomenon in space and time, subsequently implying that this dataset contains information with scientific value. The assessment of validity of the tick bites collection was important for the development of chapters 4 and 5.

RQ3: How to devise a novel indicator of human exposure to ticks, enabling the geographical identification of clusters of high exposure? In Chapter 4, we present a first-of-its-kind map of human exposure to tick bites, which stems from the combination of the tick dynamics work in Chapter 2 and the volunteered tick bites collected by the NK and TR citizen science projects. In Chapter 1 we explain that there is a close relationship between the R, H, and E components, which means that it should be possible to get a (rough) estimation of each of them as a linear combination of the remaining two. In this work, we used this principle to obtain a novel map product showing the human exposure to tick bites. This operation enables mapping the different types of locations in which humans are exposed to ticks and provide an estimation on the intensity of this exposure. The map reveals forest patches frequently visited by citizens, especially at the interface between urban and natural areas. In addition, this map shows locations outside forests in which tick bites are reported, and also forested locations with a very low recreational pressure. Our results indicate that the risk of tick bite is strongly influenced by the human exposure, more than to the existing hazard in a location. This suggests that hazard maps alone are insufficient to identify risky locations for LB infection, which motivates creating exposure maps for public health specialists and forest managers that could assist them at designing tick-borne mitigation campaigns. This work confirms the versatility and usefulness of citizen science projects at monitoring fine-grained and elusive public health threats, and encourages further research at calculating human exposure metrics to have a better overview on tick bite risk.

RQ4: How to integrate hazard and exposure metrics to devise a tick bite risk model, capable of handling the skewness and zero-inflation inherent to the volunteered tick bite reports? In Chapter 5, we take the knowledge acquired in the previous chapters to develop a data-driven tick bite risk model, integrating H and E metrics in a single analysis. We combine the tick dynamics model (H) developed in Chapter 2 and we derive a series of human exposure metrics (E) rooted in the findings of Chapter 3 and 4, to build a robust indicator of tick bite risk. We derived a series of accessibility and attractiveness metrics from publicly available geodata sources (official, user-contributed) and we modelled them using a well-known ensemble learning algorithm (i.e. RF). A major challenge in this work was learning from a highly skewed and zero-inflated distribution of the tick bites, since canonical machine learning algorithms have difficulties at modelling non-normal data. We solved this problem by modifying the chosen method and combining the built-in segmentation task with count data probability models (i.e. Poisson family), which are better suited to learn from imbalanced data. The contributions of this work are double fold: first we propose an integral analytical framework to model risk of tick bite based on H and E metrics, and second, we propose a methodological step forward by combining a machine learning-method with count data models, which enables an ensemble learning algorithm to model skewed and zero-inflated distributions.

6.4 Main contributions

In this thesis we propose guidelines to carry out an integral spatio-temporal assessment of tick bite risk considering human exposure and tick hazard factors. The individual and integrated modelling of these three components has produced four main **scientific contributions** and insights. First, we illustrate the intrinsic relationship between the R, E, and H components. This is important, since it suggests that other researchers in the field could derive a component from the other two and opens the way to re-use components to keep advancing modelling human-tick interactions. Second, we found that tick dynamics are consistently more influenced by water-related features (i.e. evapotranspiration, relative humidity) at all time scales, than to temperature or vegetation features. The latter types of features have been the preferred modelling choice in the past decades, but recent studies suggested that tick activity could be driven by atmospheric water levels. Our study supports these hypotheses at a fixed spatial scale and multiple temporal scales and recommends the inclusion of water-related features to model tick dynamics. Third, the study of frequent patterns reveals that favorable weather conditions for outdoor recreation are positively correlated with the number of tick bites per year and illustrates that tick bites tend to be geographically clustered at the interface between forest and urban areas. Both results suggest that residential exposure to ticks might play a significant role at triggering new tick bites. Fourth, we found out that tick bite risk is strongly influenced by human exposure, more than to the existing hazard in a location. This contribution led us to create a robust human exposure indicator that we used to create the last contribution, in which we provide a nationwide tick bite risk map that might help at fostering new research lines.

During this PhD journey, we implemented three **methodological contributions** that might serve a step-forward for new research in the fields of machine learning and environmental modelling. First, we bridge the inability of RF to produce time-aware predictions. We made a first attempt at applying a data-level modification of this method, in which we apply Z-scores to the tick activity count, and subsequently we proceed at modelling the data as it was initially planned. In this way, we provide 'context' to the model to be able to distinguish between different weather conditions (e.g. winter vs summer) yielding a similar tick activity count. In this way, we force RF to learn the mapping between the covariates and the Z-score'd tick activity, thus better 'understanding' the temporal dimension. Second, we solve RF's inability to learn from skewed and zero-inflated distributions. We propose an algorithm-level modification in which we combine the 'segmentation' capabilities of RF with count data models (e.g. Poisson family) which are better suited to model imbalanced distributions. Third, we developed a simple method to estimate human exposure to ticks, which overcomes the unavailability of ICT data (e.g. mobile phones locations) due to privacy laws, to create a nationwide coverage of human exposure to ticks.

We believe that this thesis also contains two **contributions to the citizen science community**. First, we would like to acknowledge the importance of VGI data at monitoring, with unprecedented levels of detail, elusive phenomena such as tick dynamics of tick bites. Our work has consistently demonstrated that citizen science projects can produce volunteered information ready to feed scientific workflows. Second, despite the challenges attributed to VGI described in Section 1.3, especially concerning data quality and representativeness, we think that geodata collections contributed by citizens are promising at modeling fine-grained and health-related phenomena. Thus, we have advanced at validating VGI data by devising creative methodologies and workflows checking whether the volunteered collections contain a signal that can be recovered. We consider this a practical but very important contribution, since there is no single manual on the proper treatment of VGI, thus each success case help researchers at validating VGI collections.

6.5 Prospective research lines and applications

In general, we think that possible future lines of research should (also) aim at quantifying uncertainties in the R, E, and H components, which may help reducing the error metrics reported during the development of this thesis. However, looking beyond error metrics, there are some interesting experiments ahead outlined in the remaining of this section.

6.5.1 Improving prediction of tick dynamics

In Section 1.2 we explained that LB infections are the outcome of a complex ecological system involving the interaction of several biotic (e.g. environment, wildlife) and abiotic factors (e.g. weather, landscape). The hazard model presented in Chapter 2 is able to capture general tick dynamics (e.g. large scale dynamics), but our predictions might be uncertain in locations in which weather conditions are not the main driver of tick activity. The missing piece of the LB ecological system that we were not able to include in the model, is the dynamics of different wildlife species. We believe that the inclusion of population dynamics on mammals (e.g. rodents, ungulates, foxes) and birds, could help at better predicting the peaks of tick activity present in data. During this research, we used the best possible collection of weather data for the Netherlands, provided by the Dutch Met Office (KNMI). These weather datasets are daily gridded layers at 1km resolution interpolated from the 34 automatic weather stations of the country. Although this resolution is reasonable for the type of experiments that we are doing, we suspect that this yet-coarse spatio-temporal resolution might not be capturing correctly local radiative or evaporative processes, which determine tick activity. We believe that the introduction of weather data at a finer spatio-temporal resolution (e.g. hourly, grid cells smaller than 1km) could help at pinpointing these local processes and improve the model metrics.

6.5.2 Creating a dynamic tick bite risk model

In Section 1.4 and Chapter 5 we introduced the concept of risk and we explained how to obtain risk maps from the volunteered reports. One of the limitations of this model is that is temporally static, which means that the risk of tick bite in a location does not change. Since most of the tick bites are reported during the period May-Aug accross years, we believe that this risk map is showing the 'maximum potential' tick bite risk in each location. However, a more advanced risk model should include some improvements, such as fluctuating between 'low risk' and 'high risk' state from winter to summer seasons, decrease (or increase) the level of risk if the weather conditions are harmful for the tick survival (e.g. too hot, too dry) or preventing humans to go outdoors (e.g. heavy rain, severe wind gust conditions). In addition, we recommend that, given the tick bite risk map, explore what are the main drivers of tick bites at each location (i.e. high hazard or high exposure), since this can help planning mitigation strategies.

6.5.3 Linking machine learning and statistics

In Chapter 5 we proposed an algorithm-level modification for RF, in which we combine its built-in segmentation capabilities with count data models. In 5.4 we illustrate how data-driven count data models are plugged to each tree estimator in the ensemble to learn from a highly skewed and zero-inflated distribution. The count data models of the Poisson family are defined by two parameters, π and λ, which correspond to the probability of ocurrence of an event and the rate at which this event occurs. The application of these count data models to each leaf node in the ensemble implies that after finishing the training phase of the ensemble, there is an array of pairs (π, λ) that are defining the shape of the tick bite risk distribution. We represented this array in a 2D space as Figure 6.1 illustrates. This figure shows the same analytical set up used in 5, but keeping the (π, λ) fitted to each leaf node. The point clouds in the figure suggest that there could be an overarching 2D Gaussian curve wrapping the points. This is interesting, since approximating the tick bite risk with this well-known model might pose some benefits regarding interpretability and fast computation of the risk. Further analysis could include the interpretation of these coefficients (e.g. assessing the value of parameters when risk is min/max), and carrying out a statistical analysis, including fitting the Gaussian curve, assessment of the uncertainty, and comparing the results with the current data-driven tick bite risk model.

6.5.4 Applications and other spatio-temporal analyses

- **Assesing and improving the quality of VGI:** As seen in Chapter 3, the tick bite reports contain an unknown factor of citizens' reporting errors (e.g. positional inaccuracies, over- or under-reporting tick bites) that we recommend tackling in further research. To overcome these issues further research could aim at making a full assessment of the data quality and representativeness of tick bites. Also, crossing the volunteered data with ICT data (e.g. mobile phones locations), or geolocated data streams from social networks (e.g. Twitter, Instagram) could be helpful at improving the quality of the reported positions, since tick bites could be repositioned to fall within the intended real-world feature.
- **Human exposure: residential vs. recreational:** In Chapter 4 we suggest that the exposure to tick bites is driven by two types of activities, namely "recreational exposure" and "residential exposure". The existing volunteered tick bites collection might be helpful to analyze the residential exposure in peri-urban areas. This would enable a high-resolution analysis of tick bite risk and find out locations (e.g. gardens, grasslands, small suburban forests) that might pose a high risk for urban dwellers.
- **Applications for professionals and citizens:** In the event that the models developed during this thesis could be operationalized, that is, running in a server and reading from a real-time tick bites and tick activity database, we believe that some web/smartphone applications

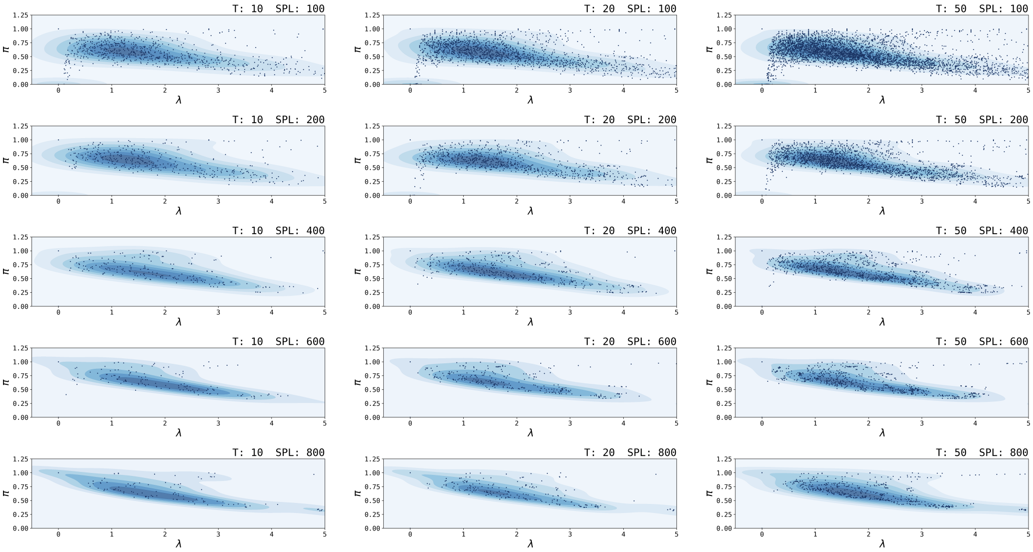

Figure 6.1: Visualization of π and λ coefficients for the selected ensemble. This figure shows the same analytical set up used in 5, in which we train ensembles for an increasing number of tress (T) and samples per leaf node (SPL). The X and Y axes show the fitted values of λ and π, respectively.

could be devised to deliver the R, E, and H maps. We envision that public health specialists and/or forest managers could be interested at receiving timely updates on R and E, since it can help planning tick mitigation strategies or design public health campaigns. Ecologists and biologists could be interested in studying the tick dynamics at the large-scale to locate regions of interest that enable a finer-scale analysis. Finally, citizens could be interested in receiving alerts in their phone in case they are entering a risky area for tick bites or the levels of tick hazard are high.

Summary

7

Human activities have induced global changes, which among other impacts are leading to the re-emergence of vector-borne diseases. Climate change, human and demographic developments, socio-economic exchanges, and the increase of human outdoor recreational activities, are some of the relevant factors that are contributing to the occurrence of vectors (e.g. mosquitoes, ticks) outside their endemic enclosements and facilitating the global transmission of vector-borne diseases between regions. The World Health Organization (WHO) has identified nine type of vectors that can cause, at least, 16 major vector-borne diseases in humans. Major tick-borne diseases comprise two bacterial infections (i.e. Lyme borreliosis, tick-borne encephalitis) and one viral infection (i.e. Crimea-Congo haemorrhagic fever), although there are several minor tick-borne diseases (e.g. rickettsial diseases, relapsing fever) with local importance.

Ticks are pervasive ectoparasites with a limited motility that require a wide array of biotic (e.g. environment, wildlife) and abiotic (e.g. weather, landscape structure) conditions ensuring their survival. Scientists have reported a latitudinal and altitudinal expansion of tick range in the last decades, which has been attributed to two related factors: global warming has turned unsuitable habitats into suitable ones and, subsequently, various wildlife species (e.g. rodents, birds, ungulates) have expanded their ranges and introduced ticks in new locations.

The expansion of the range of ticks is not the sole element prompting tick-borne diseases. Our planet is experiencing an increasing urbanization. Urban sprawl has increased the amount of residential areas in the periphery of cities, which are in closer contact with green spaces and nature. As a response, several bird and mammal species have adapted their ethology to live at the interface between forests and urban regions, where the chances of finding more food and less predators are higher. However, the proximity between nature and urban areas also means that pathogens and parasites carried by wildlife are getting closer to citizens. This phenomenon also explains the increasing hazard for tick-borne diseases. In addition, the progressive adoption of healthier lifestyles prompts citizens to carry more outdoor activities leading to a higher exposure to tick-borne diseases.

Lyme borreliosis (LB) is a tick-borne disease that has experienced a substantial spatio-temporal expansion in the Northern hemisphere in the last 20 years. Scientists and clinicians in nine European countries, USA, and Canada, have reported that the incidence of LB has steadily increased. In recent years, however, sub European sentinel networks of general practitioners have identified the first signs of stabilization. Yet, each year, roughly 25,000 Dutch citizens are diagnosed with LB. Albeit most of them respond well to the antibiotic treatment, there is a minority of patients reporting persisting symptoms that might lead to chronic symptoms and disability.

Ticks are the vehicle that pathogens utilize to infect new organisms, therefore, it is utterly important to monitor tick dynamics that enable the identification of hazardous locations for LB infection. The ubiquity of humans in natural spaces and the small size of ticks poses multiple challenges for public health organizations to monitor this disease. This is why two Dutch organizations started citizen science-based initiatives to collect data that can shed light on tick bite risk and tick dynamics.

During the period 2006-2012 the educational phenology platform *Natuurkalender*, linked to Wageningen University, gathered nearly 10,000 volunteered tick bites. This pioneering project attracted the attention of the Dutch National Institute for Public Health and the Environment (RIVM), and in 2012, the platform *Tekenradar* was launched by these two organizations. *Tekenradar* has collected over 50,000 volunteered tick bite reports in the Netherlands. To the best of our knowledge, these initiatives constitute the first citizen science projects that specifically focus on ticks and tick-borne diseases. Also in 2006, a group of scientists from Wageningen University started a country wide project to assess tick dynamics. A group of trained volunteers sampled a transect of forest on a monthly basis using a method called blanket dragging, counting the caught ticks every 25 meters. This project was carried out during the period 2006-2016 and created a unique collection of volunteered data.

These unique collections of volunteered observations on tick bites and tick dynamics enable the study of public health threats, such as LB. Volunteered data has the potential of monitoring complex and elusive environmental phenomena at an unprecedented spatio-temporal resolution. However, volunteered data is not exempt of several issues and challenges that require attention. Data quality and representativeness are some of the challenges that required attention before feeding these observations to scientific workflows.

In this PhD thesis, we focus on extracting spatial patterns and temporal trends from these volunteered data collections with two objectives in mind: modelling tick dynamics to identify the major drivers of tick populations, and modelling human activities to identify potentially risky factors and regions for LB. This disease is the product of a complex zoonotic cycle in which biotic and abiotic factors are weaved together. In our approach, we use machine learning methods to combine these data-sources about these factors and to account for the non-linearity of the interactions within the zoonotic cycle. In this way, we are able to create data-driven models capable of predicting daily tick activity and identifying risky locations for LB.

This journey towards the modelling of tick dynamics and tick bite risk has been structured in four steps. First, we present an approach to calculate tick dynamics stemming from the volunteered tick activity data. For this, we investigate the impact that long- and short-term variables have on tick activity, and we provide a model capable of predicting daily tick activity in forests at the national scale . Second, we introduce an extensive exploratory analysis to identify the most recurrent human and environmental patterns found in the tick bites dataset. This analysis enables the assessment of whether the tick bites collection is representative of the phenomenon under study, and helps elucidating whether this collection contains a clear tick bite signal. Third, we created a novel map of human exposure to tick bites in forested areas. We demonstrate that the risk of tick bite is strongly influenced by human behaviour, rather than by the particular tick dynamics in a given location. This map can also be used to identify locations where citizens are most exposed to ticks. Fourth, we develop a tick bite risk model that integrates information on tick dynamics (hazard) and human exposure to tick bites. This final analysis relies on the tick dynamic model developed previously, the knowledge acquired after the second and third analytical workflows, and on a series of human exposure indicators based on accessibility and landscape attractiveness. All these analytical workflows contribute at creating actionable geo-information products that can assist decision makers at designing tick bite prevention campaigns.

This PhD thesis has led to several scientific and methodological contributions that constitute a step-forward in the fields of spatio-temporal machine learning and environmental modelling. From a scientific point of view, our data-driven approaches showed that water-related features determine tick activity at multiple time scales, that tick bite risk is strongly influenced by human exposure rather than by tick activity, and that it is possible to localize and map the riskiest locations for tick bites, often located along the forest-urban interface. Regarding methodological contributions, we propose two modifications for a well-known machine learning method (i.e. Random Forest). These modifications extend the canonical functionalities of this method, enabling it to learn from skewed and zero-inflated distributions and timeseries data. To conclude, we believe that our work demonstrates that citizen science projects can produce valuable volunteered information that can feed scientific workflows to study elusive geographic phenomena, and that data-driven machine learning methods are paving the road for novel research lines on ticks and vector-borne diseases.

Samenvatting 8

Menselijke activiteiten hebben mondiale veranderingen veroorzaakt, waardoor onder andere ziekten die door vectoren worden overgedragen weer zijn toegenomen. Klimaatverandering, menselijke en demografische ontwikkelingen, socio-economische uitwisselingen en de toename van recreatieve activiteiten buitenshuis, zijn een aantal van de relevante factoren die bijdragen aan het vóórkomen van vectoren (bijvoorbeeld muggen en teken) buiten hun endemische gebieden en die de mondiale verspreiding van vectorziektes faciliteren. De wereldgezondheidsorganisatie (WHO) heeft negen typen vectoren geïdentificeerd die ten minste zestien heftige ziektes aan mensen kunnen overdragen. Voor ziektes overdraagbaar door teken vallen hieronder twee bacteriële infecties (de ziekte van Lyme en tekenencefalitis) en een virale infectie (Krim-Congovirus), met daarnaast verscheidene andere ziektes die nog niet wijdverspreid voorkomen maar lokaal wel van belang zijn (bv. ziektes veroorzaakt door rickettsiae, febris recurrens – terugkerende koorts).

Teken zijn algemeen voorkomende ectoparasieten met een beperkte mobiliteit die om te overleven een omgeving nodig hebben die moet voldoen aan een aantal biotische (waaronder milieu, voorkomen van diersoorten) en abiotische (o.a. weer, landschapssamenstelling) factoren. Wetenschappers hebben over het vóórkomen van teken in de afgelopen decennia een verspreiding naar meer geografische breedtes en een groter hoogtebereik gerapporteerd, wat gerelateerd wordt aan twee factoren: door mondiale opwarming zijn voorheen ongeschikte habitats geschikt worden en vervolgens zijn verschillende diersoorten (o.a. knaagdieren, vogels en hoefdieren) naar deze nieuwe locaties getrokken en hebben daar teken geïntroduceerd.

De uitbreiding van de plekken waar teken voorkomen is niet de enige reden waarom door teken overdraagbare ziektes in opmars zijn. Onze planeet ondergaat ook een toename in urbanisatie. Het bouwen van woningen aan de randen van steden leidt tot een intensiever contact met de groene ruimte en natuur. Een gevolg hiervan is dat verschillende vogel- en zoogdiersoorten hun leefpatronen hebben aangepast aan de bebouwde omgeving waar ze minder roofdieren tegenkomen en meer voedsel kunnen vinden. Dit mixen van natuur en bebouwde omgeving heeft echter ook tot gevolg dat pathogenen en parasieten die door wilde dieren worden verspreid dichterbij mensen komen. Dit fenomeen verklaart het toegenomen risico op ziektes overgedragen door teken. Daarnaast is door het promoten van een gezondere leefstijl het aantal activiteiten buitenshuis toegenomen wat ook leidt tot een hogere blootstelling aan teken en de ziektes die zij kunnen overdragen.

De ziekte van Lyme (LB) is een door teken overdraagbare ziekte die een enorme toename heeft laten zien op het noordelijk halfrond in de afgelopen twintig jaar. Wetenschappers en medici in negen Europese landen, de Verenigde Staten en Canada hebben gerapporteerd dat het vóórkomen van LB constant is toegenomen. In recente jaren wordt er echter een stabilisatie gezien door een groep van Europese huisartsen. Toch wordt elk jaar bij ongeveer 25 duizend Nederlanders de ziekte van Lyme vastgesteld. De meesten van hen reageren goed op een behandeling met antibiotica, maar een minderheid heeft last van blijvende symptomen die leiden tot chronische klachten en invaliditeit.

Teken zijn het middel waarmee pathogenen nieuwe organismen infecteren, en daarom is het zeer belangrijk om de tekendynamiek te monitoren om gevaarlijke plekken, waar het risico op LB infectie hoog is, te kunnen identificeren. Dat mensen veel in de natuur komen en het kleine formaat van teken maakt het zeer lastig voor volksgezondheidsorganisaties om de verspreiding van deze ziekte goed in kaart te kunnen brengen. Hierom hebben twee Nederlandse organisaties citizen science initiatieven gelanceerd zodat zij data kunnen verzamelen om de tekendynamiek en factoren die het risico op tekenbeten beïnvloeden beter te kunnen begrijpen.

In de periode 2006-2012 zijn bijna tienduizend tekenbeten gemeld door vrijwilligers aan het educatieve fenologienetwerk *Natuurkalender* (gelinkt aan Wageningen universiteit). Dit pionierende project trok de aandacht van het Rijksinstituut voor Volksgezondheid en het Milieu (RIVM) en in 2012 werd gezamenlijk de website *Tekenradar* gelanceerd. *Tekenradar* heeft meer dan vijftigduizend meldingen van vrijwilligers over tekenbeten in Nederland geregistreerd. Voor zover ons bekend is, waren deze initiatieven de eerste citizen science projecten die specifiek op teken en door teken overdraagbare ziektes gericht zijn. Tussen 2006 en 2016 heeft een groep wetenschappers van Wageningen universiteit ook een nationaal project uitgevoerd om te kijken naar de tekendynamiek. Hiervoor werd maandelijks een traject in een bos bemonsterd door getrainde vrijwilligers, waarbij elke 25 meter werd geteld hoeveel teken er aanwezig waren op het laken dat door de vrijwilligers getrokken werd over het traject.

Deze unieke verzamelingen van observaties door vrijwilligers met betrekking tot tekenbeten en tekendynamiek helpen ons om bedreigingen voor de volksgezondheid zoals LB te bestuderen. Het verzamelen van data door vrijwilligers heeft de potentie om complexe fenomenen op een ruimtelijke en tijdsschaal zonder precedent te kunnen bemonsteren. Deze metingen zijn echter niet geheel vrij van problemen en vragen om een zorgvuldige kwaliteitscontrole waarbij representativiteit van data ook moet worden meegenomen voordat deze observaties in wetenschappelijke analyses kunnen worden gebruikt.

In dit proefschrift hebben we de nadruk gelegd op het extraheren van ruimtelijke patronen en trends in de tijd gebaseerd op deze dataverzamelingen, waarbij er twee achterliggende objectieven zijn: het modelleren van tekendynamiek om de belangrijkste invloeden op tekenpopulaties te identificeren; en het modelleren van menselijke activiteiten om potentiële risicofactoren en regio's voor LB te onderkennen. Deze ziekte is het product van een complexe zoönotische cyclus waarin biotische en abiotische factoren met elkaar verbonden zijn. We hebben machine learning methodes gebruikt om deze databronnen over deze factoren te combineren en om te kunnen werken met de nonlineariteit van de interacties in de zoönotische cyclus. Zo hebben we modellen gebaseerd op de data gecreëerd die dagelijkse tekenactiviteit kunnen voorspellen samen met de locaties waar het risico op het oplopen van LB hoog is.

Om tot dit resultaat te kunnen komen, zijn vier stappen ondernomen. Ten eerste presenteren we een methodologie om de tekendynamiek te kunnen berekenen gebaseerd op de tekenactiviteit gerapporteerd door de vrijwilligers. Hiervoor hebben we gekeken naar het belang van variabelen met korte en lange termijn op de tekenactiviteit en daaruit een model afgeleid wat op de nationale schaal de dagelijkse tekenactiviteit voorspellen kan. Ten tweede hebben we een uitgebreide verkennende analyse uitgevoerd om de vaakst terugkerende menselijke en omgevingspatronen te vinden in de dataset over gerapporteerde tekenbeten. Deze analyse zorgt ervoor dat we kunnen beoordelen of de tekenbetenverzameling representatief is van het bestudeerde fenomeen en maakt duidelijk of er een duidelijk "tekenbeetsignaal" in de data zit. Ten derde hebben we een nieuwe kaart gemaakt van menselijke blootstelling aan teken(beten) in bosrijke omgevingen. Deze kaart laat zien dat het risico op een tekenbeet sterker door het gedrag van mensen wordt beïnvloed dan door de lokale tekendynamiek. Deze kaart kan ook worden gebruikt om locaties te identificeren waar burgers het meest worden blootgesteld aan teken. Daar kan bijvoorbeeld door volkgsgezondheidsorganisaties actief voorlichting over preventie van tekenbeten worden gegeven. Ten vierde hebben we een model ontwikkeld wat het risico op tekenbeten voorspelt waarvoor de informatie over tekendynamiek en menselijke blootstelling aan teken(beten) wordt geïntegreerd. Deze laatste analyse gebruikt het model over tekendynamiek dat eerder werd ontwikkeld, de kennis die in de tweede en derde analyses werd opgedaan, en een serie indicatoren van menselijke blootstelling gebaseerd op toegankelijkheid en aantrekkelijkheid van het landschap. Al deze analytische methodologieën leiden tot bruikbare geoinformatieproducten die beleidsmakers kunnen helpen om effectieve tekenbeetpreventiecampagnes op te zetten.

Dit proefschrift heeft geleid tot een aantal wetenschappelijke en methodologische bijdragen die een stap vooruit vormen in de velden ruimtelijke modellering en spatiotemporele machine learning. Vanuit een wetenschappelijk oogpunt hebben onze methodologieën die gedreven worden door data laten zien dat factoren gerelateerd aan water (zoals relatieve vochtigheid) tekenactiviteit bepalen op meerdere tijdschalen. Daarnaast hebben we gezien dat het risico op het oplopen van een tekenbeet sterker door menselijke blootstelling dan door tekenactiviteit bepaald wordt, en dat het mogelijk is om de meest riskante locaties voor het oplopen van een tekenbeet te identificeren en te karteren. Deze plekken zijn veelal de overgang tussen bos en bebouwing. Met betrekking tot de methodologische bijdragen hebben we twee aanpassingen voorgesteld voor een veelgebruikte machine learning methode (random forests). Deze aanpassingen breiden de standaardfuncties van deze methode uit, waardoor het kan leren van data met een scheve verdeling en statistische verdelingen met veel nullen, en van data uit tijdseries. We geloven dat ons werk laat zien dat citizen science projecten waardevolle informatie kan produceren die in wetenschappelijke analyses gebruikt kan worden om veelzijdige geografische fenomenen te bestuderen en dat machine learning methoden gedreven door data nieuwe onderzoeksrichtingen naar teken en vectorziektes mogelijk maken.

Bibliography

R. Agrawal and Srikant. Fast Algorithms for Mining Association Rules. *Proceedings of the 20th VLDB Conference*, 42:487–499, 1994. ISSN 1542-6270. doi: 10.1.1.40.7506. URL `http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.40.7506`.

B. F. Allan, F. Keesing, and R. S. Ostfeld. Effect of Forest Fragmentation on Lyme Disease Risk. *Conservation Biology*, 17(1):267–272, feb 2003. ISSN 0888-8892. doi: 10.1046/j.1523-1739.2003.01260.x. URL `http://doi.wiley.com/10.1046/j.1523-1739.2003.01260.x`.

E. Altpeter, H. Zimmermann, J. Oberreich, O. Péter, and C. Dvoøák. Tick related diseases in Switzerland, 2008 to 2011. *Swiss Medical Weekly*, 143 (January):1–13, 2013. ISSN 14247860. doi: 10.4414/smw.2013.13725.

J. Barrios, W. Verstraeten, P. Maes, J. Clement, J. Aerts, J. Farifteh, K. Lagrou, M. Van Ranst, and P. Coppin. Remotely sensed vegetation moisture as explanatory variable of Lyme borreliosis incidence. *International Journal of Applied Earth Observation and Geoinformation*, 18:1–12, aug 2012. ISSN 03032434. doi: 10.1016/j.jag.2012.01.023. URL `http://linkinghub.elsevier.com/retrieve/pii/S0303243412000256`.

J. M. Barrios González. *Spatio-temporal modelling of the epidemiology of Nephropathia Epidemica and Lyme Borreliosis*. PhD thesis, Leuven, Belgium, 2013.

L. Bennet, a. Halling, and J. Berglund. Increased incidence of Lyme borreliosis in southern Sweden following mild winters and during warm, humid summers. *European journal of clinical microbiology & infectious diseases : official publication of the European Society of Clinical Microbiology*, 25(7): 426–32, jul 2006. ISSN 0934-9723. doi: 10.1007/s10096-006-0167-2. URL `http://www.ncbi.nlm.nih.gov/pubmed/16810531`.

K. a. Berger, H. S. Ginsberg, K. D. Dugas, L. H. Hamel, and T. N. Mather. Adverse moisture events predict seasonal abundance of Lyme disease vector ticks (Ixodes scapularis). *Parasites & vectors*, 7: 181, jan 2014a. ISSN 1756-3305. doi: 10.1186/1756-3305-7-181. URL `http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3991885{&}tool=pmcentrez{&}rendertype=abstract`.

K. A. Berger, H. S. Ginsberg, L. Gonzalez, and T. N. Mather. Relative humidity and activity patterns of Ixodes scapularis (Acari: Ixodidae).

Journal of medical entomology, 51(4):769–76, 2014b. ISSN 0022-2585. URL http://www.ncbi.nlm.nih.gov/pubmed/25118408.

C. Bleyenheuft, T. Lernout, N. Berger, J. Rebolledo, M. Leroy, A. Robert, and S. Quoilin. Epidemiological situation of Lyme borreliosis in Belgium, 2003 to 2012. *Archives of Public Health*, 73(1):1–8, 2015. ISSN 20493258. doi: 10.1186/s13690-015-0079-7. URL http://dx.doi.org/10.1186/s13690-015-0079-7.

R. A. Boria, L. E. Olson, S. M. Goodman, and R. P. Anderson. Spatial filtering to reduce sampling bias can improve the performance of ecological niche models. *Ecological Modelling*, 275:73–77, 2014. ISSN 03043800. doi: 10.1016/j.ecolmodel.2013.12.012. URL http://dx.doi.org/10.1016/j.ecolmodel.2013.12.012.

M. Bouzid, F. J. Colón-González, T. Lung, I. R. Lake, and P. R. Hunter. Climate change and the emergence of vector-borne diseases in Europe: Case study of dengue fever. *BMC Public Health*, 14(1):1–12, 2014. ISSN 14712458. doi: 10.1186/1471-2458-14-781.

M. Braks, S. van Wieren, W. Takken, and H. Sprong. *Ecology and prevention of Lyme borreliosis*. Wageningen Academic Publishers, 2016. ISBN 978-90-8686-293-1. doi: http://dx.doi.org/10.3920/978-90-8686-838-4.

L. Breiman. Bagging predictors. *Machine Learning*, 24(2):123–140, 1996. ISSN 0885-6125. doi: 10.1007/BF00058655. URL http://link.springer.com/10.1007/BF00058655.

L. Breiman. Random Forests. *Machine Learning*, 45:5–32, 2001.

J. S. Brownstein, T. R. Holford, and D. Fish. A climate-based model predicts the spatial distribution of the Lyme disease vector Ixodes scapularis in the United States. *Environmental health perspectives*, 111(9):1152–7, 2003. ISSN 0091-6765. doi: 10.1289/ehp.6052. URL http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=1241567{&}tool=pmcentrez{&}rendertype=abstract.

J. S. Brownstein, D. K. Skelly, T. Holford, and D. Fish. Forest fragmentation predicts local scale heterogeneity of Lyme disease risk. 146(3):469–475, 2005. doi: 10.1007/s00442-005-0251-9.

J. P. Buonaccorsi, J. Elkinton, W. Koenig, R. P. Duncan, D. Kelly, and V. Sork. Measuring mast seeding behavior: relationships among population variation, individual variation and synchrony. *Journal of Theoretical Biology*, 224(1):107–114, 2003. doi: 10.1016/s0022-5193(03)00148-6.

W. Burgdorfer, A. Barbour, S. Hayes, J. Benach, E. Grunwaldt, and J. Davis. Lyme Disease: a tick-borne spirochetosis? *Science*, 18(216):1317–1319, 1982. doi: 10.1126/science.7043737.

J. A. Cardona-Ospina, F. A. Diaz-Quijano, and A. J. Rodríguez-Morales. Burden of chikungunya in Latin American countries: Estimates of disability-adjusted life-years (DALY) lost in the 2014 epidemic. *International Journal of Infectious Diseases*, 38:60–61, 2015. ISSN 18783511. doi: 10.1016/j.ijid.2015.07.015.

A. Chakravarti and R. Kumaria. Eco-epidemiological analysis of dengue infection during an outbreak of dengue fever, India. *Virology Journal*, 2: 1–7, 2005. ISSN 1743422X. doi: 10.1186/1743-422X-2-32.

L. Chapman, C. Bell, and S. Bell. Can the crowdsourcing data paradigm take atmospheric science to a new level? A case study of the urban heat island of London quantified using Netatmo weather stations. *International Journal of Climatology*, 37(9):3597–3605, 2017. ISSN 10970088. doi: 10.1002/joc.4940.

D. Cianci, N. Hartemink, and A. Ibáñez-Justicia. Modelling the potential spatial distribution of mosquito species using three different techniques. *International Journal of Health Geographics*, 14(1):10, Feb 2015. ISSN 1476-072X. doi: 10.1186/s12942-015-0001-0. URL `https://doi.org/10.1186/s12942-015-0001-0`.

K. M. Clow, P. A. Leighton, N. H. Ogden, L. R. Lindsay, P. Michel, D. L. Pearl, and C. M. Jardine. Northward range expansion of Ixodes scapularis evident over a short timescale in Ontario, Canada. *PLoS ONE*, 12(12):1–15, 2017. ISSN 19326203. doi: 10.1371/journal.pone.0189393.

M. Craglia and L. Shanley. Data democracy – increased supply of geospatial information and expanded participatory processes in the production of data. *International Journal of Digital Earth*, 8(9):679–693, 2015. ISSN 17538955. doi: 10.1080/17538947.2015.1008214. URL `http://dx.doi.org/10.1080/17538947.2015.1008214`.

L. E. M. Crommentuijn, J. M. J. Farjon, C. den Dekker, and N. van der Wulp. Belevingswaardenmonitor Nota Ruimte 2006: Nulmeting landschap en groen in en om de stad. 2007.

G. S. Cumming. Comparing Climate and Vegetation As Limiting Factors for Species Ranges of African Ticks. *Ecology*, 83(1):255–268, 2002. ISSN 0012-9658. doi: 10.1890/0012-9658(2002)083[0255:CCAVAL]2.0.CO;2. URL `http://www.esajournals.org/doi/abs/10.1890/0012-9658(2002)083[0255:CCAVAL]2.0.CO;2`.

R. Curtis-Robles, E. J. Wozniak, L. D. Auckland, G. L. Hamer, S. A. Hamer, G. Lawrence, R. Gorchakov, H. Alamgir, E. Dotson, B. Sissel, S. Sarkar, and K. O. Murray. Combining Public Health Education and Disease Ecology Research: Using Citizen Science to Assess Chagas Disease Entomological Risk in Texas. *Journal of Parasitology*, 9(12):520–528, 2015. ISSN 0022-3395. doi: 10.1371/journal.pntd.0004235. URL `http://www.bioone.org/doi/10.1645/15-748`.

G. J. Dammin. Erythema Migrans : A Chronicle. *Reviews of infectious diesease*, 11(1):142–151, 1989.

F. Dantas-Torres and D. Otranto. Species diversity and abundance of ticks in three habitats in southern Italy. *Ticks and tick-borne diseases*, 4(3):251–5, 2013. ISSN 1877-9603. doi: 10.1016/j.ttbdis.2012.11.004. URL `http://www.sciencedirect.com/science/article/pii/S1877959X12001409`.

M. de Groot. Personal protection for people with occupational risk in the Netherlands. volume 4, pages 389–407. 2016. ISBN 978-90-8686-293-1. doi:

10.3920/978-90-8686-838-4. URL http://www.wageningenacademic.com/doi/book/10.3920/978-90-8686-838-4.

M. De Keukeleire, S. O. Vanwambeke, E. Somassè, B. Kabamba, V. Luyasu, and A. Robert. Scouts, forests, and ticks: Impact of landscapes on human-tick contacts. *Ticks and Tick-borne Diseases*, 6(5):636–644, 2015. ISSN 18779603. doi: 10.1016/j.ttbdis.2015.05.008. URL http://dx.doi.org/10.1016/j.ttbdis.2015.05.008.

Y. M. Didyk, L. Blaňárová, S. Pogrebnyak, I. Akimov, B. Peťko, and B. Víchová. Emergence of tick-borne pathogens (Borrelia burgdorferi sensu lato, Anaplasma phagocytophilum, Ricketsia raoultii and Babesia microti) in the Kyiv urban parks, Ukraine. *Ticks and Tick-borne Diseases*, 8 (2):219–225, 2017. ISSN 18779603. doi: 10.1016/j.ttbdis.2016.10.002.

M. A. Diuk-Wasser, E. Vannier, and P. J. Krause. Coinfection by Ixodes Tick-Borne Pathogens: Ecological, Epidemiological, and Clinical Consequences. *Trends in Parasitology*, 32(1):30–42, 2016. ISSN 14715007. doi: 10.1016/j.pt.2015.09.008. URL http://dx.doi.org/10.1016/j.pt.2015.09.008.

E. E. A. EEA. *The impacts of urban sprawl*. Number 10. 2006. ISBN 1725-9177. doi: 10.1126/science.279.5352.860.

E. E. A. EEA. Landscape Fragmentation in Europe. Technical Report 2, 2011. URL http://www.ilpoe.uni-stuttgart.de/files/Landscape{_}Fragmentation{_}in{_}Europe.pdf.

S. Ehrmann, J. Liira, S. Gärtner, K. Hansen, J. Brunet, S. A. O. Cousins, M. Deconchat, G. Decocq, P. D. Frenne, P. D. Smedt, M. Diekmann, E. G. Moron, A. Kolb, J. Lenoir, J. Lindgren, T. Naaf, T. Paal, A. Valdés, K. Verheyen, M. Wulf, and M. S. Lorenzen. Environmental drivers of Ixodes ricinus abundance in forest fragments of rural European landscapes. *BMC Ecology*, pages 1–14, 2017. ISSN 1472-6785. doi: 10.1186/s12898-017-0141-0.

R. J. Eisen, L. Eisen, Y. A. Girard, N. Fedorova, J. Mun, B. Slikas, S. Leonhard, U. Kitron, and R. S. Lane. A spatially-explicit model of acarological risk of exposure to Borrelia burgdorferi-infected Ixodes pacificus nymphs in northwestern California based on woodland type, temperature, and water vapor. *Ticks Tick Borne Dis*, 1(1):35–43, 2010. doi: 10.1016/j.ttbdis.2009.12.002.A.

S. Elwood. Volunteered geographic information: future research directions motivated by critical, participatory, and feminist GIS. *GeoJournal*, 72(3-4): 173–183, jul 2008a. ISSN 0343-2521. doi: 10.1007/s10708-008-9186-0. URL http://link.springer.com/10.1007/s10708-008-9186-0.

S. Elwood. Geographic Information Science: new geovisualization technologies – emerging questions and linkages with GIScience research. *Progress in Human Geography*, 33(2):256–263, aug 2008b. ISSN 0309-1325. doi: 10.1177/0309132508094076. URL http://phg.sagepub.com/cgi/doi/10.1177/0309132508094076.

A. Estrada-Peña. Distribution, Abundance, and Habitat Preferences of Ixodes ricinus (Acari: Ixodidae) in Northern Spain. *Journal of Medical Entomology*, 38(3):361–370, 2001. doi: 10.1603/0022-2585-38.3.361.

A. Estrada-Peña and J. de la Fuente. Species interactions in occurrence data for a community of tick-transmitted pathogens. *Scientific Data*, 3, 2016. doi: 10.1038/sdata.2016.56.

A. Estrada-Peña, C. Ortega, N. Sánchez, L. Desimone, B. Sudre, J. E. Suk, and J. C. Semenza. Correlation of Borrelia burgdorferi sensu lato prevalence in questing Ixodes ricinus ticks with specific abiotic traits in the western palearctic. *Applied and environmental microbiology*, 77(11): 3838–45, jun 2011. ISSN 1098-5336. doi: 10.1128/AEM.00067-11. URL `http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3127598{&}tool=pmcentrez{&}rendertype=abstract`.

A. Estrada-Peña, N. Ayllón, and J. de la Fuente. Impact of climate trends on tick-borne pathogen transmission. *Frontiers in physiology*, 3(March): 64, jan 2012. ISSN 1664-042X. doi: 10.3389/fphys.2012.00064. URL `http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3313475{&}tool=pmcentrez{&}rendertype=abstract`.

A. Estrada-Peña, J. S. Gray, O. Kahl, R. S. Lane, and A. M. Nijhof. Research on the ecology of ticks and tick-borne pathogens–methodological principles and caveats. *Frontiers in cellular and infection microbiology*, 3 (August):29, 2013. ISSN 2235-2988. doi: 10.3389/fcimb.2013.00029. URL `http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3737478{&}tool=pmcentrez{&}rendertype=abstract`.

A. Estrada-Peña, J. de la Fuente, T. Latapia, and C. Ortega. The Impact of Climate Trends on a Tick Affecting Public Health: A Retrospective Modeling Approach for Hyalomma marginatum (Ixodidae). *PloS one*, 10(5):e0125760, 2015. ISSN 1932-6203. doi: 10.1371/journal.pone.0125760. URL `http://journals.plos.org/plosone/article?id=10.1371/journal.pone.0125760`.

R. C. Falco and D. Fish. Potential for Exposure to Tick Bites in Recreational Parks in a Lyme Disease Endemic Area. 79(1), 1989. ISSN 00900036. doi: 10.2105/AJPH.79.1.12.

C. A. Fiebrich. History of surface weather observations in the United States. *Earth-Science Reviews*, 93(3-4):77–84, 2009. ISSN 00128252. doi: 10.1016/j.earscirev.2009.01.001. URL `http://dx.doi.org/10.1016/j.earscirev.2009.01.001`.

D. Fink, W. M. Hochachka, B. Zuckerberg, David W. Winkler, Ben Shaby, M. Arthur Munson, Giles Hooker, Mirek Riedewald, Daniel Sheldon, and S. Kelling. Spatiotemporal exploratory models for broad-scale survey data. *Journal of Ecological Applications*, 20(8):2131–2147, 2010.

D. Fink, Theodoros Damoulas, and J. Dave. Adaptive Spatio-Temporal Exploratory Models: Hemisphere-wide species distributions from massively crowdsourced eBird data, 2013.

P. Fournier-Viger, A. Gomariz, T. Gueniche, A. Soltani, C.-W. Wu, and V. S. Tseng. SPMF : A Java Open-Source Pattern Mining Library. *Journal of Machine Learning Research*, 15:3389–3393, 2014.

I. Garcia-Martí, R. Zurita-Milla, A. Swart, C. C. van den Wijngaard, A. J. H. van Vliet, S. Bennema, and M. Harms. Identifying environmental and human factors associated with tick bites using volunteered reports and frequent pattern mining. *Transactions in GIS*, 2016. doi: 10.13140/RG.2.1.3702.8566.

I. Garcia-Martí, R. Zurita-Milla, A. van Vliet, and W. Takken. Modelling and mapping tick dynamics using volunteered observations. *International Journal of Health Geographics*, 16(1), 2017. ISSN 1476072X. doi: 10.1186/s12942-017-0114-8.

I. Garcia-Marti, R. Zurita-Milla, M. G. Harms, and A. Swart. Using volunteered observations to map human exposure to ticks. *Scientific Reports*, 8 (1):15435, 2018. ISSN 2045-2322. doi: 10.1038/s41598-018-33900-2. URL `http://www.nature.com/articles/s41598-018-33900-2`.

I. Garcia-Marti, R. Zurita-Milla, and A. Swart. Modelling tick bite risk by combining random forests and count data regression models. *bioRxiv*, 2019. doi: 10.1101/642728. URL `https://www.biorxiv.org/content/early/2019/05/24/642728`.

F. Gassner, A. J. H. van Vliet, S. L. G. E. Burgers, F. Jacobs, P. Verbaarschot, E. K. E. Hovius, S. Mulder, N. O. Verhulst, L. S. van Overbeek, and W. Takken. Geographic and temporal variations in population dynamics of Ixodes ricinus and associated Borrelia infections in The Netherlands. *Vector borne and zoonotic diseases (Larchmont, N.Y.)*, 11(5): 523–32, may 2011. ISSN 1557-7759. doi: 10.1089/vbz.2010.0026. URL `http://www.ncbi.nlm.nih.gov/pubmed/21083369`.

F. Gassner, K. M. Hansford, and J. Medlock. Greener cities, a wild card for ticks? In M. A. Braks, S. E. van Wieren, W. Takken, and H. Sprong, editors, *Ecology and prevention of Lyme borreliosis*, volume 4, chapter 13, pages 187–203. Wageningen Academic Publishers, 2016. ISBN 9789086868384. doi: 10.3920/978-90-8686-838-4.

GDAL Development Team. GDAL Geospatial Data Abstraction Library: Version 2.1.0, Open Source Geospatial Foundation, 2018. URL `http://gdal.osgeo.org/`.

M. Gold, L. Robinson, and F. Sanz. *Citizen Science for environmental policy: development of an EU-wide inventory and analysis of selected practices*. 2018. ISBN 9789279981043.

M. F. Goodchild. Editorial : Citizens as Voluntary Sensors : Spatial Data Infrastructure in the World of Web 2 . 0. 2:24–32, 2007a.

M. F. Goodchild. Citizens as sensors: the world of volunteered geography. *GeoJournal*, 69(4):211–221, nov 2007b. ISSN 0343-2521. doi: 10.1007/s10708-007-9111-y. URL `http://link.springer.com/10.1007/s10708-007-9111-y`.

M. F. Goodchild and L. Li. Assuring the quality of volunteered geographic information. *Spatial Statistics*, 1:110–120, may 2012. ISSN 22116753. doi: 10.1016/j.spasta.2012.03.002. URL `http://linkinghub.elsevier.com/retrieve/pii/S2211675312000097`.

W. H. Greene. Accounting for Excess Zeros and Sample Selection in Poisson and Negative Binomial Regression Models. *Biology & Philosophy*, 9(3):265–265, 1994. ISSN 0169-3867. doi: 10.1007/BF00857937. URL http://papers.ssrn.com/sol3/papers.cfm?abstract{_}id=1293115{%}5Cnhttp://www.springerlink.com/index/10.1007/BF00857937.

D. J. Gubler. Resurgent Vector-Borne Diseases as a Global Health Problem. *Emerging Infectious Diseases*, 4(3), 1998. ISSN 00428450.

D. J. Gubler, N. Vasilakis, and D. Musso. History and Emergence of Zika Virus. *Journal of Infectious Diseases*, 216(January):S860–S867, 2017. ISSN 15376613. doi: 10.1093/infdis/jix451.

M. G. Guzman, D. J. Gubler, A. Izquierdo, E. Martinez, and S. B. Halstead. Dengue infection. *Nature Reviews Disease Primer*, 2:301–311, 2016. ISSN 2056676X. doi: 10.1007/978-1-4614-4496-1_10. URL http://dx.doi.org/10.1038/nrdp.2016.55.

F. Hadiji, A. Molina, S. Natarajan, and K. Kersting. Poisson Dependency Networks: Gradient Boosted Models for Multivariate Count Data. *Machine Learning*, 100(2-3):477–507, 2015. ISSN 15730565. doi: 10.1007/s10994-015-5506-z.

M. B. Hahn, J. K. Bjork, D. F. Neitzel, F. M. Dorr, T. Whitemarsh, K. A. Boegler, C. B. Graham, T. L. Johnson, S. E. Maes, and R. J. Eisen. Evaluating acarological risk for exposure to Ixodes scapularis and Ixodes scapularis-borne pathogens in recreational and residential settings in Washington County, Minnesota. *Ticks and Tick-borne Diseases*, (November):0–1, 2017. ISSN 18779603. doi: 10.1016/j.ttbdis.2017.11.010. URL http://dx.doi.org/10.1016/j.ttbdis.2017.11.010.

A. Haines, R. S. Kovats, D. Campbell-Lendrum, and C. Corvalan. Climate change and human health: impacts, vulnerability and public health. *Public health*, 120(7):585–96, jul 2006. ISSN 0033-3506. doi: 10.1016/j.puhe.2006.01.002. URL http://www.ncbi.nlm.nih.gov/pubmed/16542689.

M. Haklay. How good is volunteered geographical information? A comparative study of OpenStreetMap and Ordnance Survey datasets. *Environment and Planning B: Planning and Design*, 37(4):682–703, 2010. ISSN 0265-8135. doi: 10.1068/b35097. URL http://www.envplan.com/abstract.cgi?id=b35097.

M. Haklay, S. Basiouka, V. Antoniou, and A. Ather. How Many Volunteers Does It Take To Map An Area Well? The validity of Linus' law to Volunteered Geographic Information. pages 1–13, 2009.

J. L. Hall, K. Alpers, K. J. Bown, S. J. Martin, and R. J. Birtles. Use of Mass-Participation Outdoor Events to Assess Human Exposure to Tickborne Pathogens. 23(3):463–467, 2017.

K. M. Hansford, M. Fonville, E. L. Gillingham, E. Claudia, M. E. Pietzsch, A. I. Krawczyk, A. G. C. Vaux, B. Cull, H. Sprong, and J. M. Medlock. Ticks and Borrelia in urban and peri-urban green space habitats in a city in southern England. *Ticks and Tick-borne Diseases*, 8(3):353–361, 2017. ISSN 1877-959X. doi: 10.1016/j.ttbdis.2016.12.009.

N. Hartemink and W. Takken. Trends in tick population dynamics and pathogen transmission in emerging tick-borne pathogens in europe: an introduction. *Experimental and Applied Acarology*, 68(3):269–278, Mar 2016. ISSN 1572-9702. doi: 10.1007/s10493-015-0003-4. URL `https://doi.org/10.1007/s10493-015-0003-4`.

N. Hartemink, S. O. Vanwambeke, B. V. Purse, M. Gilbert, and H. Van Dyck. Towards a resource-based habitat approach for spatial modelling of vector-borne disease risks. *Biological Reviews*, 90(4):1151–1162, 2015. doi: 10.1111/brv.12149. URL `https://onlinelibrary.wiley.com/doi/abs/10.1111/brv.12149`.

B. Haworth. Emergency management perspectives on volunteered geographic information: Opportunities, challenges and change. *Computers, Environment and Urban Systems*, 57:189–198, 2016. ISSN 01989715. doi: 10.1016/j.compenvurbsys.2016.02.009. URL `http://dx.doi.org/10.1016/j.compenvurbsys.2016.02.009`.

H. He and E. A. Garcia. Learning from imbalanced data. *IEEE Transactions on knowledge and data engineering*, 21(9):1263–1284, 2009. ISSN 15680266. doi: 10.2174/156802608786786589. URL `http://www.eurekaselect.com/openurl/content.php?genre=article{&}issn=1568-0266{&}volume=8{&}issue=18{&}spage=1691`.

G. M. Hengeveld, E. Schüll, R. Trubins, and O. Sallnäs. Forest Policy and Economics Forest Landscape Development Scenarios (FoLDS) – A framework for integrating forest models , owners ′ behaviour and socio-economic developments. *Forest Policy and Economics*, 2017. ISSN 1389-9341. doi: 10.1016/j.forpol.2017.03.007.

D. Heylen, E. Tijsse, M. Fonville, E. Matthysen, and H. Sprong. Transmission dynamics of Borrelia burgdorferi s.l. in a bird tick community. *Environmental microbiology*, 15(2):663–73, feb 2013. ISSN 1462-2920. doi: 10.1111/1462-2920.12059. URL `http://www.ncbi.nlm.nih.gov/pubmed/23279105`.

P. Heyman, C. Cochez, A. Hofhuis, J. van der Giessen, H. Sprong, S. R. Porter, B. Losson, C. Saegerman, O. Donoso-Mantke, M. Niedrig, and A. Papa. A clear and present danger: tick-borne diseases in Europe. *Expert review of anti-infective therapy*, 8(1):33–50, jan 2010. ISSN 1744-8336. doi: 10.1586/eri.09.118. URL `http://www.ncbi.nlm.nih.gov/pubmed/20014900`.

T. K. Ho. The random subspace method for constructing decision forests. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 20(8):832–844, 1998. ISSN 01628828. doi: 10.1109/34.709601.

A. Hofhuis, M. Harms, S. Bennema, C. C. Van Den Wijngaard, and W. Van Pelt. Physician reported incidence of early and late Lyme borreliosis. *Parasites and Vectors*, 8(1):1–8, 2015a. ISSN 17563305. doi: 10.1186/s13071-015-0777-6161.

A. Hofhuis, M. Harms, C. van den Wijngaard, H. Sprong, and W. van Pelt. Continuing increase of tick bites and Lyme disease between 1994 and

2009. *Ticks and Tick-borne Diseases*, 6:69–74, 2015b. ISSN 1877959X. doi: 10.1016/j.ttbdis.2014.09.006. URL http://linkinghub.elsevier.com/retrieve/pii/S1877959X14001903.

A. Hofhuis, S. Bennema, M. Harms, A. J. Van Vliet, W. Takken, C. C. Van Den Wijngaard, and W. Van Pelt. Decrease in tick bite consultations and stabilization of early Lyme borreliosis in the Netherlands in 2014 after 15 years of continuous increase. *BMC Public Health*, 16(1):1–6, 2016. ISSN 14712458. doi: 10.1186/s12889-016-3105-y. URL http://dx.doi.org/10.1186/s12889-016-3105-y.

Hofhuis A, van der Giessen JW, Borgsteede F, Wielinga PR, Notermans DW, and van Pelt W. Lyme borreliosis in the Netherlands: strong increase in GP consultations and hospital admissions in past 10 years. *Eurosurveillance*, 11 (25):1–5, 2006. URL http://library.wur.nl/WebQuery/wurpubs/fulltext/157590.

M. A. Hoogstra-Klein. Exploring the financial rationales of Dutch forest holdings and their relation with financial results. *European Journal of Forest Research*, 2016. ISSN 1612-4677. doi: 10.1007/s10342-016-0991-6.

Z. Hubálek, J. Halouzka, Z. Juricová, S. Sikutová, and I. Rudolf. Effect of forest clearing on the abundance of Ixodes ricinus ticks and the prevalence of Borrelia burgdorferi s.l. *Medical and veterinary entomology*, 20(2):166–72, jun 2006. ISSN 0269-283X. doi: 10.1111/j.1365-2915.2006.00615.x. URL http://www.ncbi.nlm.nih.gov/pubmed/16796612.

J. D. Hunter. Matplotlib: A 2D graphics environment, 2007. ISSN 15219615. URL https://matplotlib.org/index.html.

T. G. Jaenson, D. G. Jaenson, L. Eisen, E. Petersson, and E. Lindgren. Changes in the geographical distribution and abundance of the tick Ixodes ricinus during the past 30 years in Sweden. *Parasites and Vectors*, 5(1):1–15, 2012. ISSN 17563305. doi: 10.1186/1756-3305-5-8.

T. G. T. Jaenson, L. Eisen, P. Comstedt, H. a. Mejlon, E. Lindgren, S. Bergström, and B. Olsen. Risk indicators for the tick Ixodes ricinus and Borrelia burgdorferi sensu lato in Sweden. *Medical and veterinary entomology*, 23(3):226–37, sep 2009. ISSN 1365-2915. doi: 10.1111/j.1365-2915.2009.00813.x. URL http://www.ncbi.nlm.nih.gov/pubmed/19712153.

N. Japkowicz and S. Stephen. The class imbalance problem: a systematic study. *Intelligent Data Analysis*, pages 426–449, 2002. ISSN 11720360. doi: 10.2165/00042310-199812070-00003.

G. F. Jenks. The Data Model Concept in Statistical Mapping. *International yearbook of Cartography*, 7:186–190, 1967.

C. G. Jones, R. S. Ostfeld, M. P. Richard, and J. O. Wolff. Mast seeding and. 9 (12):5347, 1998.

S. Jore, H. Viljugrein, M. Hofshagen, H. Brun-Hansen, A. B. Kristoffersen, K. Nygård, E. Brun, P. Ottesen, B. K. Sævik, and B. Ytrehus. Multi-source analysis reveals latitudinal and altitudinal shifts in range of Ixodes ricinus at its northern distribution limit. *Parasites and Vectors*, 4(1):1–11, 2011. ISSN 17563305. doi: 10.1186/1756-3305-4-84.

S. Jore, S. O. Vanwambeke, H. Viljugrein, K. Isaksen, A. B. Kristoffersen, Z. Woldehiwet, B. Johansen, E. Brun, H. Brun-Hansen, S. Westermann, I.-L. Larsen, B. Ytrehus, and M. Hofshagen. Climate and environmental change drives Ixodes ricinus geographical expansion at the northern range margin. *Parasites & vectors*, 7: 11, jan 2014. ISSN 1756-3305. doi: 10.1186/1756-3305-7-11. URL `http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3895670{&}tool=pmcentrez{&}rendertype=abstract`.

A. K. Kala, C. Tiwari, A. R. Mikler, and S. F. Atkinson. A comparison of least squares regression and geographically weighted regression modeling of west nile virus risk based on environmental parameters. *PeerJ*, 5:e3070, Mar. 2017. ISSN 2167-8359. doi: 10.7717/peerj.3070. URL `https://doi.org/10.7717/peerj.3070`.

M. N. Kamel Boulos. On geography and medical journalology: a study of the geographical distribution of articles published in a leading medical informatics journal between 1999 and 2004. *International journal of health geographics*, 4(1):7, mar 2005. ISSN 1476-072X. doi: 10.1186/1476-072X-4-7. URL `http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=1079922{&}tool=pmcentrez{&}rendertype=abstract`.

M. N. Kamel Boulos, B. Resch, D. N. Crowley, J. G. Breslin, G. Sohn, R. Burtner, W. A. Pike, E. Jezierski, and K. Y. S. Chuang. Crowdsourcing, citizen sensing and sensor web technologies for public and environmental health surveillance and crisis management: Trends, OGC standards and application examples. *International Journal of Health Geographics*, 10, 2011. ISSN 1476072X. doi: 10.1186/1476-072X-10-67.

P. Kanaroglou, E. Delmelle, and A. Paez, editors. *Spatial Analysis in Health Geography*. Taylor and Francis Group, London, 2015. ISBN 9781472416193. doi: 10.4324/9781315610252.

D. Kelly, W. D. Koenig, and A. M. Liebhold. An intercontinental comparison of the dynamic behavior of mast seeding communities. *Population Ecology*, 50(4):329–342, 2008. doi: 10.1007/s10144-008-0114-4.

M. D. Keukeleire, A. Robert, B. Kabamba, E. Dion, V. Luyasu, and S. O. Vanwambeke. Individual and environmental factors associated with the seroprevalence of Borrelia burgdorferi in Belgian farmers and veterinarians. *Infection Ecology & Epidemiology*, 6:1–10, 2016. doi: 10.3402/iee.v6.32793.

D. Kiewra, E. Stefańska-Krzaczek, M. Szymanowski, and A. Szczepańska. Local-scale spatio-temporal distribution of questing Ixodes ricinus L. (Acari: Ixodidae)-A case study from a riparian urban forest in Wrocław, SW Poland. *Ticks and Tick-borne Diseases*, 8(3):362–369, 2017. ISSN 18779603. doi: 10.1016/j.ttbdis.2016.12.011.

M. Kosmala, A. Wiggins, A. Swanson, and B. Simmons. Assessing data quality in citizen science. *Frontiers in Ecology and the Environment*, 14(10): 551–560, 2016. ISSN 15409309. doi: 10.1002/fee.1436.

M. Kowalec, T. Szewczyk, R. Welc-Falciak, E. Siński, G. Karbowiak, and A. Bajer. Ticks and the city - Are there any differences between city parks and natural forests in terms of tick abundance and prevalence of spirochaetes? *Parasites and Vectors*, 10(1):1–19, 2017. ISSN 17563305. doi: 10.1186/s13071-017-2391-2.

B. Krawczyk. Learning from imbalanced data: open challenges and future directions. *Progress in Artificial Intelligence*, 5(4):221–232, 2016. ISSN 2192-6352. doi: 10.1007/s13748-016-0094-0. URL `http://link.springer.com/10.1007/s13748-016-0094-0`.

K. Krishnamoorthy, K. Harichandrakumar, A. Kumari, and L. Das. Burden of Chikungunya in India: Estimates of disability adjusted life years (DALY) lost in 2006 epidemic. *Journal of Vector Borne Diseases*, 46(1): 26–35, 2009. ISSN 0972-9062. doi: 10.1109/PICMET.2009.5262001. URL `http://www.embase.com/search/results?subaction=viewrecord{&}from=export{&}id=L354519848{%}5Cnhttp://www.mrcindia.org/journal/issues/461026.pdf`.

C. Kullenberg and D. Kasperowski. What is citizen science? - A scientometric meta-analysis. *PLoS ONE*, 11(1):1–16, 2016. ISSN 19326203. doi: 10.1371/journal.pone.0147152.

F. A. La Sorte, D. Fink, P. J. Blancher, A. D. Rodewald, V. Ruiz-Gutierrez, K. V. Rosenberg, W. M. Hochachka, P. H. Verburg, and S. Kelling. Global change and the distributional dynamics of migratory bird populations wintering in Central America. *Global Change Biology*, 23(12):5284–5296, 2017. ISSN 13652486. doi: 10.1111/gcb.13794.

D. Lambert. Zero-inflated poisson regression, with an application to defects in manufacturing. *Technometrics*, 34(1):1–14, 1992. ISSN 15372723. doi: 10.1080/00401706.1992.10485228.

E. F. Lambin, A. Tran, S. O. Vanwambeke, C. Linard, and V. Soti. Pathogenic landscapes: interactions between land, people, disease vectors, and their animal hosts. *International journal of health geographics*, 9(1): 54, jan 2010. ISSN 1476-072X. doi: 10.1186/1476-072X-9-54. URL `http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2984574{&}tool=pmcentrez{&}rendertype=abstract`.

J. M. Last. *A Dictionary for Epidemiology*. Oxford University Press, 2001. ISBN 9780195141696.

B. Y. Lee, K. M. Bacon, M. E. Bottazzi, and P. J. Hotez. Global economic burden of Chagas disease: A computational simulation model. *The Lancet Infectious Diseases*, 13(4):342–348, 2013. ISSN 14733099. doi: 10.1016/S1473-3099(13)70002-1. URL `http://dx.doi.org/10.1016/S1473-3099(13)70002-1`.

Lemon, Stanley M., Sparling, Frederick P., Hamburg, Margaret A., Relman, David A., E. R. Choffnes, and A. Mack. Vector-Borne Diseases: Understanding the environmental, human health and ecological connections. In *Workshop Summary*, Washington, 2008.

S. Li, N. Hartemink, N. Speybroeck, and S. O. Vanwambeke. Consequences of landscape fragmentation on Lyme disease risk: a cellular automata approach. *PloS one*, 7(6):e39612, jan 2012.

ISSN 1932-6203. doi: 10.1371/journal.pone.0039612. URL http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3382467{&}tool=pmcentrez{&}rendertype=abstract.

S. Li, V. Colson, P. Lejeune, N. Speybroeck, and S. O. Vanwambeke. Agent-based modelling of the spatial pattern of leisure visitation in forests: A case study in wallonia, south belgium. *Environmental Modelling and Software*, 71:111 – 125, 2015. ISSN 1364-8152. doi: https://doi.org/10.1016/j.envsoft.2015.06.001. URL http://www.sciencedirect.com/science/article/pii/S1364815215001632.

S. Li, V. Colson, P. Lejeune, and S. O. Vanwambeke. On the distance travelled for woodland leisure via different transport modes in Wallonia, south Belgium. *Urban Forestry and Urban Greening*, 15:123–132, 2016. ISSN 16108167. doi: 10.1016/j.ufug.2015.12.007. URL http://dx.doi.org/10.1016/j.ufug.2015.12.007.

S. Li, L. Juhász-Horváth, A. Trájer, L. Pintér, M. D. Rounsevell, and P. A. Harrison. Lifestyle, habitat and farmers' risk of exposure to tick bites in an endemic area of tick-borne diseases in Hungary. *Zoonoses and Public Health*, 65(1):e248–e253, 2018. ISSN 18632378. doi: 10.1111/zph.12413.

C. Linard, P. Lamarque, P. Heyman, G. Ducoffre, V. Luyasu, K. Tersago, S. O. Vanwambeke, and E. F. Lambin. Determinants of the geographic distribution of Puumala virus and Lyme borreliosis infections in Belgium. *International journal of health geographics*, 6:15, jan 2007. ISSN 1476-072X. doi: 10.1186/1476-072X-6-15. URL http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=1867807{&}tool=pmcentrez{&}rendertype=abstract.

E. Lindgren and T. G. T. Jaenson. Lyme borreliosis in Europe : influences of climate and climate change , epidemiology , ecology and adaptation measures By :. Technical report, 2006.

K. LoGiudice, R. S. Ostfeld, K. a. Schmidt, and F. Keesing. The ecology of infectious disease: effects of host diversity and community composition on Lyme disease risk. *Proceedings of the National Academy of Sciences of the United States of America*, 100(2):567–71, jan 2003. ISSN 0027-8424. doi: 10.1073/pnas.0233733100. URL http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=141036{&}tool=pmcentrez{&}rendertype=abstract.

G. Louppe, L. Wehenkel, A. Sutera, and P. Geurts. Understanding variable importances in forests of randomized trees. *Neural Information Processing Systems*, pages 1–9, 2013. ISSN 1098-6596. doi: NIPS2013_4928.

M. Madder and L. Baeten. The abundance of Ixodes ricinus ticks depends on tree species composition and shrub cover. (April), 2012. doi: 10.1017/S0031182012000625.

L. A. Magnarelli, A. Denicola, K. C. Stafford, and J. F. Anderson. Borrelia burgdorferi in an urban environment: White-tailed deer with infected ticks and antibodies. *Journal of Clinical Microbiology*, 33(3):541–544, 1995. ISSN 00951137.

T. N. Mather, M. C. Nicholson, E. F. Donnelly, and B. T. Matyas. Entomologic index for human risk of Lyme disease. *American Journal of Epidemiology*, 144(11):1066–1069, 1996. ISSN 00029262. doi: 10.1093/oxfordjournals.aje.a008879.

G. J. McCabe and J. E. Bunnell. Precipitation and the Occurrence of Lyme Disease in the Northeastern United States. *Vector borne and zoonotic diseases*, 4(2):1–7, 2011. doi: 10.1089/ast.2006.0095.

J. M. Medlock, 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, and W. Van Bortel. Driving forces for changes in geographical distribution of Ixodes ricinus ticks in Europe. *Parasites & vectors*, 6:1, jan 2013. ISSN 1756-3305. doi: 10.1186/1756-3305-6-1. URL `http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3549795{&}tool=pmcentrez{&}rendertype=abstract`.

H. Mehdipoor, R. Zurita-Milla, A. Rosemartin, K. L. Gerst, and J. F. Weltzin. Developing a workflow to identify inconsistencies in volunteered geographic information: A phenological case study. *PLOS ONE*, 10(10):1–14, 10 2015. doi: 10.1371/journal.pone.0140811. URL `https://doi.org/10.1371/journal.pone.0140811`.

H. Mehdipoor, R. Zurita-Milla, E.-W. Augustijn, and A. J. H. van Vliet. Checking the Consistency of Volunteered Phenological Observations While Analysing Their Synchrony. *International Journal of Geo-Information*, 7(487):22, 2018a. doi: 10.3390/ijgi7120487.

H. Mehdipoor, R. Zurita-Milla, E.-W. Augustijn, and A. J. H. Van Vliet. Checking the consistency of volunteered phenological observations while analysing their synchrony. *ISPRS International Journal of Geo-Information*, 7(12), 2018b. ISSN 2220-9964. doi: 10.3390/ijgi7120487. URL `https://www.mdpi.com/2220-9964/7/12/487`.

Met Office UK. Cartopy: a cartographic python library with a matplotlib interface, 2010. URL `http://scitools.org.uk/cartopy`.

L. Mokraoui, N. M. Noor, and A. Abdullah. Developing dengue index through the integration of crowdsourcing approach (X-Waba). *IOP Conference Series: Earth and Environmental Science*, 169(1), 2018. ISSN 17551315. doi: 10.1088/1755-1315/169/1/012058.

P. Mooney, A.-M. Olteanu-Raimond, G. Touya, J. Niels, S. Alvanides, and N. Kerle. Considerations of Privacy, Ethics and Legal Issues in Volunteered Geographic Information. *Mapping and the Citizen Sensor*, pages 119–135, 2017. doi: https://doi.org/10.5334/bbf.f.License. URL `http://www.oapen.org/download?type=document{&}docid=637890{#}page=128`.

S. Mulder, A. J. H. van Vliet, W. A. Bron, F. Gassner, and W. Takken. High risk of tick bites in Dutch gardens. *Vector Borne Zoonotic Dis*, 13(12):865–871, 2013. doi: 10.1089/vbz.2012.1194. URL `http://www.ncbi.nlm.nih.gov/pubmed/24107214`.

C. J. Murray, T. Vos, R. Lozano, M. Naghavi, A. D. Flaxman, C. Michaud, M. Ezzati, K. Shibuya, J. A. Salomon, S. Abdalla, V. Aboyans, J. Abraham, I. Ackerman, R. Aggarwal, S. Y. Ahn, M. K. Ali, M. A. AlMazroa, M. Alvarado, H. R. Anderson, L. M. Anderson, K. G. Andrews, C. Atkinson, L. M. Baddour, A. N. Bahalim, S. Barker-Collo, L. H. Barrero, D. H. Bartels, M. G. Basáñez, A. Baxter, M. L. Bell, E. J. Benjamin, D. Bennett, E. Bernabé, K. Bhalla, B. Bhandari, B. Bikbov, A. Bin Abdulhak, G. Birbeck, J. A. Black, H. Blencowe, J. D. Blore, F. Blyth, I. Bolliger, A. Bonaventure, S. Boufous, R. Bourne, M. Boussinesq, T. Braithwaite, C. Brayne, L. Bridgett, S. Brooker, P. Brooks, T. S. Brugha, C. Bryan-Hancock, C. Bucello, R. Buchbinder, G. Buckle, C. M. Budke, M. Burch, P. Burney, R. Burstein, B. Calabria, B. Campbell, C. E. Canter, H. Carabin, J. Carapetis, L. Carmona, C. Cella, F. Charlson, H. Chen, A. T. A. Cheng, D. Chou, S. S. Chugh, L. E. Coffeng, S. D. Colan, S. Colquhoun, K. E. Colson, J. Condon, M. D. Connor, L. T. Cooper, M. Corriere, M. Cortinovis, K. Courville De Vaccaro, W. Couser, B. C. Cowie, M. H. Criqui, M. Cross, K. C. Dabhadkar, M. Dahiya, N. Dahodwala, J. Damsere-Derry, G. Danaei, A. Davis, D. De Leo, L. Degenhardt, R. Dellavalle, A. Delossantos, J. Denenberg, S. Derrett, D. C. Des Jarlais, S. D. Dharmaratne, M. Dherani, C. Diaz-Torne, H. Dolk, E. R. Dorsey, T. Driscoll, H. Duber, B. Ebel, K. Edmond, A. Elbaz, S. Eltahir Ali, H. Erskine, P. J. Erwin, P. Espindola, S. E. Ewoigbokhan, F. Farzadfar, V. Feigin, D. T. Felson, A. Ferrari, C. P. Ferri, E. M. Fèvre, M. M. Finucane, S. Flaxman, L. Flood, K. Foreman, M. H. Forouzanfar, F. G. R. Fowkes, M. Fransen, M. K. Freeman, B. J. Gabbe, S. E. Gabriel, E. Gakidou, H. A. Ganatra, B. Garcia, F. Gaspari, R. F. Gillum, G. Gmel, D. Gonzalez-Medina, R. Gosselin, R. Grainger, B. Grant, J. Groeger, F. Guillemin, D. Gunnell, R. Gupta, J. Haagsma, H. Hagan, Y. A. Halasa, W. Hall, D. Haring, J. M. Haro, J. E. Harrison, R. Havmoeller, R. J. Hay, H. Higashi, C. Hill, B. Hoen, H. Hoffman, P. J. Hotez, D. Hoy, J. J. Huang, S. E. Ibeanusi, K. H. Jacobsen, S. L. James, D. Jarvis, R. Jasrasaria, S. Jayaraman, N. Johns, J. B. Jonas, G. Karthikeyan, N. Kassebaum, N. Kawakami, A. Keren, J. P. Khoo, C. H. King, L. M. Knowlton, O. Kobusingye, A. Koranteng, R. Krishnamurthi, F. Laden, R. Lalloo, L. L. Laslett, T. Lathlean, J. L. Leasher, Y. Y. Lee, J. Leigh, D. Levinson, S. S. Lim, E. Limb, J. K. Lin, M. Lipnick, S. E. Lipshultz, W. Liu, M. Loane, S. Lockett Ohno, R. Lyons, J. Mabweijano, M. F. MacIntyre, R. Malekzadeh, L. Mallinger, S. Manivannan, W. Marcenes, L. March, D. J. Margolis, G. B. Marks, R. Marks, A. Matsumori, R. Matzopoulos, B. M. Mayosi, J. H. McAnulty, M. M. McDermott, N. McGill, J. McGrath, M. E. Medina-Mora, M. Meltzer, Z. A. Memish, G. A. Mensah, T. R. Merriman, A. C. Meyer, V. Miglioli, M. Miller, T. R. Miller, P. B. Mitchell, C. Mock, A. O. Mocumbi, T. E. Moffitt, A. A. Mokdad, L. Monasta, M. Montico, M. Moradi-Lakeh, A. Moran, L. Morawska, R. Mori, M. E. Murdoch, M. K. Mwaniki, K. Naidoo, M. N. Nair, L. Naldi, K. M. Narayan, P. K. Nelson, R. G. Nelson, M. C. Nevitt, C. R. Newton, S. Nolte, P. Norman, R. Norman, M. O'Donnell, S. O'Hanlon, C. Olives, S. B. Omer, K. Ortblad, R. Osborne, D. Ozgediz, A. Page, B. Pahari, J. D. Pandian, A. Panozo Rivero, S. B. Patten, N. Pearce, R. Perez Padilla, F. Perez-Ruiz, N. Perico, K. Pesudovs, D. Phillips, M. R. Phillips, K. Pierce, S. Pion, G. V. Polanczyk,

S. Polinder, C. A. Pope, S. Popova, E. Porrini, F. Pourmalek, M. Prince, R. L. Pullan, K. D. Ramaiah, D. Ranganathan, H. Razavi, M. Regan, J. T. Rehm, D. B. Rein, G. Remuzzi, K. Richardson, F. P. Rivara, T. Roberts, C. Robinson, F. Rodriguez De Leòn, L. Ronfani, R. Room, L. C. Rosenfeld, L. Rushton, R. L. Sacco, S. Saha, U. Sampson, L. Sanchez-Riera, E. Sanman, D. C. Schwebel, J. G. Scott, M. Segui-Gomez, S. Shahraz, D. S. Shepard, H. Shin, R. Shivakoti, D. Silberberg, D. Singh, G. M. Singh, J. A. Singh, J. Singleton, D. A. Sleet, K. Sliwa, E. Smith, J. L. Smith, N. J. Stapelberg, A. Steer, T. Steiner, W. A. Stolk, L. J. Stovner, C. Sudfeld, S. Syed, G. Tamburlini, M. Tavakkoli, H. R. Taylor, J. A. Taylor, W. J. Taylor, B. Thomas, W. M. Thomson, G. D. Thurston, I. M. Tleyjeh, M. Tonelli, J. A. Towbin, T. Truelsen, M. K. Tsilimbaris, C. Ubeda, E. A. Undurraga, M. J. Van Der Werf, J. Van Os, M. S. Vavilala, N. Venketasubramanian, M. Wang, W. Wang, K. Watt, D. J. Weatherall, M. A. Weinstock, R. Weintraub, M. G. Weisskopf, M. M. Weissman, R. A. White, H. Whiteford, N. Wiebe, S. T. Wiersma, J. D. Wilkinson, H. C. Williams, S. R. Williams, E. Witt, F. Wolfe, A. D. Woolf, S. Wulf, P. H. Yeh, A. K. Zaidi, Z. J. Zheng, D. Zonies, and A. D. Lopez. Disability-adjusted life years (DALYs) for 291 diseases and injuries in 21 regions, 1990-2010: A systematic analysis for the Global Burden of Disease Study 2010. *The Lancet*, 380(9859):2197–2223, 2012. ISSN 1474547X. doi: 10.1016/S0140-6736(12)61689-4.

F. W. Murray. On the computation of saturation vapor pressure. *Journal of Applied Meteorology*, 6(1):203–204, 1967. doi: 10.1175/1520-0450(1967)006<0203:OTCOSV>2.0.CO;2. URL https://doi.org/10.1175/1520-0450(1967)006<0203:OTCOSV>2.0.CO;2.

D. Musso, V. M. Cao-Lormeau, and D. J. Gubler. Zika virus: following the path of dengue and chikungunya? *The Lancet*, 386 (9990):243–244, 2015. ISSN 1474547X. doi: 10.1016/S0140-6736(15)61273-9. URL http://linkinghub.elsevier.com/retrieve/pii/S0140673615612739.

NASEM. Global health impacts of Vector-borne diseases: workshop summary. In *Global Health Impacts of Vector-Borne Diseases: Workshop Summary*, pages 221–257, 2016. ISBN 9780309377591. doi: 10.17226/21792. URL https://www.ncbi.nlm.nih.gov/books/NBK390439/.

A. B. Nielsen, E. Heyman, and G. Richnau. Liked, disliked and unseen forest attributes: Relation to modes of viewing and cognitive constructs. *Journal of Environmental Management*, 113:456–466, 2012. ISSN 03014797. doi: 10.1016/j.jenvman.2012.10.014. URL http://dx.doi.org/10.1016/j.jenvman.2012.10.014.

N. C. Nieto, W. T. Porter, J. C. Wachara, T. J. Lowrey, L. Martin, P. J. Motyka, and D. Salkeld. Using citizen science to describe the prevalence and distribution of tick bite and exposure to tick-borne diseases in the United States. *PloS one*, 13(7), 2018. ISSN 18173195. doi: 10.1371/journal.pone.0199644.

C. P. Oechslin, D. Heutschi, N. Lenz, W. Tischhauser, O. Péter, O. Rais, C. M. Beuret, S. L. Leib, S. Bankoul, and R. Ackermann-Gäumann. Prevalence of tick-borne pathogens in questing Ixodes ricinus ticks in urban and suburban areas of Switzerland. *Parasites and Vectors*, 10(1):1–18, 2017. ISSN 17563305. doi: 10.1186/s13071-017-2500-2.

N. H. Ogden, a. Maarouf, I. K. Barker, M. Bigras-Poulin, L. R. Lindsay, M. G. Morshed, C. J. O'callaghan, F. Ramay, D. Waltner-Toews, and D. F. Charron. Climate change and the potential for range expansion of the Lyme disease vector Ixodes scapularis in Canada. *International journal for parasitology*, 36(1):63–70, jan 2006. ISSN 0020-7519. doi: 10.1016/j.ijpara.2005.08.016. URL `http://www.ncbi.nlm.nih.gov/pubmed/16229849`.

N. H. Ogden, J. K. Koffi, Y. Pelcat, and L. R. Lindsay. Lyme disease Surveillance Environmental risk from Lyme disease in central and eastern Canada : a summary of recent surveillance information. *Canada Communicable Disease Report*, 40(5), 2014.

O. Okwa, F. I. Akinmolayan, V. Carter, and H. Hurd. Original Article T Ransmission D Ynamics of M Alaria in F Our S Elected E Cological Z Ones of N Igeria in the R Ainy S Eason. *Annals of African Medicine*, 8(1):1–9, 2009.

T. E. Oliphant. Guide to NumPy, 2006. URL `http://www.numpy.org/`.

T. E. Oliphant. Python for scientific computing. *Computing in Science and Engineering*, pages 10–20, 2007. ISSN 1521-9615. doi: 10.1109/MCSE.2007.58.

E. Olivieri, A. L. Gazzonis, S. A. Zanzani, F. Veronesi, and M. T. Manfredi. Seasonal dynamics of adult Dermacentor reticulatus in a peri-urban park in southern Europe. *Ticks and Tick-borne Diseases*, 8(5):772–779, 2017. ISSN 18779603. doi: 10.1016/j.ttbdis.2017.06.002. URL `http://dx.doi.org/10.1016/j.ttbdis.2017.06.002`.

T. O'Reilly. What is Web 2.0: Design patterns and business models for the next generation of software. *MPRA*, 65, 2007. ISSN 00219606. doi: 10.1063/1.3153845.

R. Ostfeld. *Lyme Disease: the ecology of a complex system*. Oxford University Press, New York, New York, USA, 1st editio edition, 2012. ISBN 978-0199928477.

R. Ostfeld, C. Canham, K. Oggenfuss, R. Winchcombe, and F. Keesing. Climate, deer, rodents, and acorns as determinants of variation in lyme-disease risk. *PLoS biology*, 4(6):e145, jun 2006. ISSN 1545-7885. doi: 10.1371/journal.pbio.0040145. URL `http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=1457019{&}tool=pmcentrez{&}rendertype=abstract`.

E. Ozdenerol. Gis and remote sensing use in the exploration of lyme disease epidemiology. *International Journal of Environmental Research and Public Health*, 12(12):15182–15203, 2015. ISSN 1660-4601. doi: 10.3390/ijerph121214971. URL `http://www.mdpi.com/1660-4601/12/12/14971`.

K. A. Padgett and D. L. Bonilla. Novel exposure sites for nymphal Ixodes pacificus within picnic areas. *Ticks and Tick-borne Diseases*, 2(4):191–195,

2011. ISSN 1877959X. doi: 10.1016/j.ttbdis.2011.07.002. URL http://dx.doi.org/10.1016/j.ttbdis.2011.07.002.

N. Pasquier, Y. Bastide, R. Taouil, and L. Lakhal. Discovering Frequent Closed Itemsets for Association Rules. *Icdt*, pages 398–416, 1999. doi: 10.1007/3-540-49257-7_25.

R. E. Paul, M. Cote, E. Le Naour, and S. I. Bonnet. Environmental factors influencing tick densities over seven years in a French suburban forest. *Parasites and Vectors*, 9(1):1–10, 2016. ISSN 17563305. doi: 10.1186/s13071-016-1591-5. URL http://dx.doi.org/10.1186/s13071-016-1591-5.

M. Perry. Jenks Natural Breaks, 2014. URL https://github.com/perrygeo/jenks.

S. E. Randolph. Ticks and Tick-borne Disease Systems in Space and from Space. *Advances in parasitology*, 47, 2000.

S. E. Randolph. Is expert opinion enough? a critical assessment of the evidence for potential impacts of climate change on tick-borne diseases. *Animal Health Research Reviews*, 14(2):133–137, 2013. doi: 10.1017/S1466252313000091.

S. E. Randolph and K. Storey. Impact of Microclimate on Immature Tick-Rodent Host Interactions (Acari : Ixodidae): Implications for Parasite Transmission. pages 741–748, 1999.

S. E. Randolph, L. Asokliene, T. Avsic-Zupanc, A. Bormane, C. Burri, L. Gern, I. Golovljova, Z. Hubalek, N. Knap, M. Kondrusik, A. Kupca, M. Pejcoch, V. Vasilenko, and M. Zygutiene. Variable spikes in tick-borne encephalitis incidence in 2006 independent of variable tick abundance but related to weather. *Parasites & vectors*, 1(1): 44, jan 2008. ISSN 1756-3305. doi: 10.1186/1756-3305-1-44. URL http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2614985{&}tool=pmcentrez{&}rendertype=abstract.

V. F. Rodriguez-Galiano, M. Chica-Olmo, and M. Chica-Rivas. Predictive modelling of gold potential with the integration of multisource information based on random forest: a case study on the Rodalquilar area, Southern Spain. *International Journal of Geographical Information Science*, 28(7): 1336–1354, 2014. ISSN 13623087. doi: 10.1080/13658816.2014.885527.

J. Roos-Klein Lankhorst, S. de Vries, A. E. Buijs, M. H. I. Bloemmen, and C. Schuiling. BelevingsGIS versie 2: waardering van het Nederlandse landschap door de bevolking op kaart. 2005.

A. H. Rosemartin, E. G. Denny, J. F. Weltzin, R. Lee Marsh, B. E. Wilson, H. Mehdipoor, R. Zurita-Milla, and M. D. Schwartz. Lilac and honeysuckle phenology data 1956–2014. *Scientific Data*, 2:150038, 2015. ISSN 2052-4463. doi: 10.1038/sdata.2015.38. URL http://www.nature.com/articles/sdata201538.

F. Ruiz-Fons, I. G. Fernández-de Mera, P. Acevedo, C. Gortázar, and J. de la Fuente. Factors driving the abundance of ixodes ricinus ticks and the prevalence of zoonotic i. ricinus-borne pathogens in natural foci. *Applied and environmental microbiology*, 78:8, 2012. doi: 10.1128/AEM.06564-11. URL `https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3318823/`.

E. L. Rulison, I. Kuczaj, G. Pang, G. J. Hickling, J. I. Tsao, and H. S. Ginsberg. Flagging versus dragging as sampling methods for nymphal ixodes scapularis (acari: Ixodidae). *Journal of Vector Ecology*, 38(1):163–167, 2013. doi: 10.1111/j.1948-7134.2013.12022.x. URL `https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1948-7134.2013.12022.x`.

P. A. Sandifer, A. E. Sutton-grier, and B. P. Ward. Exploring connections among nature , biodiversity , ecosystem services , and human health and well-being : Opportunities to enhance health and biodiversity conservation $. *Ecosystem Services*, 12:1–15, 2015. ISSN 2212-0416. doi: 10.1016/j.ecoser.2014.12.007.

A. S. Santos, A. de Bruin, A. R. Veloso, C. Marques, I. Pereira da Fonseca, R. de Sousa, H. Sprong, and M. M. Santos-Silva. Detection of Anaplasma phagocytophilum, Candidatus Neoehrlichia sp., Coxiella burnetii and Rickettsia spp. in questing ticks from a recreational park, Portugal. *Ticks and Tick-borne Diseases*, 9(6):1555–1564, 2018. doi: 10.1016/j.ttbdis.2018.07.010.

J. Sawyer. Man-made Carbon Dioxide and the “Greenhouse” Effect. *Nature new biology*, 239:23–26, 1972. ISSN 0028-0836. doi: 10.1038/239137a0.

M. J. Schelhaas, W. P. Daamen, J. F. Oldenburger, G. Velema, and P. Schnitger. Zesde Nederlandse Bosinventarisatie : methoden en basisresultaten. 2014.

A. M. Schwartz, A. F. Hinckley, P. S. Mead, S. A. Hook, and K. J. Kugeler. Surveillance for Lyme Disease — United States, 2008–2015. Technical Report 22, 2017. URL `http://www.cdc.gov/mmwr/volumes/66/ss/ss6622a1.htm`.

S. Seabold and J. Perktold. Statsmodels: Econometric and Statistical Modeling with Python. *PROC. OF THE 9th PYTHON IN SCIENCE CONF*, (Scipy):57, 2010. ISSN 0276-0460. doi: 10.1007/s00367-011-0258-7. URL `http://statsmodels.sourceforge.net/`.

L. See, S. Fritz, E. Dias, E. Hendriks, B. Mijling, F. Snik, P. Stammes, F. D. Vescovi, G. Zeug, P. P. Mathieu, Y. L. Desnos, and M. Rast. Supporting earth-observation calibration and validation: A new generation of tools for crowdsourcing and citizen science. *IEEE Geoscience and Remote Sensing Magazine*, 4(3):38–50, 2016a. ISSN 21686831. doi: 10.1109/MGRS.2015.2498840.

L. See, P. Mooney, G. Foody, L. Bastin, A. Comber, J. Estima, S. Fritz, N. Kerle, B. Jiang, M. Laakso, H.-Y. Liu, G. Milčinski, M. Nikšič, M. Painho, A. Pődör, A.-M. Olteanu-Raimond, and M. Rutzinger. Crowdsourcing, Citizen Science or Volunteered Geographic Information? The Current State of Crowdsourced Geographic Information. *ISPRS International Journal of Geo-Information*, 5(5):55, 2016b. ISSN 2220-9964. doi: 10.3390/ijgi5050055. URL `http://www.mdpi.com/2220-9964/5/5/55`.

H. Senaratne, A. Mobasheri, A. L. Ali, C. Capineri, and M. M. Haklay. A review of volunteered geographic information quality assessment methods. *International Journal of Geographical Information Science*, 31(1):139–167, 2017. ISSN 13623087. doi: 10.1080/13658816.2016.1189556.

D. S. Shepard, E. A. Undurraga, and Y. A. Halasa. Economic and Disease Burden of Dengue in Southeast Asia. *PLOS Neglected Tropical Diseases*, 7 (2), 2013. ISSN 0035001X. doi: 10.1371/journal.pntd.0002055.

H. Sprong, A. Hofhuis, F. Gassner, W. Takken, F. Jacobs, A. J. H. van Vliet, M. van Ballegooijen, J. van der Giessen, and K. Takumi. Circumstantial evidence for an increase in the total number and activity of Borrelia-infected Ixodes ricinus in the Netherlands. *Parasites & vectors*, 5(1): 294, jan 2012. ISSN 1756-3305. doi: 10.1186/1756-3305-5-294. URL `http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3562265{&}tool=pmcentrez{&}rendertype=abstract`.

J. D. Stanaway and G. Roth. The Burden of Chagas Disease Estimates and Challenges. *Global Heart*, 10(3):139–144, 2015. ISSN 22118179. doi: 10.1016/j.gheart.2015.06.001. URL `http://dx.doi.org/10.1016/j.gheart.2015.06.001`.

P. Stefanoff, B. Rubikowska, J. Bratkowski, Z. Ustrnul, S. O. Vanwambeke, and M. Rosinska. A predictive model has identified tick-borne encephalitis high-risk areas in regions where no caseswere reported previously, Poland, 1999-2012. *International Journal of Environmental Research and Public Health*, 15(4):1–17, 2018. ISSN 16604601. doi: 10.3390/ijerph15040677.

S. Su, C. Lei, A. Li, J. Pi, and Z. Cai. Coverage inequality and quality of volunteered geographic features in Chinese cities: Analyzing the associated local characteristics using geographically weighted regression. *Applied Geography*, 78:78–93, 2017. ISSN 01436228. doi: 10.1016/j.apgeog.2016.11.002. URL `http://dx.doi.org/10.1016/j.apgeog.2016.11.002`.

S. Subak. Effects of climate on variability in Lyme disease incidence in the northeastern United States. *American Journal of Epidemiology*, 157(6): 531–538, 2003. ISSN 00029262. doi: 10.1093/aje/kwg014.

B. L. Sullivan, C. L. Wood, M. J. Iliff, R. E. Bonney, D. Fink, and S. Kelling. eBird: A citizen-based bird observation network in the biological sciences. *Biological Conservation*, 142(10):2282–2292, 2009. ISSN 00063207. doi: 10.1016/j.biocon.2009.05.006. URL `http://dx.doi.org/10.1016/j.biocon.2009.05.006`.

A. Swart, A. Ibañez-Justicia, J. Buijs, S. E. van Wieren, T. R. Hofmeester, H. Sprong, and K. Takumi. Predicting Tick Presence by Environmental Risk Mapping. *Frontiers in Public Health*, 2(November): 1–8, nov 2014. ISSN 2296-2565. doi: 10.3389/fpubh.2014.00238. URL `http://www.frontiersin.org/Epidemiology/10.3389/fpubh.2014.00238/abstract`.

S. Szekeres, E. C. Coipan, K. Rigó, G. Majoros, S. Jahfari, H. Sprong, and G. Földvári. Eco-epidemiology of Borrelia miyamotoi and Lyme borreliosis spirochetes in a popular hunting and recreational forest area in Hungary. *Parasites and Vectors*, 8(1):1–8, 2015. ISSN 17563305. doi: 10.1186/s13071-015-0922-2. URL http://dx.doi.org/10.1186/s13071-015-0922-2.

S. Szekeres, A. Docters van Leeuwen, K. Rigó, M. Jablonszky, G. Majoros, H. Sprong, and G. Földvári. Prevalence and diversity of human pathogenic rickettsiae in urban versus rural habitats, Hungary. *Experimental and Applied Acarology*, 68(2):223–226, 2016. ISSN 15729702. doi: 10.1007/s10493-015-9989-x.

S. Szekeres, A. Docters van Leeuwen, E. Tóth, G. Majoros, H. Sprong, and G. Földvári. Road-killed mammals provide insight into tick-borne bacterial pathogen communities within urban habitats. *Transboundary and Emerging Diseases*, (October), 2018. ISSN 18651682. doi: 10.1111/tbed.13019.

W. Tack. *Impact of Forest Conversion on the Abundance of Ixodes Ricinus Ticks*. PhD thesis, Belgium, 2013.

W. Tack, M. Madder, L. Baeten, P. De Frenne, and K. Verheyen. The abundance of Ixodes ricinus ticks depends on tree species composition and shrub cover. *Parasitology*, 139(10):1273–81, sep 2012. ISSN 1469-8161. doi: 10.1017/S0031182012000625. URL http://www.ncbi.nlm.nih.gov/pubmed/22717041.

W. Tack, M. Madder, L. Baeten, M. Vanhellemont, and K. Verheyen. Shrub clearing adversely affects the abundance of Ixodes ricinus ticks. *Experimental & applied acarology*, 60(3):411–20, jul 2013. ISSN 1572-9702. doi: 10.1007/s10493-013-9655-0. URL http://www.ncbi.nlm.nih.gov/pubmed/23344639.

A. J. Tatem, S. I. Hay, and D. J. Rogers. Global traffic and disease vector dispersal. *PNAS*, 103(16):6242–6247, 2006. ISSN 0939351X. doi: 10.1073/pnas.0508391103.

A. J. Tatem, Z. Huang, A. Das, Q. Qi, J. Roth, and Y. Qiu. Air travel and vector-borne disease movement. *Parasitology*, 139(14):1816–1830, 2012. ISSN 00311820. doi: 10.1017/S0031182012000352.

J. Tournadre. Anthropogenic pressure on the open ocean: The growth. *Geophysical Research Letters*, 41(22):7924–7932, 2014. ISSN 00948276. doi: 10.1002/2014GL061786.

P. Tran and L. Tran. Validating negative binomial lyme disease regression model with bootstrap resampling. *Environmental Modelling and Software*, 82:121 – 127, 2016. ISSN 1364-8152. doi: https://doi.org/10.1016/j.envsoft.2016.04.019. URL http://www.sciencedirect.com/science/article/pii/S1364815216301141.

A. R. Tuite, A. L. Greer, and D. N. Fisman. Effect of latitude on the rate of change in incidence of Lyme disease in the United States. *CMAJ Open*, 1 (1):E43–E47, 2013. ISSN 2291-0026. doi: 10.9778/cmajo.20120002. URL http://cmajopen.ca/cgi/doi/10.9778/cmajo.20120002.

UNDRR. Report of the open-ended intergovernmental expert working group on indicators and terminology relating to disaster risk reduction.

Technical report, UN Disaster Risk Reduction, 2016. URL `https://www.unisdr.org/we/inform/publications/51748`.

I. V. Uspensky. Blood-sucking ticks (Acarina, Ixodoidea) as an essential component of the urban environment. *Entomological Review*, 97(7):941–969, 2017. ISSN 0013-8738. doi: 10.1134/S0013873817070107. URL `http://link.springer.com/10.1134/S0013873817070107`.

C. C. van den Wijngaard, A. Hofhuis, M. G. Harms, J. A. Haagsma, A. Wong, G. A. de Wit, A. H. Havelaar, A. K. Lugnér, A. W. M. Suijkerbuijk, and W. van Pelt. The burden of Lyme borreliosis expressed in disability-adjusted life years. *European journal of public health*, 5:1–8, 2015. ISSN 1464-360X. doi: 10.1093/eurpub/ckv091.

C. C. van den Wijngaard, A. Hofhuis, A. Wong, M. G. Harms, G. A. De Wit, A. K. Lugnér, A. W. Suijkerbuijk, M. J. J. Mangen, and W. Van Pelt. The cost of Lyme borreliosis. *European Journal of Public Health*, 27(3):538–547, 2017. ISSN 1464360X. doi: 10.1093/eurpub/ckw269.

A. J. H. van Vliet, W. A. Bron, and S. Mulder. The how and why of societal publications for citizen science projects and scientists. *Int J Biometeorol*, 2014. doi: http://dx.doi.org/10.1007/s00484-014-0821-9.

A. Vandenesch, C. Turbelin, E. Couturier, C. Arena, B. Jaulhac, E. Ferquel, V. Choumet, C. Saugeon, E. Coffinieres, T. Blanchon, V. Vaillant, and T. Hanslik. Incidence and hopstialisation rates of lyme borreliosis, Frans, 2004 to 2012. *Eurosurveillance*, 19:20883, 2014. ISSN 1560-7917 (Electronic). doi: 20883. URL `e:http://www.eurosurveillance.org`.

S. O. Vanwambeke, D. Sumilo, A. Bormane, E. F. Lambin, and S. E. Randolph. Landscape predictors of tick-borne encephalitis in Latvia: land cover, land use, and land ownership. *Vector Borne Zoonotic Dis*, 10(5):497–506, 2010. doi: 10.1089/vbz.2009.0116[doi].

S. Varela, R. P. Anderson, R. García-Valdés, and F. Fernández-González. Environmental filters reduce the effects of sampling bias and improve predictions of ecological niche models. *Ecography*, 37(11):1084–1091, 2014. ISSN 16000587. doi: 10.1111/j.1600-0587.2013.00441.x.

U. Vesco, N. Knap, M. B. Labruna, T. Avšič-Županc, A. Estrada-Peña, A. A. Guglielmone, G. H. Bechara, A. Gueye, A. Lakos, A. Grindatto, V. Conte, and D. de Meneghi. An integrated database on ticks and tick-borne zoonoses in the tropics and subtropics with special reference to developing and emerging countries. *Experimental and Applied Acarology*, 54(1):65–83, 2011. ISSN 01688162. doi: 10.1007/s10493-010-9414-4.

J. G. Vos, E. Dybing, H. A. Greim, O. Ladefoged, C. Lambré, J. V. Tarazona, I. Brandt, and D. A. Vethaak. Health effects of endocrine-disrupting chemicals on wildlife , with special re ... *Critical Reviews in Toxicology*, 30 (1):71–133, 2000.

M. Waskom, O. Botvinnik, D. O'Kane, P. Hobson, S. Lukauskas, D. C. Gemperline, T. Augspurger, Y. Halchenko, J. B. Cole, J. Warmenhoven, J. de Ruiter, C. Pye, S. Hoyer, J. Vanderplas, S. Villalba, G. Kunter, E. Quintero, P. Bachant, M. Martin, K. Meyer, A. Miles, Y. Ram, T. Yarkoni,

M. L. Williams, C. Evans, C. Fitzgerald, Brian, C. Fonnesbeck, A. Lee, and A. Qalieh. Seaborn, sep 2017. URL https://doi.org/10.5281/zenodo.883859{#}.W0CLrtVh4{_}R.mendeley.

M. Welvaert and P. Caley. Citizen surveillance for environmental monitoring: combining the efforts of citizen science and crowdsourcing in a quantitative data framework. *SpringerPlus*, 5(1), 2016. ISSN 21931801. doi: 10.1186/s40064-016-3583-5.

WHO. Preventive and control of dengue hemmorhagic fever. 2009. ISSN 0074-0276. doi: 10.1590/S0074-02761992000700024.

WHO. A global brief on vector-borne diseases. Technical report, 2014.

WHO. IARC Monographs evaluate DDT, lindane, and 2,4-D. Technical Report June, 2015. URL http://www.iarc.fr/en/media-centre/pr/2015/pdfs/pr236{_}E.pdf.

WHO. Overlapping global distribution of nine major vector-borne diseases, 2016. Technical report, 2016.

WHO. Global Vector Control Response 2017 - 20130. Technical report, World Health Organization, 2017. URL http://apps.who.int/iris/bitstream/handle/10665/259205/9789241512978-eng.pdf?sequence=1.

WMO. Declaration of the World Climate Conference. Technical report, Geneva, 1979.

World Bank. World Development Report (1993): Investing in Health. Technical report, 1993.

X. Wu, V. R. S. K. Duvvuri, and J. Wu. Modeling dynamical temperature influence on tickixodes scapularis population. In *Proceedings of the International Congress on Environmental Modelling and Software*, 2010.

A. Yang, H. Fan, and N. Jing. Amateur or Professional: Assessing the Expertise of Major Contributors in OpenStreetMap Based on Contributing Behaviors. *ISPRS International Journal of Geo-Information*, 5(2):21, 2016. ISSN 2220-9964. doi: 10.3390/ijgi5020021. URL http://www.mdpi.com/2220-9964/5/2/21.

D. Yang, C. Xu, J. Wang, and Y. Zhao. Spatiotemporal epidemic characteristics and risk factor analysis of malaria in Yunnan Province , China. *BMC Public Health*, pages 1–10, 2017. ISSN 1471-2458. doi: 10.1186/s12889-016-3994-9. URL http://dx.doi.org/10.1186/s12889-016-3994-9.

C. Zeimes, G. E. Olsson, M. Hjertqvist, and S. O. Vanwambeke. Shaping zoonosis risk : landscape ecology vs . landscape attractiveness for people , the case of tick- borne encephalitis in Sweden. pages 1–10, 2014.

P. Zeman and C. Benes. Peri-urbanisation, counter-urbanisation, and an extension of residential exposure to ticks: A clue to the trends in Lyme borreliosis incidence in the Czech Republic? *Ticks and Tick-borne Diseases*, 5(6):907–916, 2014. ISSN 1877-959X. doi: 10.1016/j.ttbdis.2014.07.006.

P. Zeman, C. Benes, and K. Markvart. Increasing Residential Proximity of Lyme Borreliosis Cases to High-Risk Habitats: A Retrospective Study in Central Bohemia, the Czech Republic, 1987–2010. *EcoHealth*, 12(3):519–522, 2015. ISSN 16129210. doi: 10.1007/s10393-015-1016-5.

G. Zhang and A. X. Zhu. The representativeness and spatial bias of volunteered geographic information: a review. *Annals of GIS*, 24(3):151–162, 2018. ISSN 19475691. doi: 10.1080/19475683.2018.1501607. URL `https://doi.org/10.1080/19475683.2018.1501607`.

X. Zhang, M. I. Meltzer, C. a. Peña, A. B. Hopkins, L. Wroth, and A. D. Fix. Economic impact of Lyme disease. *Emerging Infectious Diseases*, 12(4): 653–660, 2006. ISSN 10806040. doi: 10.3201/eid1204.050602.

B. Zhao and D. Z. Sui. True lies in geospatial big data: detecting location spoofing in social media. *Annals of GIS*, 23(1):1–14, 2017. ISSN 19475691. doi: 10.1080/19475683.2017.1280536. URL `http://dx.doi.org/10.1080/19475683.2017.1280536`.

A. X. Zhu, G. Zhang, W. Wang, W. Xiao, Z. P. Huang, G. S. Dunzhu, G. Ren, C. Z. Qin, L. Yang, T. Pei, and S. Yang. A citizen data-based approach to predictive mapping of spatial variation of natural phenomena. *International Journal of Geographical Information Science*, 29(10):1864–1886, 2015. ISSN 13623087. doi: 10.1080/13658816.2015.1058387.