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

Irene Garcia-Marti[1], **Raul Zurita-Milla**[1]*, **Arno Swart**[2]

1 Department of Geo-Information Processing, Faculty of Geo-Information and Earth Observation (ITC), University of Twente, Enschede, the Netherlands, **2** Centre for Infectious Disease Control, National Institute for Public Health and the Environment (RIVM), Bilthoven, the Netherlands

* r.zurita-milla@utwente.nl

Abstract

The socio-economic and demographic changes that occurred over the past 50 years have dramatically expanded urban areas around the globe, thus bringing urban settlers in closer contact with nature. Ticks have trespassed the limits of forests and grasslands to start inhabiting green spaces within metropolitan areas. Hence, the transmission of pathogens causing tick-borne diseases is an important threat to public health. Using volunteered tick bite reports collected by two Dutch initiatives, here we present a method to model tick bite risk using human exposure and tick hazard predictors. Our method represents a step forward in risk modelling, since we combine a well-known ensemble learning method, Random Forest, with four count data models of the (zero-inflated) Poisson family. This combination allows us to better model the disproportions inherent in the volunteered tick bite reports. Unlike canonical machine learning models, our method can capture the overdispersion or zero-inflation inherent in data, thus yielding tick bite risk predictions that resemble the original signal captured by volunteers. Mapping model predictions enables a visual inspection of the spatial patterns of tick bite risk in the Netherlands. The Veluwe national park and the Utrechtse Heuvelrug forest, which are large forest-urban interfaces with several cities, are areas with high tick bite risk. This is expected, since these are popular places for recreation and tick activity is high in forests. However, our model can also predict high risk in less-intensively visited recreational areas, such as the patchy forests in the northeast of the country, the natural areas along the coastline, or some of the Frisian Islands. Our model could help public health specialists to design mitigation strategies for tick-borne diseases, and to target risky areas with awareness and prevention campaigns.

Background

Over the last couple of decades, urban areas have dramatically expanded [1]. In Europe, the development of low density residential areas in the periphery of cities has become the norm for urban growth. [1]. 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) [2]. In addition, urban

sprawl also brings urban settlers in closer contact with nature and the countryside [3]. 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) [4]. 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 [5]. In parallel to this, the progressive adoption of healthier lifestyles encourages citizens to spend more time outdoors carrying out leisure [6] or sportive activities [7]. This behavior leads to a higher exposure to tick-borne diseases [8].

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 [9]. 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 [10], Milan [11], Kiev [12], Warsaw [13] or Lisbon [14], and suburban forests in Paris [15], Budapest [16,17] or Wroclaw [18], 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 [19–21], but in the last years researchers have worked on integrating hazard and exposure metrics to model tick bite risk [22,23] because hazard maps alone are insufficient to identify locations with a high risk for LB [24].

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 [25,26], 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 [27,28], 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 [29], which can split the data into homogenous groups using decision tree rules, with GLM 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.

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 $R = H \times E$ [30]. Thus, the calculation of tick bite risk should include variables representing both the H and E components, which likely have different underlying risk factors.

In the case of ticks and LB, the first spirochetes were identified in the early 1980s [31], 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. [32–33]) 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) [34]. 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 [24] 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 [35] to predict tick bite risk.

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 [36,37], mast years [38,39], vegetation type [40], and weather variables [41] 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 [42], tick presence [43], tick activity [44], or tick dynamics [35], 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 nymphal 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 [35]. The map in Fig 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).

Human exposure

Human exposure to ticks is intrinsically linked to human behavior outdoors and to diverse socio-economic factors. For instance, [45] 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.
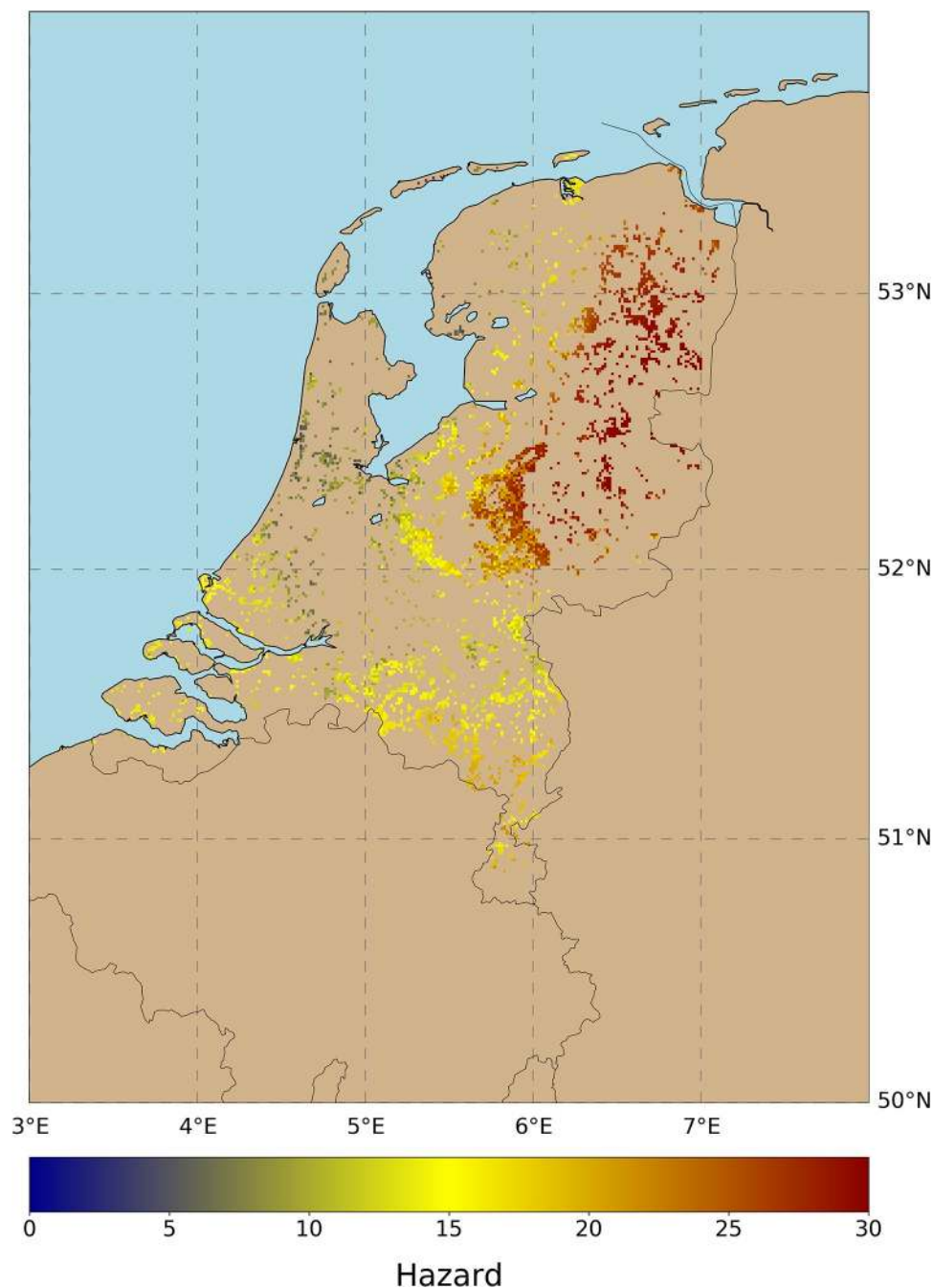
Fig 1. Tick activity per 1km grid cell. This tick activity estimation is provided by the data-driven model in [35], which is capable of predicting daily tick activity. We run this model for each day during the period (2006–2014) and we calculated a robust long-term metric of hazard, showing the maximum mean tick activity for the entire period. As seen, hazard is minimum along the coastline and maximum in the northeast of the country.

Similarly, the societal adoption of healthier lifestyles encourages citizens to spend more time outdoors carrying out physical or sportive activities. For instance, in [7] 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, in [46] the authors 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 [22]. All these examples are associated to 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 [6] and two studies in the USA [47] and Czech Republic [48] 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 [23]. 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 [49] the authors assess the willingness of residents to travel for woodland leisure, because it varies as a 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 [50], or because they have intrinsic value such as high scenic beauty or a good preservation [51].

Methods

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"; http://www.natuurkalender.nl) and the Tekenradar (TR; "tick radar"; https://www.tekenradar.nl) citizen science initiatives promoted by WUR and the RIVM, which to date are still ongoing. 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. Both initiatives continue collecting data, but we only have access to the reports until the end of 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 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 2.1. This aggregation groups together observations that are close in geographic space (Fig 2a). However, it also creates a grid with right-skewed and zero-inflated (Fig 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. It is also worth noting that the percentage of cells with over 20 tick bites is 0,42% (152 cells), and with over 30 tick bites is 0,14% (51 cells). Skewedness and zero-inflation are common real-world problems, especially when modelling count data, [52]. Thus the analysis of the tick bite reports requires a modelling approach capable of handling these characteristics [27,28].
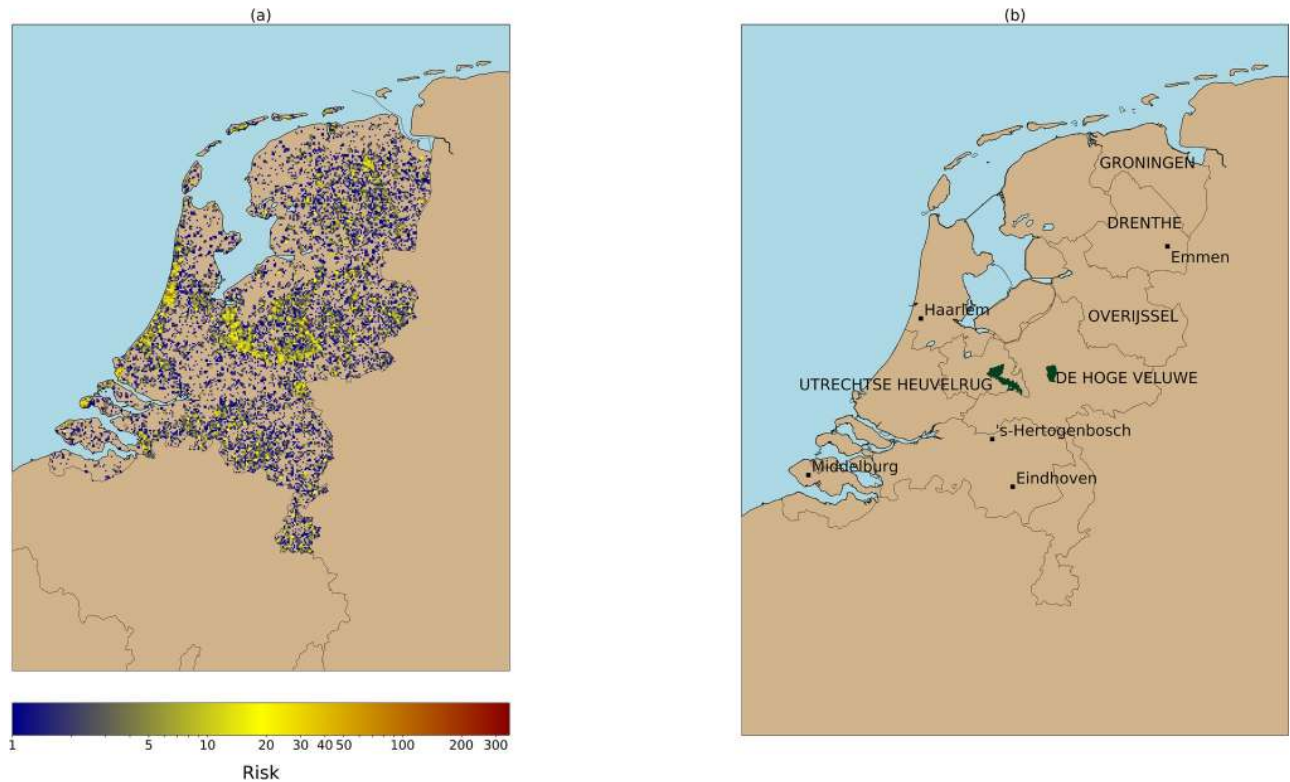
Fig 2. (a) Tick bite risk (2006–2016) as monitored by the NK and TR initiatives. We use as a proxy of tick bite risk the cumulative sum of tick bite reports per 1km grid cell. This image shows that tick bites are produced throughout the country, but the reports tend to be clustered around forests (e.g. 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.

Ensemble learning from skewed and zero-inflated datasets

Random Forest (RF) [29] 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 [53]. RF combines bagging [54] and the random subspace method [55]. 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 [27,56]. Learning from a imbalanced (classification) or skewed (regression) dataset, is a non-trivial problem that started to receive attention in the early 2000s [57]. According to [27] 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
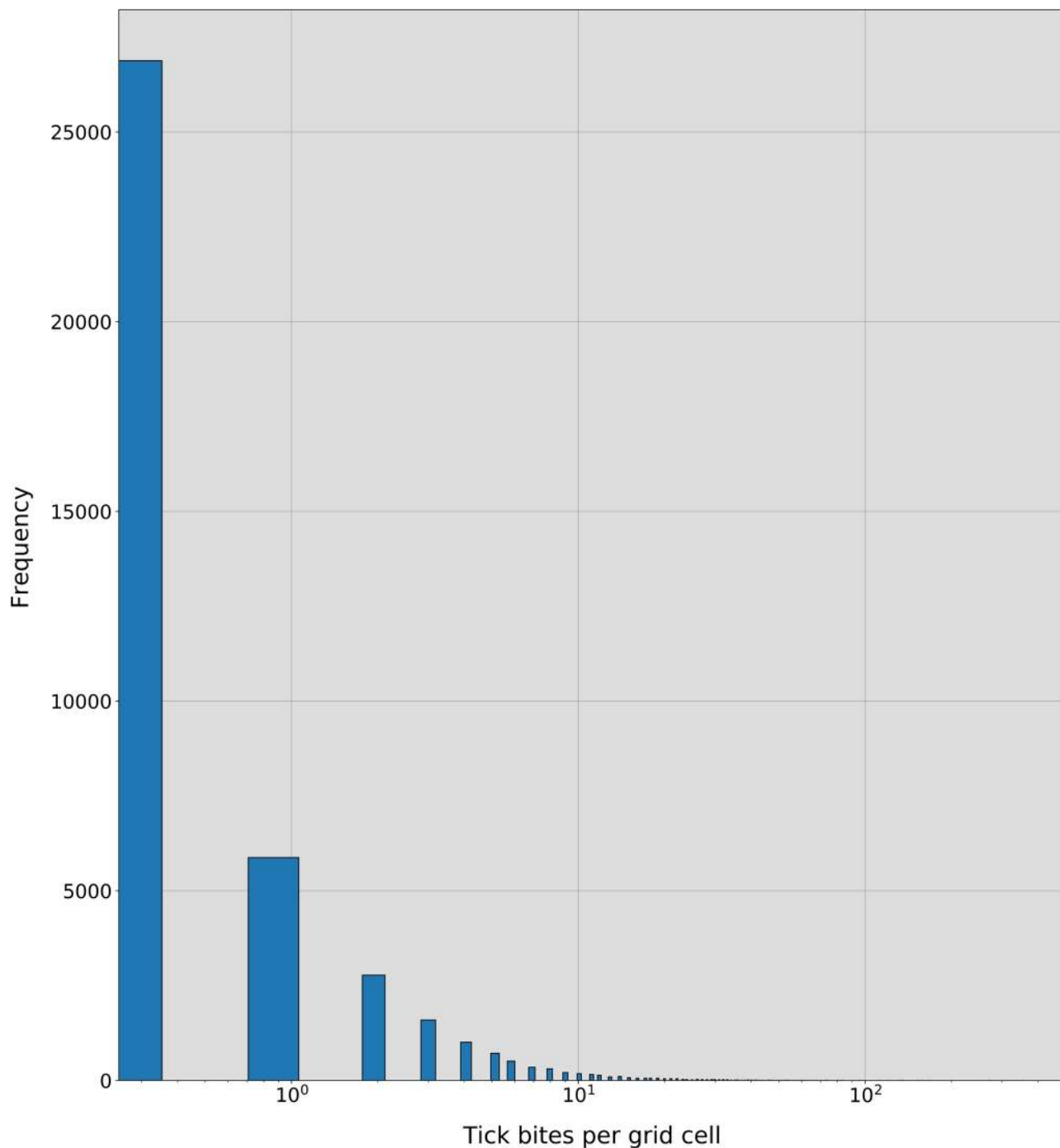
Fig 3. Histogram of the tick bites per grid cell. As seen, after the process of spatial aggregation described in Section 3.1, a skewed distribution with zero-inflation is created. Note that the X-axis is represented in log-scale to ease the visualization of the distribution. The number of grid cells containing more than 30–40 tick bite reports is almost negligible, but the distribution spans until a maximum of 353 tick bites per cell.

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 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 (Fig 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) [58], and the zero-inflated negative binomial (ZINB) [59] 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 hazard is present (presence of an established tick population), 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 the "rate of transmission", conceptually composed of visiting rates and biting rates. These count data models fit data considering the covariates described in Section 3.3.1, which will determine the value of the λ rate of transmission parameter.

Zero-inflated models have been used to predict TBE in a set-up in which the majority of the available samples had a zero [60]. 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 [28]. 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. Note that for the sake of brevity, we refer to the combination of RF with the count data models only by using the name of the count data model. Hence, the combination of RF with Poisson, negative binomial, zero-inflated Poisson, and zero-inflated negative binomial models is referred to as 'POI', 'NB', 'ZIP', and 'ZINB', respectively. Fig 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. 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). These predictions are obtained by drawing numbers from the estimated distributions computed in the training step, which are transformed into discrete values using the quantile function. After each tree has yielded the four discrete predictions, the final prediction of the ensemble is calculated. In this step, we keep the canonical behavior of RF by averaging these predictions for each model type. This final result per count data model will be rounded to provide an estimation of tick bites per grid cell which is in the integer domain.

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: scikit-learn [61] for the data-driven modelling, numpy [62] to handle the multidimensional arrays, statsmodels [63] to fit the count data models, GDAL [64] and Cartopy [65] to process geospatial data and prepare the visualizations through map layers, matplotlib [66] to prepare the rest of the figures, and SkillMetrics (https://github.com/PeterRochford/SkillMetrics)library and scipy [67] to obtain the statistical metrics used to evaluate the model.

Fig 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.

Feature engineering. In this study, we extend the ideas regarding human exposure described in [23]. 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 2.1, hazard metrics are extracted for forests and natural grasslands using the model developed by [35]. Table 1 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.

We derived accessibility metrics (Table 1, indices 1–7, Type AC) from the land-use map (BBG 2008) provided by Statistics Netherlands (CBS; https://www.cbs.nl/en-gb), and from the transportation networks contributed by volunteers in OpenStreetMaps (OSM; https://www.openstreetmap.org). 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. Note that we compute the distance between each point and the closest boundary of a polygon. We downloaded the latest snapshot of OSM for the

Table 1. Features used in this work.

Id	Name	Data type	Description	Source	Provider	Type of metric
1	dbuiltup	integer	Distance to the closest built-up area	BBG2008	Statistics Netherlands (CBS)	AC
2	dforest	integer	Distance to the closest forest patch	BBG2008	Statistics Netherlands (CBS)	AC
3	drecreation	integer	Distance to the closest recreational area	BBG2008	Statistics Netherlands (CBS)	AC
4	dbrr	integer	Distance to the closest regional bike route	OpenStreetMap	OSM Foundation	AC
5	dwrl	integer	Distance to the closest local walking path	OpenStreetMap	OSM Foundation	AC
6	dwrr	integer	Distance to the closest regional walking route	OpenStreetMap	OSM Foundation	AC
7	dwrn	integer	Distance to the closest national walking route	OpenStreetMap	OSM Foundation	AC
8	dcamping	integer	Distance to the closest camping location	TOP10NL	Cadaster Netherlands	AT
9	dcaravan	integer	Distance to the closest caravan parking location	TOP10NL	Cadaster Netherlands	AT
10	dcross	integer	Distance to the closest bike cross circuit	TOP10NL	Cadaster Netherlands	AT
11	dgolf	integer	Distance to the closest golf course	TOP10NL	Cadaster Netherlands	AT
12	dheem	integer	Distance to the closest wild garden	TOP10NL	Cadaster Netherlands	AT
13	dhaven	integer	Distance to the closest non-commercial haven	TOP10NL	Cadaster Netherlands	AT
14	dsafari	integer	Distance to the closest safari park	TOP10NL	Cadaster Netherlands	AT
15	dwater	integer	Distance to the closest waterbody (pond or lake)	TOP10NL	Cadaster Netherlands	AT
16	dbath	integer	Distance to the closest authorized bathing location	Swimming locations and water quality	National Registry	AT
17	attr	category	Attractiveness of the location	Attractiveness of the landscape	Wageningen University & Research (WUR) / National Registry	AT
18	LU	category	Land use at location	BBG2008	Statistics Netherlands (CBS)	AT
19	LC	category	Land cover at location	LGN7	Wageningen University & Research (WUR)	AT
20	maxmeanhaz	float	Max. mean hazard in the period 2006–2014 within forests	Tick dynamics	(Garcia-Marti et al, 2017)	HZ
21	maxstdhaz	float	Max. standard deviation of hazard in the period 2006–2014 within forests	Tick dynamics	(Garcia-Marti et al, 2017)	HZ

https://doi.org/10.1371/journal.pone.0216511.t001

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 1, 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 (https://zakelijk.kadaster.nl/-/top10nl) 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*; https://data.overheid.nl/data/dataset/49505-belevingswaarde-van-het-landschap), last accessed July 5th, 2018) [68,69].

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 1, indices 20–21, Type HZ) are extracted from the model outlined in Section 2.1. This model, described in more detail in [35], 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 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 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 [70], 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.

Experiments. The spatial aggregation and feature engineering described in sections 3.1 and 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, yielding the models POI, NB, ZIP, and ZINB. Each decision tree in the ensemble can be viewed as a segmentation of the data into relatively homogeneous parts, with regards to both the outcome variable and covariates. It is our expectation that most of the complicated non-linear, or even discontinuous, dependencies of the outcome on the data are handled by the decision tree. The use of a GLM at the terminal node level is meant to provide a simple model to capture the remaining structure in the data, which at this level should be fairly simple given the homogeneity of the data segments at the terminal nodes. Thus, the data-driven count data models will use the features derived in Section 3.1 for the fitting. Note that the count data models use 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. Here, two special cases can occur: 1) all samples in a leaf node are zero; 2) the majority of samples in a node, but not all of them, are zero. In the first case, we skip the part of fitting a count data model because it would be meaningless. Instead, a pure zero is returned during the testing phase. In the second case, we keep the canonical behavior of RF by returning the average of all samples falling in that tree node. Since most of these samples are zeros, the average is close to zero as well. 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, or return a zero as described above, 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) discretized 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 the lowest RMSD, the highest correlation coefficient and a standard deviation as close as possible to the tick bites per grid cell distribution, 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 Fig 5.

Results

Fig 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. However, POI and ZIP provide poor fits, since they overpredict tick bite counts in the range 1–10. 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, which 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 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 on the ability of the models to predict zeros.
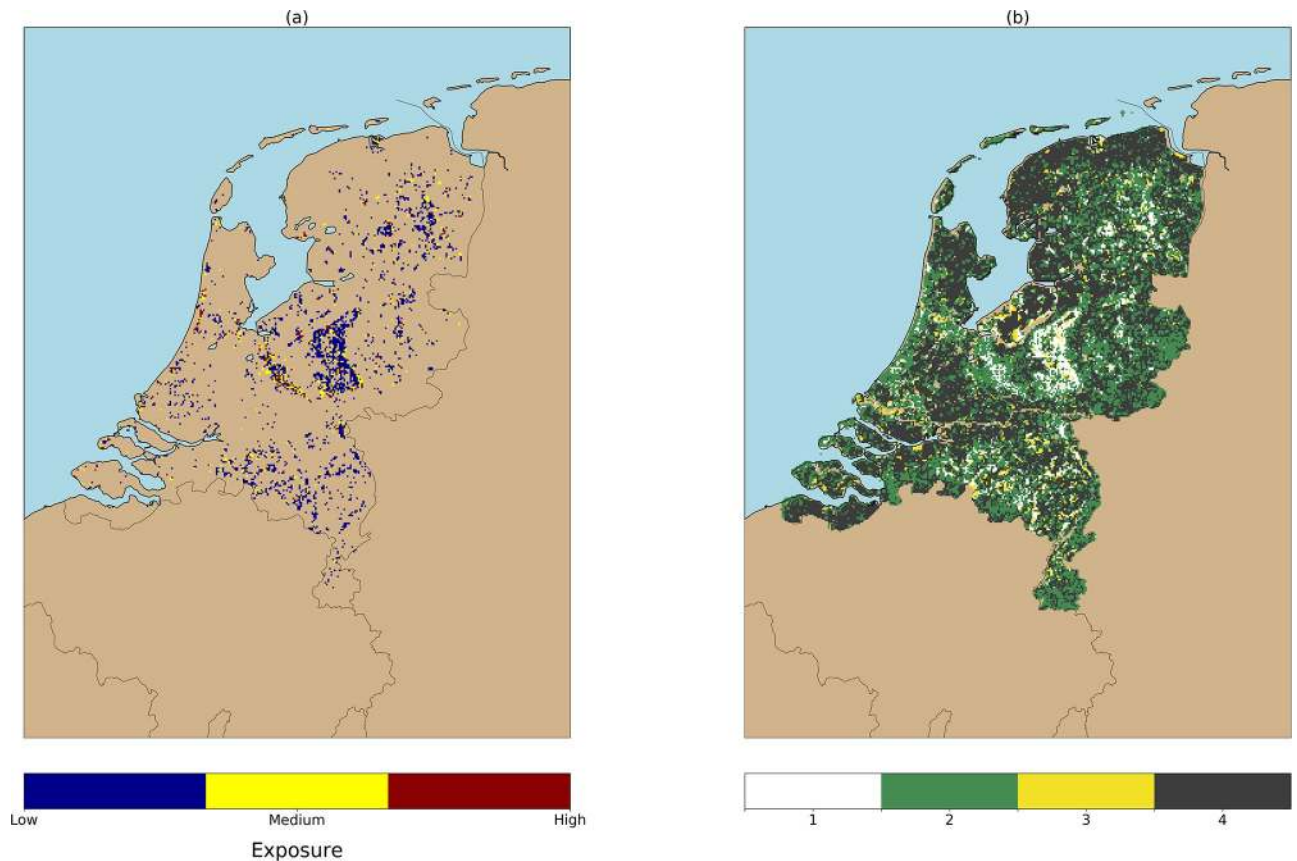
Fig 5. (a) Human exposure to tick bites classified in three categories. (b) Class 1 in this map correspond to the classes from (a), class 2 represents tick bites reported outside forests, class 3 represents forests with no tick bites recorded, and class 4 shows locations where no tick bites were reported during the study period. These results can be found in [24], and we cross them with the tick bite risk maps obtained in this work to explore the risk per human exposure category (Fig 9).

https://doi.org/10.1371/journal.pone.0216511.g005

Fig 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). Using the pink hexagon as a reference point, we can see there are NB, ZIP, and ZINB models below the arc RMSD = 2, that also present a high Pearson/Kendall correlation (i.e. >0.7) and a fair range of standard deviation (i.e. 2–6). In this work we are not only interested in models with a high correlation and low error, but also in those providing a standard deviation (stdev) that is similar to the one found in our collection of tick bites per grid cell, which equates to 3,15. The stdev of the models under the arc traced by RMSD = 2 are close to this value. A closer inspection to these models reveals that there is a small cluster of NB and ZINB models with a stdev close to 3, whereas two ZIP and ZINB models present a stdev slightly over 4. These results suggest that the cluster is fitting better the zero-inflation in data, whereas the remaining models are better suited to predict overdispersion. These Taylor diagrams also show that the optimal experiments correspond to those created with 200 to 600 Samples per terminal node and with 20 to 50 trees. The tick bite risk maps are then created with the experiments with 200 samples per leaf node and 20 trees.

Fig 8 shows the tick bite risk maps produced by the four count data models (a-d), by the canonical RF (e), and a zoom-in of the maps obtained with the ZIP and ZINB models (f-g). We applied a logarithmic scale plus one (i.e. log-10) to better visualize the location of the risk
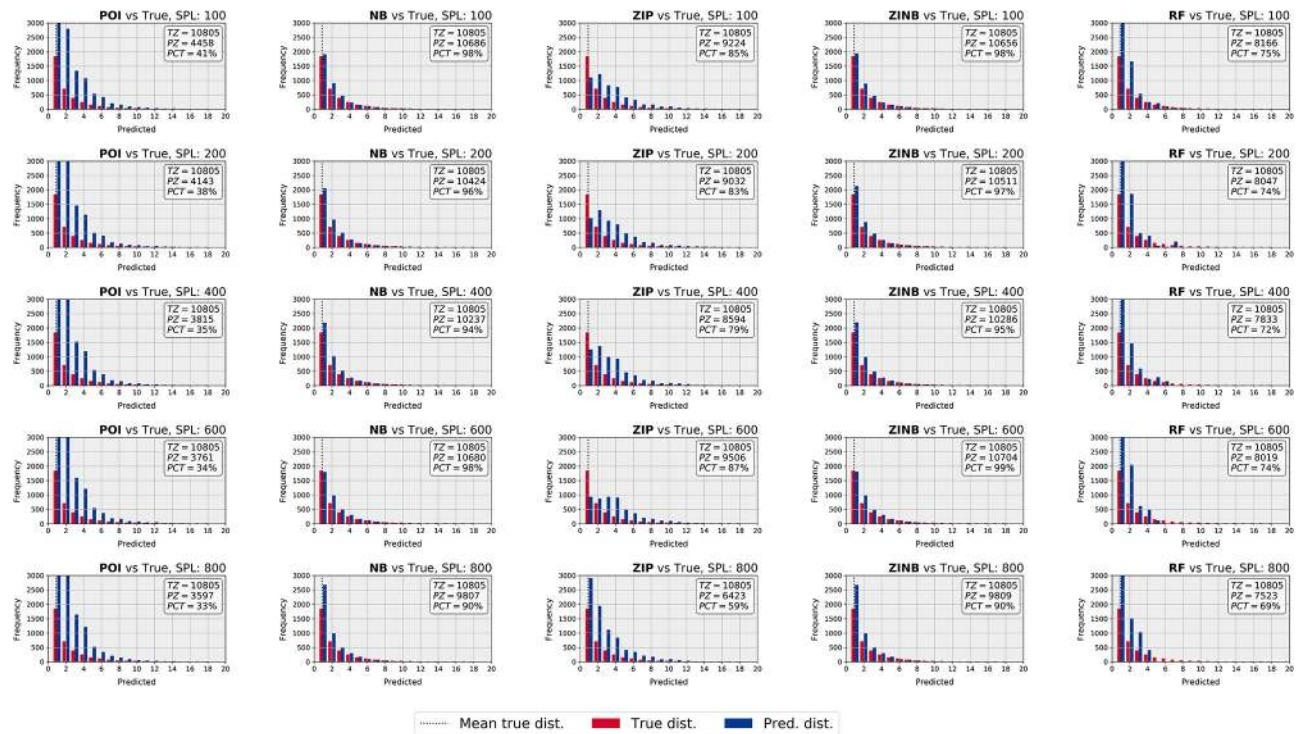
Fig 6. Histograms of original (red) and predicted (blue) distributions for an ensemble with 20 trees and an increasing number of samples per leaf node. Note that for visualization purposes, the axes have been limited and the zeros are summarized in the text box within each subplot, thus showing the number 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.

https://doi.org/10.1371/journal.pone.0216511.g006

areas. Note that this log-10 scale represents a range of 0–30 predicted tick bites per grid cell. The application of POI models at the country level create maps that present a wide range of predicted tick bites, so that most of the locations in the country present a risk of tick bites. The ZIP, NB, and ZINB models yield maps that seem able to distinguish between locations with no risk, and locations with a certain risk, due to the better handling of the inherent zero-inflation. The canonical RF yields predictions that are mostly uniform throughout the territory and due to the inability of handling zero-inflation and overdispersion simultaneously, this method does not identify well locations with no risk. In Figs 6 and 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. Figs 6 and 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 (Fig 8f and 8g) 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.

Fig 9 depicts the risk of tick bite classified by the exposure levels found in Fig 5. We show a plot for each count model (a-d), for the canonical RF (e), and the original volunteered reports
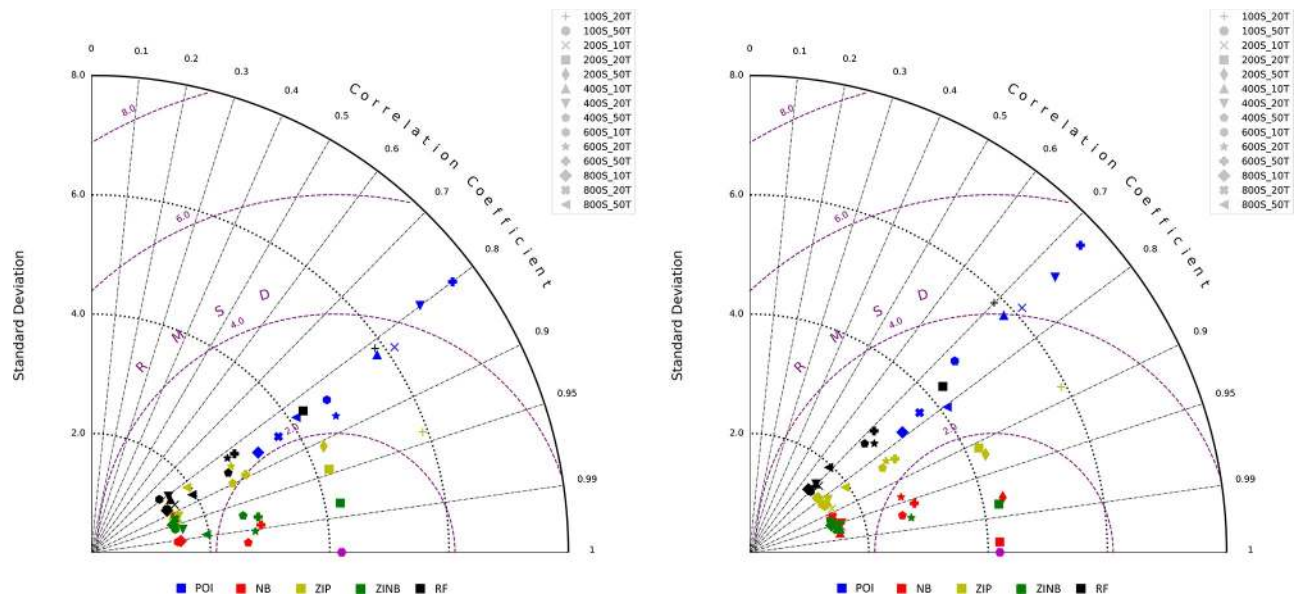
Fig 7. Modified Taylor diagrams showing the performance of the models based in three metrics: Standard deviation, RMSD, and a correlation coefficient (left: Spearman's, right: Kendall Tau), which are represented by the Y, circular, and radial axes, respectively. Each of the colored symbols represents an ensemble with a concrete number of tree estimators (T) and samples per leaf node (S). Using the pink hexagon as a reference, we can see that 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 standard deviation close to the raw tick bites per grid cell (i.e. stdev = 3.15). 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.

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 Fig 5. Considering the low, medium, and high exposure classes in Fig 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 Fig 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.

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
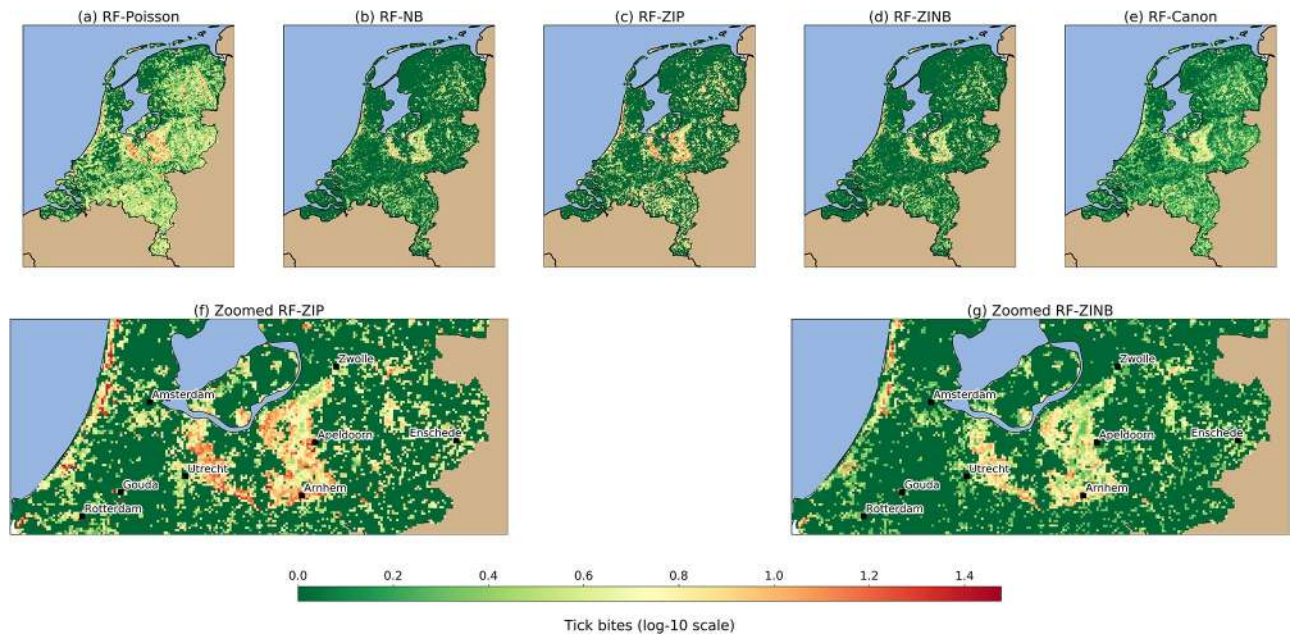
Fig 8. Tick bite risk maps produced by combining RF with each of the count data models. The logarithmic (log-10) scale corresponds to the range of 0–30 predicted tick bites per grid cell. The upper row (a-e) shows the general overview of the models, whereas the bottom row shows a close-up for the ZIP (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.

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 Fig 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.

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
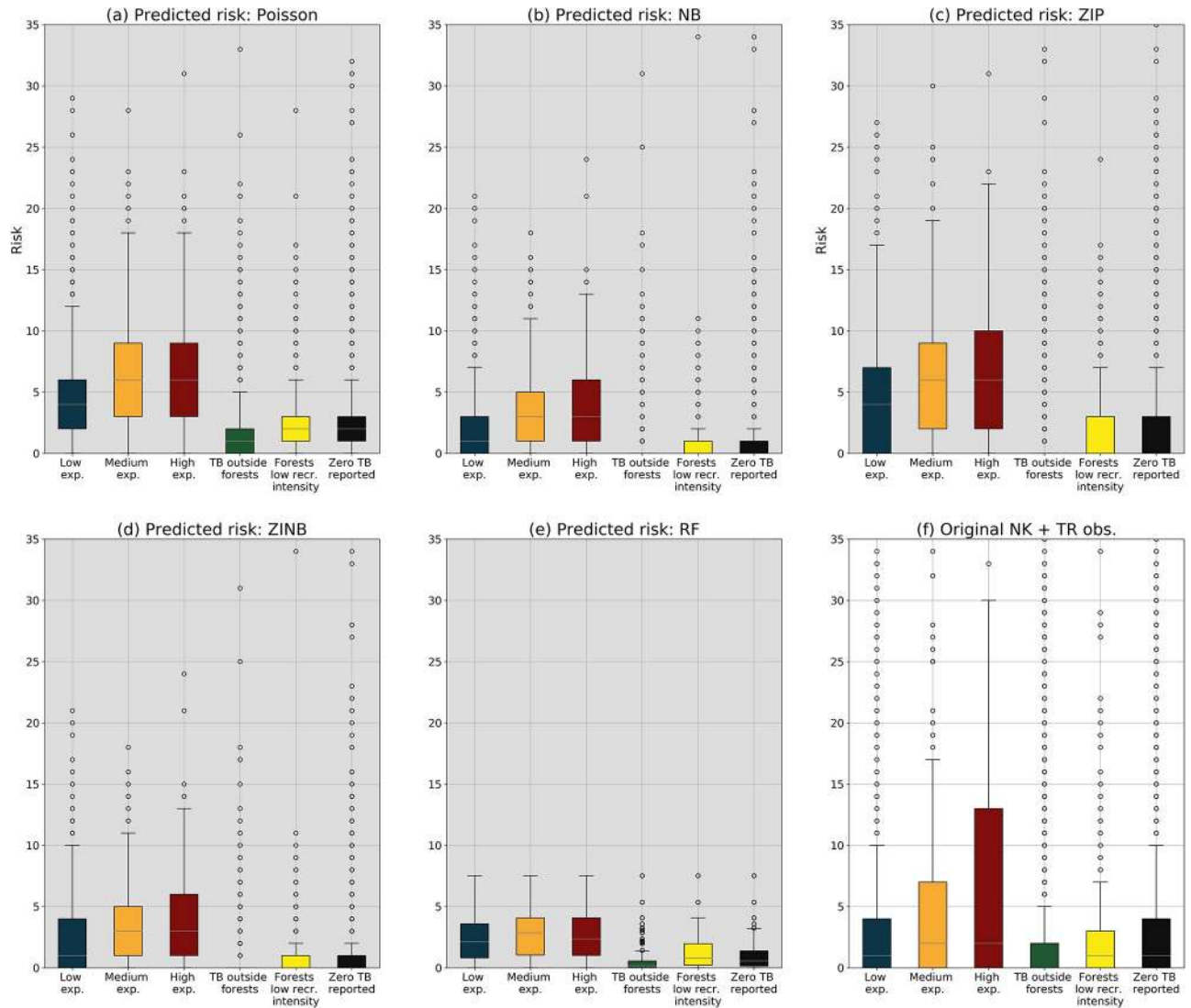
Fig 9. Tick bite risk classified based on the human exposure classes from Fig 5. The subplots show the predicted distributions per human exposure class for each of the count data models (a-d) and RF (e), and the type of human exposure using the original volunteered observations from NK and TR (f). The 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.

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 Fig 5 and the map layers in Fig 8, we inspected the risk of tick bites in function of human exposure inside and outside forests. In Fig 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. We foresee two possible research lines stemming from the present work. First, 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. Second, given that we are modelling a zero-inflated and over-dispersed variable, we could modify the canonical version of RF to return predictions based in other statistical metrics. The canonical approach uses the mean of the predictions yielded by the tree estimators, but other relevant metrics (e.g. median) could be more representative of the modelled distribution and might improve the current tick bite risk maps.

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.

Acknowledgments

Author Contributions

Conceptualization: Irene Garcia-Marti, Raul Zurita-Milla, Arno Swart.

Data curation: Irene Garcia-Marti.

Formal analysis: Irene Garcia-Marti.

Methodology: Irene Garcia-Marti, Raul Zurita-Milla, Arno Swart.

Supervision: Raul Zurita-Milla.

Validation: Irene Garcia-Marti, Raul Zurita-Milla, Arno Swart.

Visualization: Irene Garcia-Marti.

Writing – original draft: Irene Garcia-Marti.

Writing – review & editing: Irene Garcia-Marti, Raul Zurita-Milla, Arno Swart.

References

1. EEA EEA. The impacts of urban sprawl. 2006.

2. EEA EEA. Landscape Fragmentation in Europe. IlpoeUni-StuttgartDe. 2011.

3. Tack W. Impact of Forest Conversion on the Abundance of Ixodes Ricinus Ticks. Department of Forest and Water Management, Department of Biomedical Sciences. Ghent University. 2013.

4. Uspensky IV. Blood-sucking ticks (Acarina, Ixodoidea) as an essential component of the urban environment. Entomol Rev. 2017; 97: 941–969. https://doi.org/10.1134/S0013873817070107

5. Allan BF, Keesing F, Ostfeld RS. Effect of Forest Fragmentation on Lyme Disease Risk. Conserv Biol. 2003; 17: 267–272. https://doi.org/10.1046/j.1523-1739.2003.01260.x

6. Mulder S, van Vliet AJH, Bron WA, Gassner F, Takken W. High risk of tick bites in Dutch gardens. Vector Borne Zoonotic Dis. 2013; 13: 865–871. https://doi.org/10.1089/vbz.2012.1194 PMID: 24107214

7. Hall JL, Alpers K, Bown KJ, Martin SJ, Birtles RJ. Use of Mass-Participation Outdoor Events to Assess Human Exposure to Tickborne Pathogens. 2017; 23: 463–467.

8. Sandifer PA, Sutton-grier AE, Ward BP. Exploring connections among nature, biodiversity, ecosystem services, and human health and well-being: Opportunities to enhance health and biodiversity conservation $. Ecosyst Serv. 2015; 12: 1–15. https://doi.org/10.1016/j.ecoser.2014.12.007

9. Ehrmann S, Liira J, Gärtner S, Hansen K, Brunet J, Cousins SAO, et al. Environmental drivers of Ixodes ricinus abundance in forest fragments of rural European landscapes. BMC Ecol. 2017; 1–14.

10. Oechslin CP, Heutschi D, Lenz N, Tischhauser W, Péter O, Rais O, et al. Prevalence of tick-borne pathogens in questing Ixodes ricinus ticks in urban and suburban areas of Switzerland. Parasites and Vectors. 2017; 10: 1–18.

11. Olivieri E, Gazzonis AL, Zanzani SA, Veronesi F, Manfredi MT. Seasonal dynamics of adult Dermacentor reticulatus in a peri-urban park in southern Europe. Ticks Tick Borne Dis. 2017; 8: 772–779. https://doi.org/10.1016/j.ttbdis.2017.06.002 PMID: 28647128

12. Didyk YM, Blaňárová L, Pogrebnyak S, Akimov I, Peťko B, Víchová B. Emergence of tick-borne pathogens (Borrelia burgdorferi sensu lato, Anaplasma phagocytophilum, Rickettsia raoultii and Babesia microti) in the Kyiv urban parks, Ukraine. Ticks Tick Borne Dis. 2017; 8: 219–225. https://doi.org/10.1016/j.ttbdis.2016.10.002 PMID: 27923669

13. Kowalec M, Szewczyk T, Welc-Falęciak R, Siński E, Karbowiak G, Bajer A. 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. 2017; 10: 1–19.

14. Santos AS, de Bruin A, Veloso AR, Marques C, Pereira da Fonseca I, de Sousa R, et al. Detection of Anaplasma phagocytophilum, Candidatus Neoehrlichia sp., Coxiella burnetii and Rickettsia spp. in questing ticks from a recreational park, Portugal. Ticks Tick Borne Dis. 2018; 9: 1555–1564. https://doi.org/10.1016/j.ttbdis.2018.07.010 PMID: 30097348

15. Paul REL, Cote M, Le Naour E, Bonnet SI. Environmental factors influencing tick densities over seven years in a French suburban forest. Parasites and Vectors. 2016; 9: 1–10.

16. Szekeres S, Docters van Leeuwen A, Tóth E, Majoros G, Sprong H, Földvári G. Road-killed mammals provide insight into tick-borne bacterial pathogen communities within urban habitats. Transbound Emerg Dis. 2018. https://doi.org/10.1111/tbed.13019 PMID: 30230270

17. Szekeres S, Docters van Leeuwen A, Rigó K, Jablonszky M, Majoros G, Sprong H, et al. Prevalence and diversity of human pathogenic rickettsiae in urban versus rural habitats, Hungary. Exp Appl Acarol. 2016; 68: 223–226. https://doi.org/10.1007/s10493-015-9989-x PMID: 26613759

18. Kiewra D, Stefańska-Krzaczek E, Szymanowski M, Szczepańska A. 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 Tick Borne Dis. 2017; 8: 362–369. https://doi.org/10.1016/j.ttbdis.2016.12.011 PMID: 28089124

19. LoGiudice K, Ostfeld RS, Schmidt K a, Keesing F. The ecology of infectious disease: effects of host diversity and community composition on Lyme disease risk. Proc Natl Acad Sci U S A. 2003; 100: 567–71. https://doi.org/10.1073/pnas.0233733100 PMID: 12525705

20. Eisen RJ, Eisen L, Girard YA, Fedorova N, Mun J, Slikas B, et al. 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. 2010; 1: 35–43. https://doi.org/10.1016/j.ttbdis.2009.12.002 PMID: 20532183

21. Gassner F, Hansford KM, Medlock J. Greener cities, a wild card for ticks? In: Braks MA, van Wieren SE, Takken W, Sprong H, editors. Ecology and prevention of Lyme borreliosis. Wageningen Academic Publishers; 2016. pp. 187–203.

22. De Keukeleire M, Vanwambeke SO, Somassè E, Kabamba B, Luyasu V, Robert A. Scouts, forests, and ticks: Impact of landscapes on human-tick contacts. Ticks Tick Borne Dis. 2015; 6: 636–644. https://doi.org/10.1016/j.ttbdis.2015.05.008 PMID: 26055232

23. Zeimes C, Olsson GE, Hjertqvist M, Vanwambeke SO. Shaping zoonosis risk: landscape ecology vs. landscape attractiveness for people, the case of tick- borne encephalitis in Sweden. 2014; 1–10.

24. Garcia-Marti I, Zurita-Milla R, Harms MG, Swart A. Using volunteered observations to map human exposure to ticks. Sci Rep. 2018; 8: 15435. https://doi.org/10.1038/s41598-018-33900-2 PMID: 30337654

25. Senaratne H, Mobasheri A, Ali AL, Capineri C, Muki Haklay M. A review of volunteered geographic information quality assessment methods. Int J Geogr Inf Sci. 2017; 31: 139–167. https://doi.org/10.1080/13658816.2016.1189556

26. Mehdipoor H, Zurita-Milla R, Augustijn E-W, van Vliet AJH. Checking the Consistency of Volunteered Phenological Observations While Analysing Their Synchrony. Int J Geo-Information. 2018; 7: 22. https://doi.org/10.3390/ijgi7120487

27. Krawczyk B. Learning from imbalanced data: open challenges and future directions. Prog Artif Intell. 2016; 5: 221–232. https://doi.org/10.1007/s13748-016-0094-0

28. Lee SK, Jin S. Decision tree approaches for zero-inflated count data. J Appl Stat. 2006; 33: 853–865. https://doi.org/10.1080/02664760600743613

29. Breiman L. Random Forests. Mach Learn. 2001; 45: 5–32.

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

31. Burgdorfer W, Barbour A, Hayes S, Benach J, Grunwaldt E, Davis J. Lyme Disease: a tick-borne spirochetosis? Science (80-). 1982; 18: 1317–1319. https://doi.org/10.1126/science.7043737 PMID: 7043737

32. Falco RC, Fish D. Potential for Exposure to Tick Bites in Recreational Parks in a Lyme Disease Endemic Area. 1989; 79. https://doi.org/10.2105/AJPH.79.1.12 PMID: 2909174

33. Magnarelli LA, Denicola A, Stafford KC, Anderson JF. Borrelia burgdorferi in an urban environment: White-tailed deer with infected ticks and antibodies. J Clin Microbiol. 1995; 33: 541–544. PMID: 7751354

34. Ostfeld R. Lyme Disease: the ecology of a complex system. 1st Editio. New York, New York, USA: Oxford University Press; 2012.

35. Garcia-Martí I, Zurita-Milla R, van Vliet AJH, Takken W. Modelling and mapping tick dynamics using volunteered observations. Int J Health Geogr. 2017; 16. https://doi.org/10.1186/s12942-017-0114-8 PMID: 29137670

36. Ostfeld R, Canham C, Oggenfuss K, Winchcombe R, Keesing F. Climate, deer, rodents, and acorns as determinants of variation in lyme-disease risk. PLoS Biol. 2006; 4: e145. https://doi.org/10.1371/journal.pbio.0040145 PMID: 16669698

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

38. Kelly D, Koenig WD, Liebhold AM. An intercontinental comparison of the dynamic behavior of mast seeding communities. Popul Ecol. 2008; 50: 329–342. https://doi.org/10.1007/s10144-008-0114-4

39. Buonaccorsi JP, Elkinton J, Koenig W, Duncan RP, Kelly D, Sork V. Measuring mast seeding behavior: relationships among population variation, individual variation and synchrony. J Theor Biol. 2003; 224: 107–114. https://doi.org/10.1016/s0022-5193(03)00148-6 PMID: 12900208

40. Tack W, Madder M, Baeten L, De Frenne P, Verheyen K. The abundance of Ixodes ricinus ticks depends on tree species composition and shrub cover. Parasitology. 2012; 139: 1273–81. https://doi.org/10.1017/S0031182012000625 PMID: 22717041

41. Berger KA, Ginsberg HS, Gonzalez L, Mather TN. Relative humidity and activity patterns of Ixodes scapularis (Acari: Ixodidae). J Med Entomol. 2014; 51: 769–76. Available: http://www.ncbi.nlm.nih.gov/pubmed/25118408

42. Estrada-Peña A, de la Fuente J, Latapia T, Ortega C. The Impact of Climate Trends on a Tick Affecting Public Health: A Retrospective Modeling Approach for Hyalomma marginatum (Ixodidae). PLoS One. 2015; 10: e0125760. https://doi.org/10.1371/journal.pone.0125760 PMID: 25955315

43. Swart A, Ibañez-Justicia A, Buijs J, van Wieren SE, Hofmeester TR, Sprong H, et al. Predicting Tick Presence by Environmental Risk Mapping. Front Public Heal. 2014; 2: 1–8. https://doi.org/10.3389/fpubh.2014.00238 PMID: 25505781

44. Bennet L, Halling a, Berglund J. Increased incidence of Lyme borreliosis in southern Sweden following mild winters and during warm, humid summers. Eur J Clin Microbiol Infect Dis. 2006; 25: 426–32. https://doi.org/10.1007/s10096-006-0167-2 PMID: 16810531

45. Zeman P, Benes C. 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 Tick Borne Dis. 2014; 5: 907–916. https://doi.org/10.1016/j.ttbdis.2014.07.006 PMID: 25113985

46. Padgett KA, Bonilla DL. Novel exposure sites for nymphal Ixodes pacificus within picnic areas. Ticks Tick Borne Dis. 2011; 2: 191–195. https://doi.org/10.1016/j.ttbdis.2011.07.002 PMID: 22108011

47. Hahn MB, Bjork JKH, Neitzel DF, Dorr FM, Whitemarsh T, Boegler KA, et al. Evaluating acarological risk for exposure to Ixodes scapularis and Ixodes scapularis-borne pathogens in recreational and residential settings in Washington County, Minnesota. Ticks Tick Borne Dis. 2017; 0–1. https://doi.org/10.1016/j.ttbdis.2017.11.010 PMID: 29195857

48. Zeman P, Benes C, Markvart K. Increasing Residential Proximity of Lyme Borreliosis Cases to High-Risk Habitats: A Retrospective Study in Central Bohemia, the Czech Republic, 1987–2010. Ecohealth. 2015; 12: 519–522. https://doi.org/10.1007/s10393-015-1016-5 PMID: 25698296

49. Li S, Colson V, Lejeune P, Vanwambeke SO. On the distance travelled for woodland leisure via different transport modes in Wallonia, south Belgium. Urban For Urban Green. 2016; 15: 123–132. https://doi.org/10.1016/j.ufug.2015.12.007

50. Lambin EF, Tran A, Vanwambeke SO, Linard C, Soti V. Pathogenic landscapes: interactions between land, people, disease vectors, and their animal hosts. Int J Health Geogr. 2010; 9: 54. https://doi.org/10.1186/1476-072X-9-54 PMID: 20979609

51. Nielsen AB, Heyman E, Richnau G. Liked, disliked and unseen forest attributes: Relation to modes of viewing and cognitive constructs. J Environ Manage. 2012; 113: 456–466. https://doi.org/10.1016/j.jenvman.2012.10.014 PMID: 23122619

52. Hadiji F, Molina A, Natarajan S, Kersting K. Poisson Dependency Networks: Gradient Boosted Models for Multivariate Count Data. Mach Learn. 2015; 100: 477–507. https://doi.org/10.1007/s10994-015-5506-z

53. Louppe G, Wehenkel L, Sutera A, Geurts P. Understanding variable importances in forests of randomized trees. Neural Inf Process Syst. 2013; 1–9. NIPS2013_4928

54. Breiman L. Bagging predictors. Mach Learn. 1996; 24: 123–140. https://doi.org/10.1007/BF00058655

55. Ho TK. The random subspace method for constructing decision forests. IEEE Trans Pattern Anal Mach Intell. 1998; 20: 832–844. https://doi.org/10.1109/34.709601

56. Japkowicz N, Stephen S. The class imbalance problem: a systematic study. Intell Data Anal. 2002; 426–449.

57. He H, Garcia EA. Learning from imbalanced data. IEEE Trans Knowl Data Eng. 2009; 21: 1263–1284.

58. Lambert D. Zero-inflated poisson regression, with an application to defects in manufacturing. Technometrics. 1992; 34: 1–14.

59. Greene WH. Accounting for Excess Zeros and Sample Selection in Poisson and Negative Binomial Regression Models. Biol Philos. 1994; 9: 265–265.

60. Stefanoff P, Rubikowska B, Bratkowski J, Ustrnul Z, Vanwambeke SO, Rosinska M. A predictive model has identified tick-borne encephalitis high-risk areas in regions where no caseswere reported previously, Poland, 1999–2012. Int J Environ Res Public Health. 2018; 15: 1–17. https://doi.org/10.3390/ijerph15040677 PMID: 29617333

61. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, et al. Scikit-learn: Machine Learning in Python. J Mach Learn Res. 2011; 12: 2825–2830. Available: http://scikit-learn.sourceforge.net.

62. Oliphant TE. Guide to NumPy. 2006. p. 371. http://www.numpy.org/

63. Seabold S, Perktold J. Statsmodels: Econometric and Statistical Modeling with Python. PROC 9th PYTHON Sci CONF. 2010; 57.

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

65. Met Office UK. Cartopy: a cartographic python library with a matplotlib interface. Exeter, Devon; 2010. http://scitools.org.uk/cartopy

66. Hunter JD. Matplotlib: A 2D graphics environment. Computing in Science and Engineering. IEEE COMPUTER SOC; 2007. pp. 90–95.

67. Oliphant TE. Python for scientific computing. Comput Sci Eng. 2007; 10–20. https://doi.org/10.1109/MCSE.2007.58

68. Crommentuijn LEM, Farjon JMJ, den Dekker C, van der Wulp N. Belevingswaardenmonitor Nota Ruimte 2006: Nulmeting landschap en groen in en om de stad. Bilthoven; 2007.

69. Roos-Klein Lankhorst J, de Vries S, Buijs AE, Bloemmen MHI, Schuiling C. BelevingsGIS versie 2: waardering van het Nederlandse landschap door de bevolking op kaart. Wageningen; 2005.

70. Heylen D, Lasters R, Adriaensen F, Fonville M, Sprong H, Matthysen E. Ticks and tick-borne diseases in the city: Role of landscape connectivity and green space characteristics in a metropolitan area. Sci Total Environ. 2019. https://doi.org/10.1016/j.scitotenv.2019.03.235 PMID: 30921726