GROUP PROJECT

Final Report

Modelling Tick Bites in the Netherlands

Group 6

Anam Akhtar | s2244314 | a.akhtar-1@student.utwente.nl
Eqi Luo | s2244365 | e.luo@student.utwente.nl
Rhea Chib | s2337290 | r.s.chib@student.utwente.nl
Zijing Wu | s2278308 | z.wu-3@student.utwente.nl

June 2020

Faculty of Geo-Information Science and Earth Observation
University of Twente, Enschede, The Netherland

Table of Contents

1. INTRODUCTION

Ticks are spider-like animals that mainly found in areas dominated by natural vegetation like forests and meadows but can also be found in gardens and other green recreational areas. They are mostly found in the northern hemisphere and are typically not active during the winter months. To survive they need blood, they suck themselves into the skin of an animal or a human and consume blood for several hours or days. If identified in time, they can be removed easily, otherwise, they can transfer diseases like Lyme disease (a.k.a. Lyme Borreliosis). It is a serious illness that can even cause disability.

Consequently, it is essential to look more into detail about the factors that determine the risk of getting a tick bite, as they vary a lot over space and time.

1.1. Study Area

Municipality of Ede is a densely forested area in the Netherlands, with a population of 112,410 and an area of 318.62 km sq (CBS, 2020). The very famous De Hoge Veluwe National Park makes it a big tourist attraction hosting around nine million tourists every year. The extensive natural landscape in Ede makes it perfect for outdoor recreational activities. Figure 1 represents the geographical location and land-use distribution of Ede.

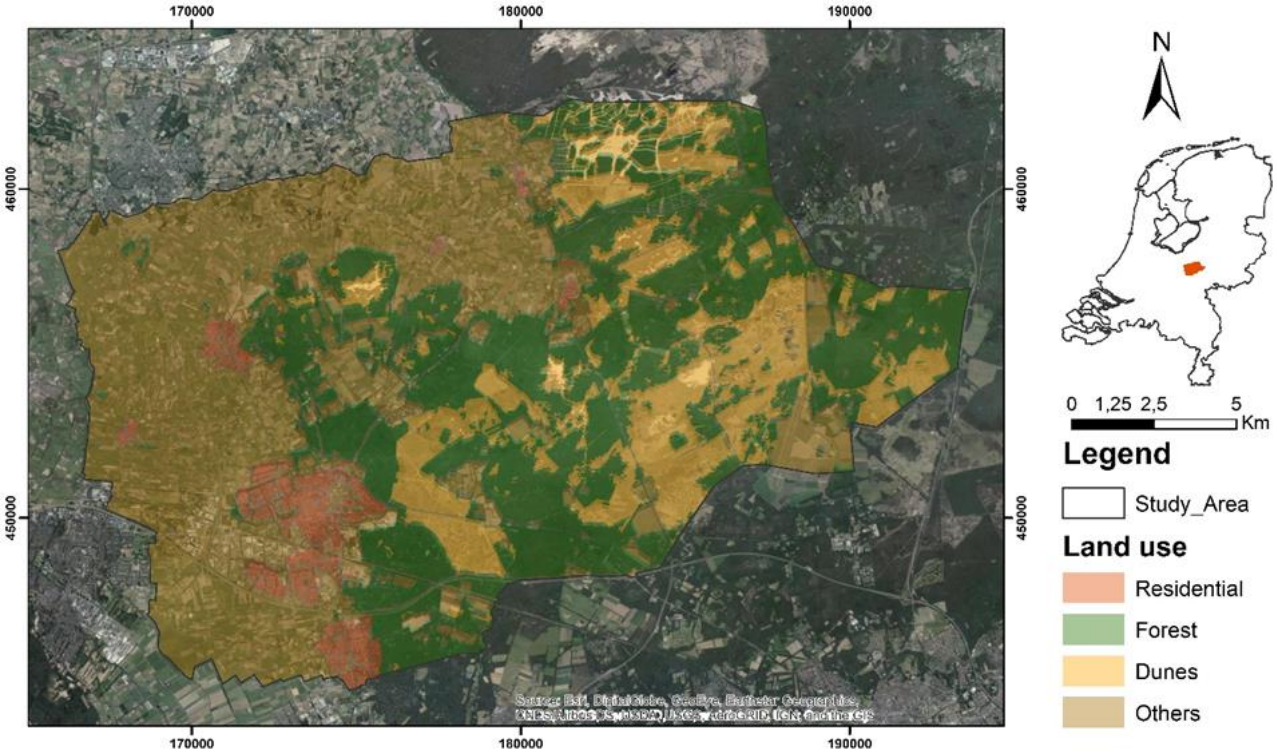

Figure 1 - Study area map of Ede, the Netherlands.

1.2. Research Problem

Every year in the Netherlands, the tick bite count and the number of people being diagnosed with Lyme disease is increasing (Kok, 2019). Scientists are working on mapping and monitoring the number of ticks by performing the blanket dragging technique. High clusters of human exposure are usually found in attractive forest locations as human activity is a major contributing factor (Garcia-Marti et al, 2018). Therefore, it is necessary to analyse the influencing factors to understand the tick activity, which can further be used as a preventive measure.

1.3. Research Objectives

- To create an Agent-Based-Model (ABM) to investigate the impact of tick activity and human activity on tick bites.
- To create a complementary Machine Learning (data-driven) model of tick activity to be used as a pre-processing step for ABM.

1.4. Research Boundaries

1. The study area in this report is Ede, a municipality in the province of Gelderland. (But this model can be adapted into other areas or upscaled to larger areas.)
2. For tick activity, the life stages of ticks will not be simulated. Only the monthly relative tick activity will be predicted by machine learning methods and used as input data in the ABM model. The tick abundance is influenced by weather factors. It is assumed the same across the study area, and only the temporal difference is considered.
3. The movement of people and tick bites will be simulated in ABM. The chance of getting tick bites is determined by tick activity, human vulnerability, and human activity in a certain land use type.
4. Every time step is one day. This model will run for two years, which is from 2015 to 2016.
5. This model will not be used to predict the exactly correct number of tick bites every year. The purpose is to explore the patterns of tick bites and explain the factors that influence the patterns.

2. DATA EXPLORATION AND PREPARATION

Considering the huge amount of data available, exploring the dataset before starting with the processing phase is extremely necessary. It helps to visually present and understand different characteristics of the data.

We started by visual inspection of our dataset (CSV files) and manually removed the missing/null and extreme values. We took into consideration that we are also dealing with 'real data' which was reported by people themselves. Thus, it is very likely to have errors and inexplicable values. Therefore, cleaning the data and preparing it for further analysis was very important.

2.1. Weather Data and Tick Activity

As demonstrated by previous studies (Berger et al, 2014), the level of tick activity is highly correlated with the weather variables such as temperature, relative humidity, precipitation etc. Therefore, to model the tick activity in Ede, the weather was downloaded data from KNMI, selecting the nearest station to our study area i.e. Deelen (275) (Figure 2). Five weather variables were retrieved, namely Maximum Temperature (TX), Minimum Temperature (TN), Precipitation (RH), Relative Humidity (UG) & Evapotranspiration (EV24). Daily data from 2007 to 2014 was used for training the ML regression model, and from 2015 to 2016 for prediction.

Figure 2 - Location of KNMI Meteorological Stations (Source: http://www.bo-mo.si/fispace/)

Moreover, Saturation Deficit (SD) was calculated using the following formula in python (SD_VPD_formula.py) (Randolph & Storey, 1999). The Vapour Pressure Deficit (VPD) was also calculated in the python.

$$SD = \left(1 - \frac{RH}{100}\right) 4.9463 e^{0.0621T} \qquad VPD = 0.611 * (1 - \frac{RH}{100}) * e^{(\frac{17.27*T}{273.3+T})}$$

Then, using SQL queries in pgAdmin (SQL_Query.txt) we calculated the monthly average of the daily weather variables. After generating the final tables, we visually represented our data to further explore and analyse it (DataExploration.ipynb).

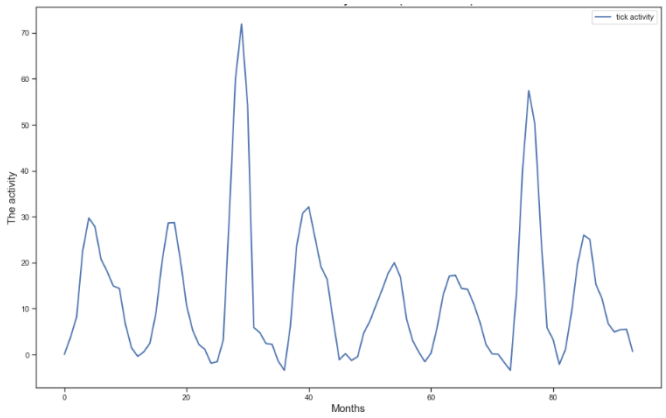

Figure 3 - Line graph for levels of tick activity in Ede (2007-14)

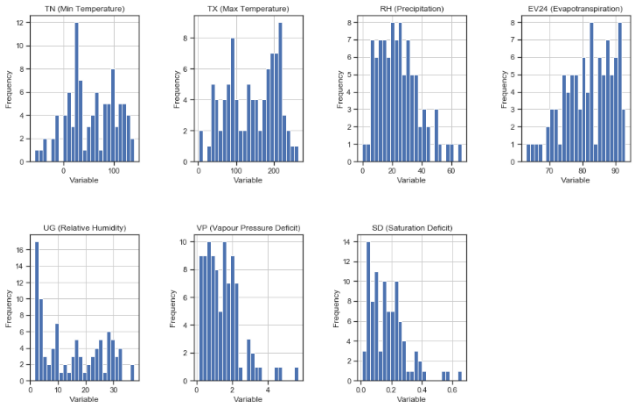

Figure 4 - Histograms for all the Weather Variables.

2.2. Tick bites data

The tick bites data "NL_TickBites_Dec16_RD_New.shp" downloaded from Canvas is collected from voluntary reports on the website of tekenradar.nl. The data shows the reported tick bites in the Netherlands from 2006 to 2016, with information about how and when each tick bite happened, including the date of the tick bite, coordinates, municipality, environment type, activity type, birth year, etc.

To explore the distribution of tick bites among different age groups, activity types and land use types, and apply the parameters into ABM, the relevant fields were extracted from the attribute table of the tick bites map.

2.2.1. Tick bites data cleaning

Since these tick bite records were reported by voluntary individual people, there are many missing values or unrealistic records that need to be cleaned for further exploration. The data was cleaned in Excel. The cleaning processes are as follows.

1. All Dutch was translated into English.
2. The age of people was calculated based on the birth year by: age = YEAR (tick bite date) - birth year. Then people were divided into three age groups: young (0-18), adult (19-64), and elderly (65-).
3. The month when tick bites happened were extracted by: month = MONTH (tick bites date).
4. The records with "unknown" activity or land use type, or NULL values were deleted. The records with the age of people larger than 100 years old were deleted too. After this step, the original 46839 records were reduced to 26604 records.
5. Some records have multiple activity types, such as dog walking-gardening, or walking-dog walking-gardening. Only the first activity was kept in these multi-type records. Eventually, there are 7 activity types, including dog walking, gardening, others, picnic, playing, professional garden maintenance, walking.
6. Some records have multiple land use types, such as forest-heather-dunes, or garden-forest-urban park. Only the first land use type was kept.
7. The land use types where tick bites happened include forest, pasture, heather, others, urban park, dunes, and garden. But in the land use map data available, there are only 4 types: forest, dunes, residential and others. So, pasture and heather were merged into the forest; urban park and garden were merged into residential.

After cleaning, the data consists of 4 fields: month, age group, activity type, and land use type:

1. three age groups: young (0-18), adult (19-64), elderly (65-)

2. seven activity types: walking, gardening, playing, dog walking, picnic, professional garden maintenance, others
3. four land use types: forest, dunes, residential and other.

2.2.2. Tick bites data exploration

The risk of getting a tick bite is defined as Risk = Hazard * Exposure * Vulnerability. Hazard is approximated by the tick activity. Exposure is determined by the activity type of people and land use type where people do this activity. It is assumed in this report that different types of activity in different land use type have different exposure (or relative risk of getting a tick bite). Vulnerability is determined by the characteristics of people, such as the age group, or the awareness of ticks. In this case, the awareness of ticks is considered in the ABM by "to learn" and "to forget" process. In order to approximate vulnerability of each age group, and exposure of every activity type in every land use type, the reported tick bites data were analyzed to extract the parameters.

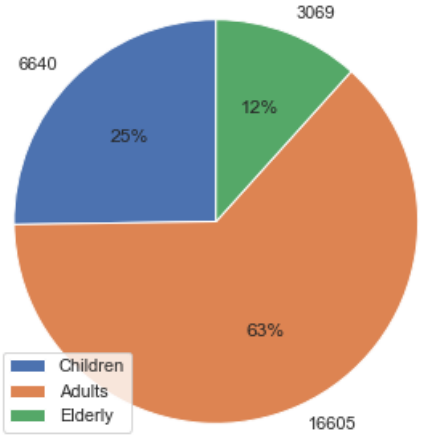

Figure 5 - The distribution of tick bites among different age groups.

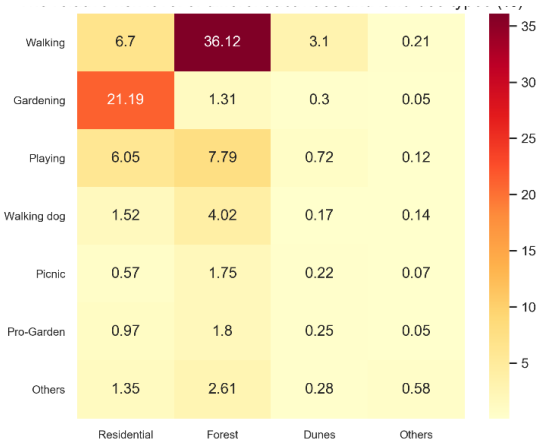

Figure 6 - The relative risk level of different activities and land-use types

Based on the distribution of tick bites among different age groups in Figure 5, the vulnerability was set as young: 25; adult: 63; elderly: 12. The exposure was set according to Figure 6. For instance, the exposure of "walking in the residential" is 6.7, and the exposure of "playing in the forest" is 7.79. The vulnerability and exposure values were used as parameters in ABM.

Besides, this tick bite data was also compared with data from the RIVM to estimate the number of tick bites in reality, because this data is only from people who were willing and had access to report their bites. The biased data was calibrated to get the total number of tick bites during 2015 and 2016 in Ede for further calibration of the ABM model.

In the tick bites data, there are 133 tick bites reported in Ede in 2015 to 2016. And 17814 tick bites reported in the whole Netherlands during 2015 and 2016. However, there are about 1350000 people who get tick bites every year ("Lyme disease RIVM," 2019). This means there are 2700000 tick bites during 2015 and 2016 in the whole Netherlands. Based on the proportion of reported bites in total bites in the whole Netherlands, which is 0.66%, there are 133 /0.66% = 20152 tick bites in Ede during 2015 and 2016.

2.3. Resident population data

As the vulnerability of each age group differs, the age group distribution data is needed for initialization of agents in ABM. Based on the population data from CBS (CBS, 2020), the age group distribution in Ede in 2016 is young: 22%; adult: 61%; elderly: 17%.

2.4. Tourist population data

Tourists are another group of people in the study area who get tick bites. The population of tourists per day and the age group of tourists are needed for the input of Agent-Based Model in NetLogo.

2.4.1. Daily tourist population

In Ede, there are 9 million tourists in total in 2017, consisting of 4570000 tourists who stayed for 2 days, 2870000 who stayed for 5 days, and 1560000 tourists who stayed for 9 days (Kok, 2019). To divide the yearly number into every day, the tourism peak and off seasons were considered. The data of tourists per month is from CBS (CBS, 2020). Here only the data of the province Gelderland is available.

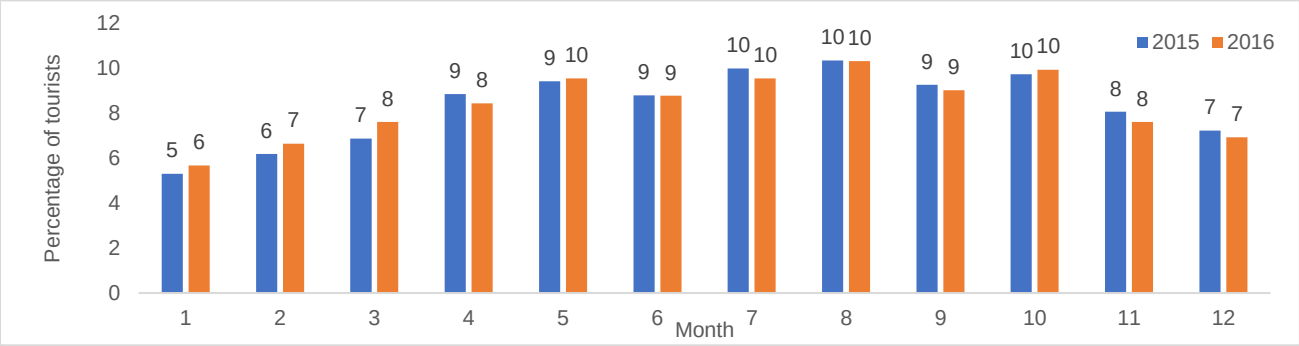

Figure 7 - The percentage of tourists per month in Gelderland in 2015 and 2016.

From April to October, the percentage varies from 0.09 to 0.10. From November to March, the percentage varies from 0.05 to 0.08. Thus, the peak seasons are from April to October, and the off seasons are from November to March. 66% of tourists visit Gelderland in April to October, and the average per month is 9.4%, per day is 0.31%. 34% of tourists visit Gelderland in November to March, and the average per month is 6.8%, per day is 0.23%. The percentage of tourists in peak and off seasons were applied to Ede. Thus, in Ede, in peak season, the number of daily tourists is 27900; in the off season, the number of daily tourists is 20700. The proportion of tourists with 2, 5 and 9 day-stay duration were assumed to be constant, that is 51%, 32%, and 17%.

Note: when this model is upscaled or applied to other areas, the proportion of tourists can be used directly in other provinces in Gelderland. But for other provinces or the whole Netherlands, the seasons or the proportion of tourists may vary.

2.4.2. Age group distribution of tourists

The age group distribution of tourists in Ede is not available, so the distribution in the whole Netherlands was applied here. The percentage of young (0 - 18), adult (19 - 64), elderly (65+) people is 19%, 67%, and 16% respectively (Gelderman, 2011).

3. ML DESCRIPTION

Random forest algorithm can be used effectively for both regression and classification (Breiman, 2001). Random Forest regression reduces the overall biasedness of the algorithm. It can handle missing values and still maintain the accuracy of a large data set. Furthermore, the output variables can be modelled for their importance which is very useful for the analysis. In this project, we used Random Forest Regression algorithm in the pre-processing step to use weather variables to predict the monthly tick activity level, aiming to integrate the ML into ABM, thus generating more accurate and meaningful results.

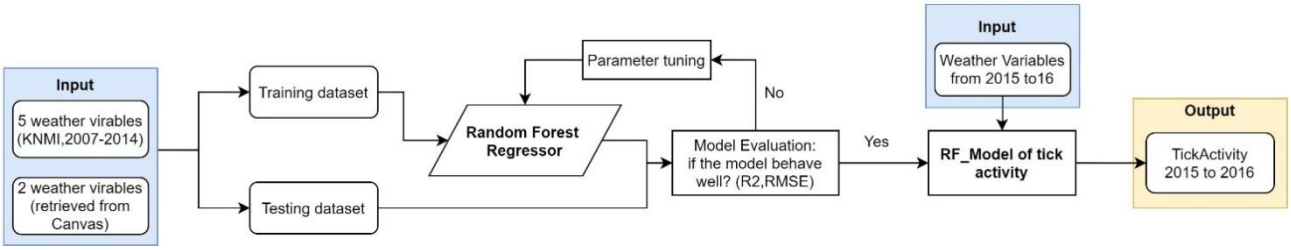

Figure 8 - The flowchart of ML pre-processing.

The flowchart was given in Figure . To make a concrete prediction, based on the previous studies done by (Garcia-Martí et al, 2017), we adopted 7 weather variables (referring to the second section) as inputs to train the RF regression model, retrieved from KNMI and Canvas.

The temporal resolution of all the input data was recalculated and scaled into monthly average, in order to fit the tick bites data. Through data exploration, we found that there are negative values of tick counts data provided on canvas. Thus, a resampling was conducted for the negative ones to convert them into positive. In addition, since we only have the monthly tick activity data available from 2007 to 2014, the size of our training data was actually quite small, with only 96 (12 months * 8 years) sets of rows. We have realized that such small dataset might not be able to sufficiently feed a random forest model, however, given the data availability and time limitation, we still decided to conduct this analysis. The input dataset was, then, split into 2 parts, of which 80% was used for training and the other 20% to validate the model, and the RF Regressor was applied and fitted to the training dataset in Python via Jupyter Notebook. To evaluate the validity of our model, the 20% testing dataset was used to calculate the R square (R) and root-mean-square error (RMSE). Based on the performance of these two values, we iteratively adjusted some key parameters (e.g. n_estimators, min_samples_split) of the RF Regressor to improve our model behaviour. Finally, an RF regression model of monthly tick activity was

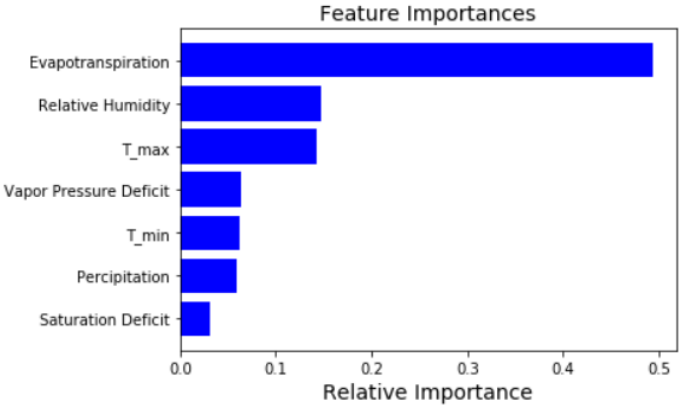

Figure 6 - Relative importance of the weather variables.

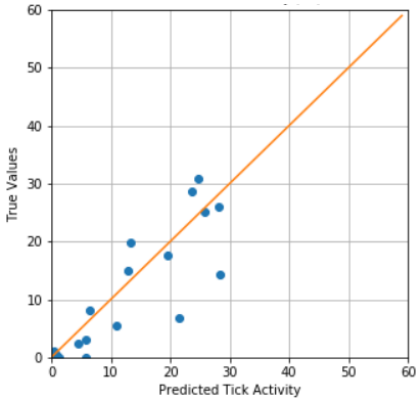

Figure 10 - Cross-plot of predicted tick bites and reality

generated, with the R^2 scored 0.71 and RMSE scored 32.07 respectively (for a more detailed overview, please refer to the file: ML_Pre-processing_TickActivity.ipynb).

Through the parameter-tuning iteration, we found that the overall model performance showed a relatively high degree of instability – i.e. when we tried to re-run our model, the R2 varied, ranging from nearly 0.30 to 0.70, sometimes even reached negative values. In other words, our model highly depends on the splitting of the training and testing dataset. The underlying reason for his, as we speculated, might be: (1) the size of our dataset, since an ML algorithm usually requires a large training data to feed in, whereas we only have less than 100 records, which are unable to provide enough variation; (2) the number of dependent variables, as we only took 7 weather variables into account, the activity level of ticks was only explained and predicted by one dimension.

Based on our analysis, the most significant variable in determining the tick activity is the evapotranspiration, explaining almost 50% of all the variations, followed by relative humidity and the maximum temperature (Figure 6). As for the rest variables, they are considered less important, with all relative importance values less than 0.1. Figure 10 presented the comparison between the model output and the real data. Though two outliers appeared in the lower middle, we still argue that our RF model is acceptable in predicting tick activity, considering the input data quality. Compared to the RMSE value of the RF model in geo-regression exercise (we also calculated that and the value was around 150), our result reduced almost 70%, further confirming the improvement of our model.

4. ABM DESCRIPTION

4.1. Purpose

This tick bites ABM aims to simulate the tick bites in the study area. The goal of this simulation is to investigate the impact of tick dynamics and human activity on the chances of getting a tick bite. This model is expected to reproduce the spatial and temporal patterns of tick bites, such as the peak and off seasons of tick bites, and spatial variety in different land use types. Our model has been designed based on certain boundaries, as discussed in section 1.4.

This model will not be used to predict the precise number of ticks bites every year. The purpose is to explore the patterns of tick bites and explain the factors that influence the patterns.

4.2. Entities, State Variables and Scales

4.2.1. Entities

1. Agent

This simulation has two breeds of agents: residents, and tourists. Residents stay in the study area, while tourists come and stay for a few days and leave this area. The goal of the residents and tourists is to do different outdoor and indoor activities based on the condition of holidays and their preferences and then move to the relevant land use type. The agent will get a tick bite if their risk is relatively high.

2. Environment

The environment in this simulation is the study area with 4 types of land use and the shape of the boundary. The number of tick bites happened on every patch is also considered. Daily weather condition (precipitation) and monthly tick activity are considered as global variables in the world.

3. Time

The timestep in this simulation is per day, and the length of the simulated period is two years, that is 720 days. The time limit can be changed according to the purposes of further exploration.

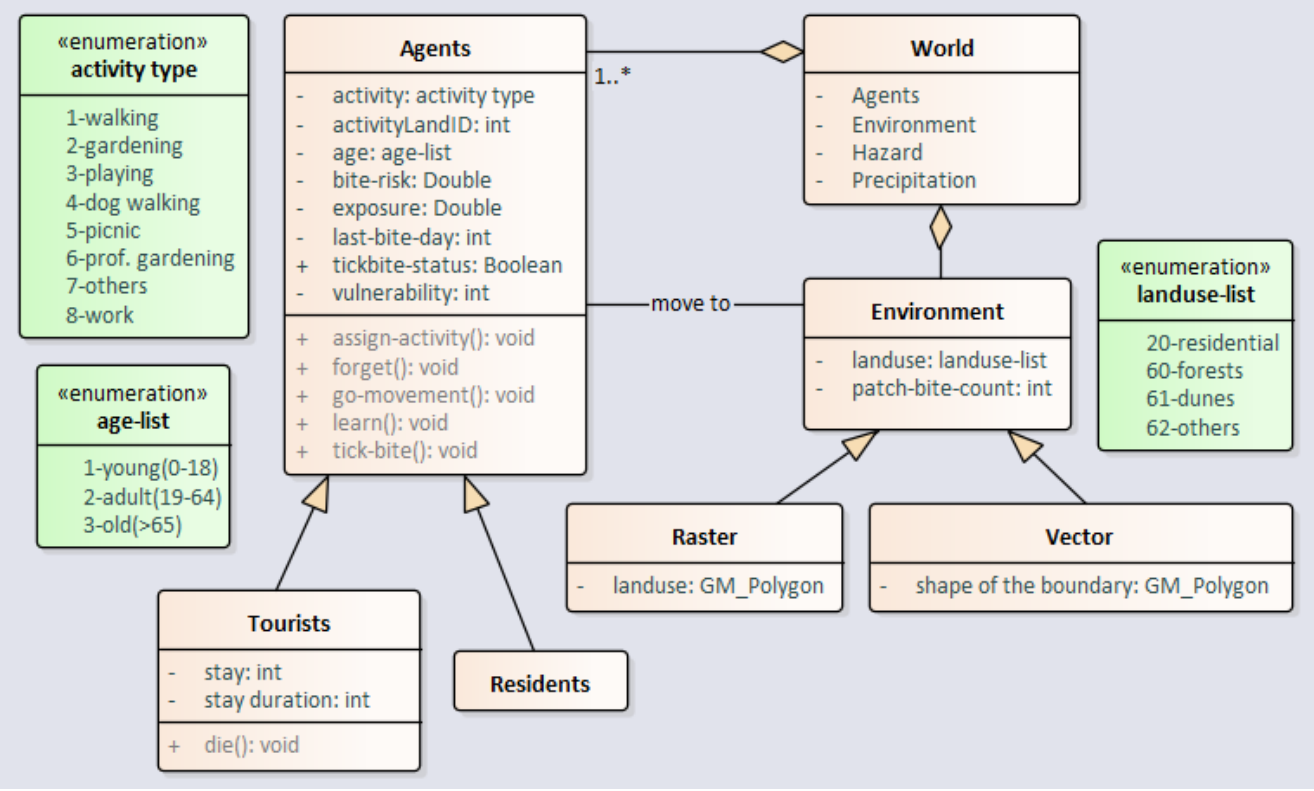

Figure 11 - The UML class diagram.

4.2.2. State Variables

For agents, the state variables are: age group (1 young, 2 adults, 3 elderly), activity (1 walking, 2 gardening, 3 playing, 4 dog walking, 5 picnic, 6 professional garden maintenance, 7 others, 8 indoors), activitylandID (201 walking in residential, 202 gardening in residential, 604, dog walking in forest, etc.), and exposure (6.7 for walking in residential, 21.19 for gardening in residential, 0 for indoors in residential, etc.).

For the environment, the state variable is the land use type (20: residential; 60: forest; 61: dunes; 62: others).

4.2.3. Scales

For the temporal scale, the total duration of the simulated period is 720 days (the year 2015 and 2016), and the length of each time step is one day. For the spatial scale, the total map extent is the Ede municipality within the boundary, and the resolution is 5*5 pixels, which is 250m*250m.

4.3. Process overview and scheduling

The behaviour of the agents: both residents and tourists can do activities; they can move to relevant land use type; they can get a tick bite; they can learn to be more aware of ticks; they can forget to be aware of ticks; residents can move back home; tourists can leave the study area.

The interactions between agents and environment: the agents can move to relevant land use type and do activities there; the tourists know when they leave the study area; the number of tick bites of agents is cumulated in relevant land where they happened. The sequence diagram is as follows.

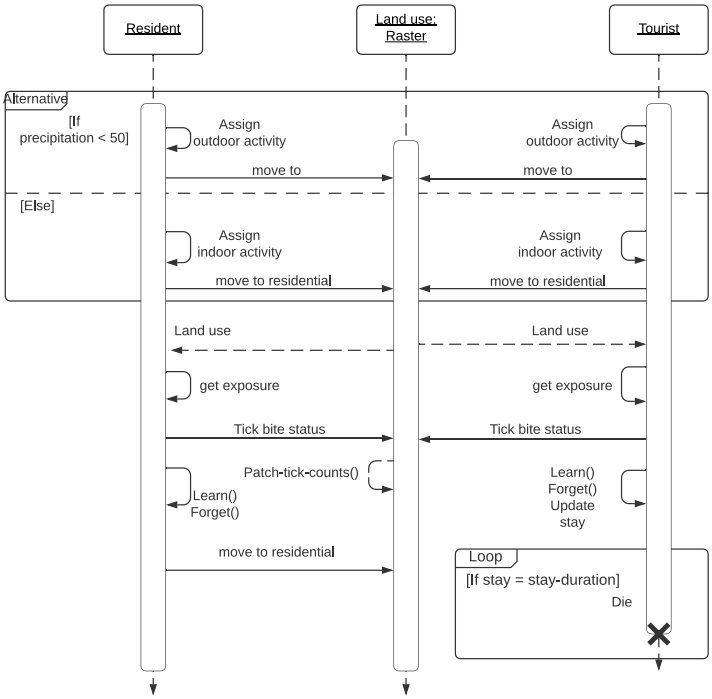

Figure 12 – The UML sequence diagram.

4.4. Design concepts

4.4.1. Basic theory

This simulation of tick bites is based on the formula of risk (R): R = Hazard (H) * Exposure (E) * Vulnerability (V). Hazard is approximated by the tick activity. Exposure is determined by the activity type of people and land use type where people do this activity. It is assumed that different types of activity in different land use type have different exposure (or relative risk of getting a tick bite). Vulnerability is determined by the characteristics of people, including the age group, and the awareness of ticks.

4.4.2. Emergence

The patterns that emerge include temporal patterns and spatial patterns of tick bites. For the number daily new tick bites when it is plotted, it shows clearly peak and off-seasons. The daily tick bites increase in spring (April), reach the peak in summer (June and July), and decrease in autumn (September and October), and then reach the lowest in winter (November to January). Different locations with different land use type have the different total number of tick bites. Forests have the highest cumulative tick bites.

4.4.3. Adaptation

Agents adapt their activities and movements according to the weather conditions. When it rains hard, all agents will stay indoors. Agents also change their own awareness of ticks in response to their tick bite status. When

they get tick bites, they will learn to be more aware of ticks, and get lower vulnerability. But the learning outcome expires when they have not gotten tick bites for a long time since the last one. They will forget to be careful about ticks, and their vulnerability will increase again.

4.4.4. Objectives

The agents adapt to their tick bite status in order to avoid getting more tick bites. Their objective is to do any activities they want and avoid getting tick bites at the same time. How they avoid getting tick bites is measured by their vulnerability to tick bites.

4.4.5. Learning

The agents repeat doing activities and move to relevant land use type every day, and they learn to be more aware of ticks when they get tick bites. To avoid getting tick bites, the agents become more careful and get lower vulnerability.

4.4.6. Prediction

In this simulation, agents do not predict future conditions. But if the user wants to extend the model and make it more accurate, we can ask agents to realize the temporal pattern of tick bites, and be more aware of tick bites in summer, or realize the spatial pattern of tick bites, like being more aware of tick bites when they go to the forest.

4.4.7. Sensing

Agents sense the weather of the environment and decide their activities. When it rains hard, agents only do an indoor activity instead. Agents also sense the land use type of environment where they do the activities.

4.4.8. Interaction

In this simulation, agents do not interact with each other. Agents change the tick bite count of the environment. When the agents get tick bites, the tick bite count will accumulate.

4.4.9. Stochasticity

At initialization step of new agents, the age, stay duration, and initial location are assigned randomly under certain conditions. The whole population and stay duration keep a certain distribution, and their locations depend on their breed and activity type, but within each group or type, the attributes and locations are set randomly. The tick bite status of agents is also randomly decided but under a certain probability.

4.4.10. Collectives

In this simulation, the agents do not form together as collectives.

4.4.11. Observation

At the level of the individual agent, their age, activity type and land use type of each movement, tick bite status are recorded to analyse the age distribution and activity exposure of tick bites.

At the level of environment, the count of tick bites that happen on every patch is recorded to understand the spatial pattern of tick bites.

At the global level, the daily new tick bites, total tick bites are recorded to understand the temporal pattern of tick bites. The hazard and precipitation value are also recorded to check if the data import is successful. The number of agents in each age group can also be recorded to test the age distribution of agents.

4.4.12. Initialization

The residents are created at the start of the simulation. The number of residents depends on the study area, which in this case, is 112410 in total in Ede. The age group of residents are assigned according to the age group distribution in the study area, which in Ede is 22% young, 61% adult, and 17% elderly.

The new tourists are created every day at the start of every time step during the simulation. As estimated in section 2.4, in the peak season from April to October, 27900 new tourists are created, while in the off-season from November to March, 20700 new tourists are created. The percentage of age group young (0 - 18), adult

(19 - 64), elderly (65+) are 19%, 67%, and 16% respectively. The proportion of tourists with 2, 5 and 9 day-stay duration are 51%, 32%, and 17% respectively.

The vulnerability of agents is set according to the age group. The tick bite status of all agents is set as False at the start of the day every day.

4.5. Input data

The input data needed in this model consists of:

1. Spatial data: land use map; study area boundary map.
2. Hazard data: the monthly tick activity data.
3. Temperature data: the daily precipitation data.
4. Agent initialization: the number of residents, percentage of young and adult age group; the number of new tourists every day in peak and off seasons, the percentage of 2-day and 5-day stay duration.
5. Parameters in simulation: vulnerability of each age group, all activity types, exposure of each activity in each land use type, and the probability of getting a tick bite.

4.6. Sub-models

4.6.1. tick activity model

The tick activity model is implemented in Machine learning as the pre-processing step for ABM. The details are described in section 3 ML description.

4.6.2. human population model

The human population is the input for initialization of residents and tourists, including the number of people, the age group distribution, and the stay duration of tourists. This model is explained in section 2 data exploration and preparation: 2.3 resident population data and 2.4 tourist population data.

4.6.3. human activity model

The activity of agents is assigned based on the precipitation, breed and age group. When it rains all agents stay where they are and do the indoor activity. On weekdays all residents do indoor activity in the residential areas.

In other conditions, the residents get one of the outdoor activities randomly, while the tourists get one of the outdoor activities except for dog walking, gardening, professional garden maintenance and others. And they move to any land use type randomly, except that people who do gardening move to the residential area.

4.6.4. tick bite model

The chance of people getting tick bites is determined by: Risk = Hazard * Exposure * Vulnerability. As described above, Hazard is approximated by the tick activity. Exposure is calculated by the activity type of people and land use type where people do this activity. Vulnerability is determined by the characteristics of people, including the age group, and the awareness of ticks. First, the risk determines the probability of people getting a tick bite, and then another general probability is set to give the agent a tick bite. For example, if the risk of an agent is 13, and the general probability is 0.005, then there is a 13% * 0.005 probability that this agent gets a tick bite. A random number ranging from 0 to 99 in ABM is generated and compared with the risk 13. If it is smaller than 13, then another random number ranging from 0 to 1000 is generated and compared with 5. If it is smaller than 5, then this agent gets a tick bite.

4.7. Calibration

The calibration step is mainly about the general probability. As described above in section 2.2, the true number of tick bites in Ede during 2015 and 2016 is around 20151. After experimenting on the probability and running the model for a few times, the probability is set as 0.005 in Ede. This probability produces nearly 20000 tick bites in one simulation.

5. RESULTS DISCUSSION

The results of the tick bite ABM mainly include two outputs: one is a raster map which records the number of total tick bites per patch in Ede, the other is a text file containing the information of each individual agent and its location as follows: month, age, activity, land-use. To generate more in-depth and comprehensive insights towards the results, two files were analysed separately.

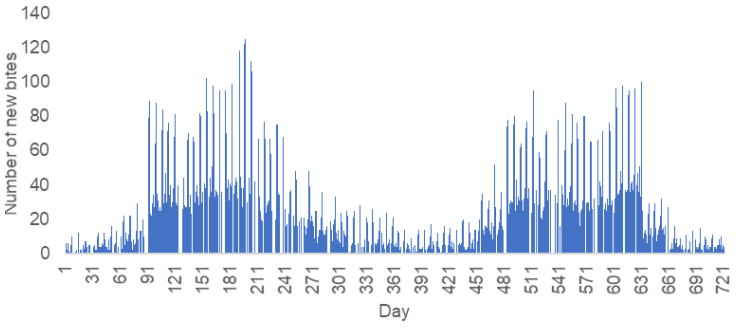

Figure 13 – Model output: the number of new tick bites per day

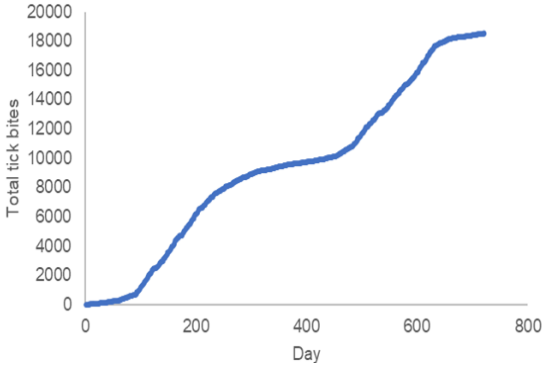

Figure 14 - The number of total tick bites per day.

The two figures above present the information of tick bites (total and daily new numbers) per day as predicted by the ABM model (Figure 13 & 14**Error! Reference source not found.**).

5.1. Spatial Analysis

The map below shows the distribution of total tick bites per patch. In order to detect the degree of spatial autocorrelation (Moran's I) in terms of the tick bites, we created a fishnet of grid cells (which have the same size as the raster cell) and transferred all the raster values to the new vector grids in ArcGIS. After this, the spatial autocorrelation analysis and hotspots analysis were performed.

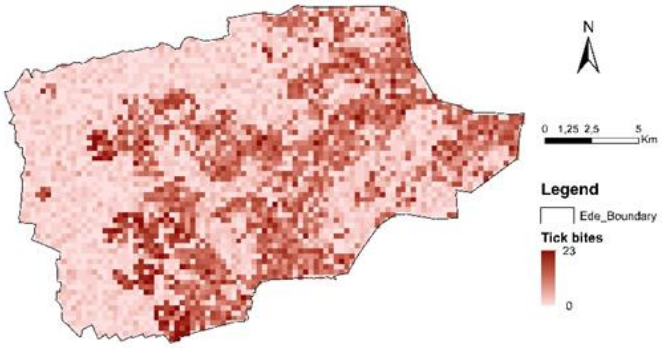

Figure 15 - The distribution of predicted tick bites in Ede.

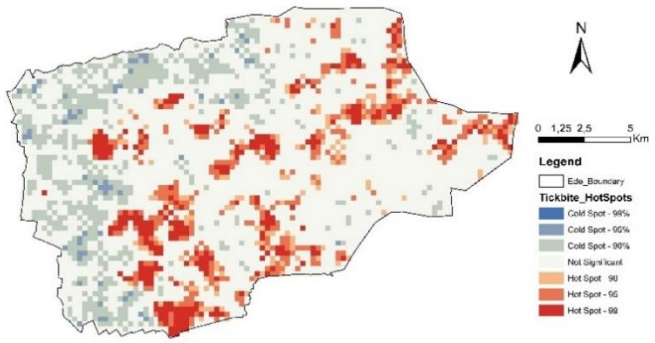

Figure 16 - The hotspot analysis of predicted tick bites.

The Moran's I of tick bites is 0.442, with a p-value < 0.001, indicating that there is a moderately positive spatial autocorrelation. Furthermore, the hotspot analysis shows the distribution of the clustering of high tick bites values within Ede. In general, the hotspots of tick bites mainly locate in residential and forest areas (especially the middle-lower part, which mainly composes the residential areas), while in dunes and other land-use types the tick bites distribution is relatively random or dispersed. It might be a bit bizarre to link residential areas with the hotspot of tick bites at first. However, this could be explained by the fact that residents usually gardening in their yards and the relative risk of getting tick bites is also high according to the reported data, besides the area of residential land is quite small compared to others.

5.2. Explanatory Data Analysis

For the analysis of the text file, since it records every agent's activity information per day for the whole 2 years, the size of this data is large, accounting for nearly 2.4 GB, of which the vast majority is non-tick-bites. Therefore, to avoid data abundance, we only take the rows where tick bite status is true by using python code (true_values.py). Then, an explanatory data analysis was conducted for this extracted data.

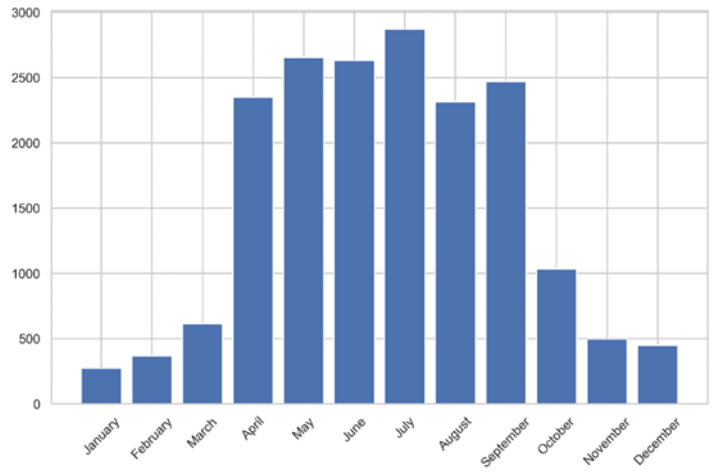

Figure 17 - The distribution of predicted tick bites (summarized by month) between 2015 and 2016.

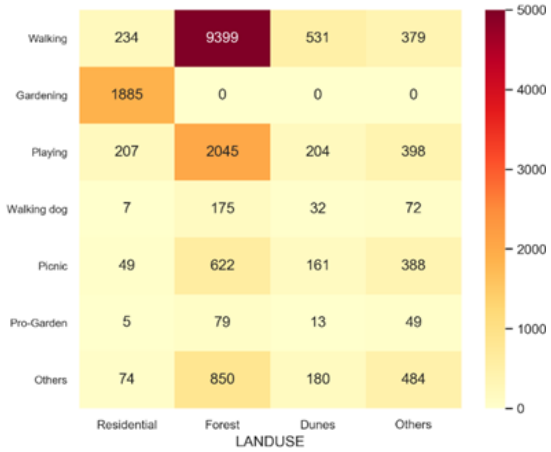

Figure 18 - The distribution of predicted tick bites in different land use types and activities.

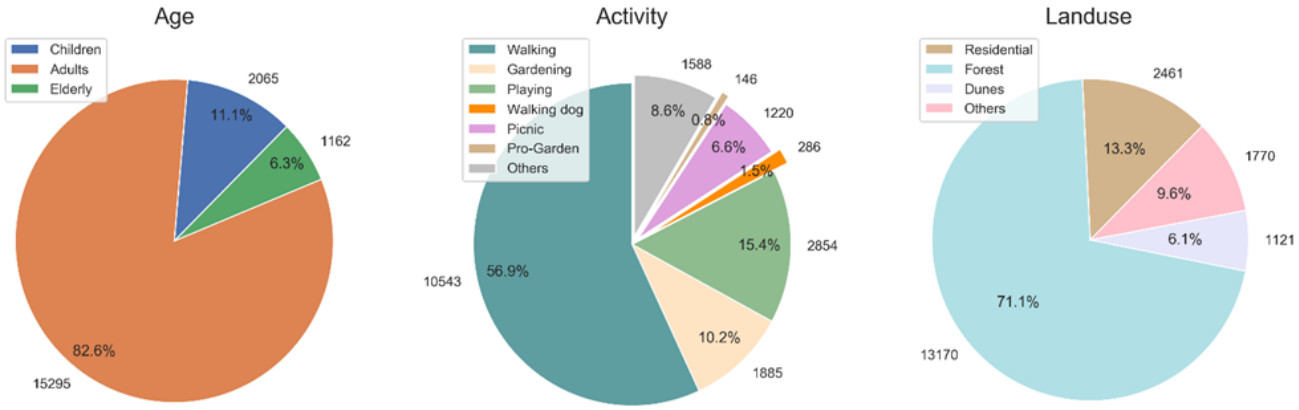

Figure 19 - The percentage/number of tick bites in different groups

With regards to the temporal distribution of tick bites, the peak seasons are mainly late spring and summer (Figure 17), which meets our assumptions of the ABM. In summer (Figure), with more tourists visiting Ede and the tick dynamics also arrived at the highest level within a year, the number of tick bites rapidly increased.

The pie charts (Figure) illustrate the predicted tick bite percentage/number among different dimensions – age groups, activities and land use types. In terms of age, 82.6% of tick bites occur in the adults, while only 11.1% and 6.3% happened in children and the elderly respectively. This huge disparity might result from the age-scale we set for each age group, where the adults include all the people from 18 to 65 years old and the fact that we assigned the population of different age groups according to the census data. Regarding activity and tick bites, we could clearly observe that people who walking have the highest risk level, accounting for more than half of the total tick bite events. Playing (15.4%) and gardening (10.2%) are identified as the second risky level, whereas it is not common to get tick bites during walking dog or professional garden maintenance. As for the distribution of tick bites in different land use, forest shows the dominance with around 70% of the total tick bites, and the second happened to be the residential areas (mainly because of gardening). Referring to the spatial analysis section above, despite residential areas were detected as the major hotspots, the forest was still predicted as the place with the largest absolute number of tick bites, which satisfies our model assumption and reality.

The heatmap (Figure) further explores the tick bites distribution across land use and activity. Apparently, in our model, walking in the forest was considered as the most dangerous activity, with 9399 cases predicted. Playing at forest and gardening in residential areas are also in a relatively high-risk level, with 2045 cases and 1885 cases, respectively. Overall, doing activities in forest and gardening are most prone to tick bites.

6. CONCLUSIONS

6.1. Integration of ABM and ML

The integration of ABM and ML was applied in the pre-processing phase, where tick activity is predicted by ML, and then used as the input data of ABM. As the proxy of hazard, tick activity data plays a vital role in the calculation of tick bite risk R = Hazard * Exposure * Vulnerability in ABM. The tick activity varies in space and time and is dependent on the environment. However, the tick activity data in our study period is missing. Thus, the unavailability of data is the motivation of our integration.

To solve this problem, the tick collection data of the year 2007 and 2014 was explored as the training dataset in ML regression to predict tick activity in the year 2015 and 2016. The predicted monthly tick activity was used as input in ABM for the hazard variable. Therefore, due to ML, the ABM has the datasets for input which would not be available otherwise.

However, although the integration of ABM with ML bridges the gap of poor data to simulations, it still has some limitations. The implementation of the ML regression algorithm requires reliable and extensive data input, but only monthly volunteered blanket dragging tick collection data is available. The tick collection only happened in forested areas, which made it impossible to model and predict the spatial variations of tick activity. Besides, tick activity is influenced by not only weather variables but also vegetation and land cover. But only weather conditions are considered as explanatory variables because other data is not available. The ideal and more accurate integration of ML and ABM would be predicting daily tick activity data that also varies among different land use types, which is inadequate in this study.

Moreover, the predicted tick activity in ML is just the number of ticks in one certain plot field. But when imported into ABM, it is directly used as a proxy of tick activity. This could be improved with more tick expert knowledge. Also, more experiments are needed to decide which ML algorithm suits the model better.

6.2. Limitation reflection

(1) Data quality. Overall, almost all the inputs for the ML and ABM models suffer from insufficient data quantity and quality, leading to the inconsistency of parameter-setting and lack of explanation of our model. For example, the assignment of age and population for the tourists was based on a national scale instead of more detailed data (due to data availability). Also, the model calibration is only based on the reported cases of tick bites in Ede. However, there are of course more tick bites occur in reality, which might lead to an underestimation of our prediction.

(2) Transferability. As our model is developed based on the data of Ede, it is then contextualised and tailored to fit the characteristics of tick bites in this study region particularly. For instance, Ede is a densely forested municipality, the tick bites show a high degree of clustering in the forest areas. However, this spatial pattern might differ substantially in other cities (e.g. coastal areas) Thus, we argue that when applying this model to another region or up-scaling it to the whole Netherlands, necessary adjustments need to be made.

(3) More elements could have been incorporated in the model like blood type of agents, sex-based vulnerability, etc. to make it more complex and accurate. Some literature was available about these parameters but due to its limited expanse, it was not used as an input for the ABM.

REFERENCES

Berger, K. A., Ginsberg, H. S., Dugas, K. D., Hamel, L. H., & Mather, T. N. (2014). Adverse moisture events predict seasonal abundance of Lyme disease vector ticks (Ixodes scapularis). *Parasites and Vectors*, *7*(1), 181. https://doi.org/10.1186/1756-3305-7-181

Breiman, L. (2001). Random forests. *Machine Learning*, *45*(1), 5–32. https://doi.org/10.1023/A:1010933404324

CBS: StatLine - Population on January 1 and average; gender, age and region. (2020). Retrieved from https://opendata.cbs.nl/statline/#/CBS/nl/dataset/03759ned/table?ts=1591732275230

Garcia-Marti, I., Zurita-Milla, R., Harms, M. G., & Swart, A. (2018). Using volunteered observations to map human exposure to ticks. *Scientific Reports*, *8*(1), 1–10. https://doi.org/10.1038/s41598-018-33900-2

Garcia-Martí, I., Zurita-Milla, R., van Vliet, A. J. H., & Takken, W. (2017). Modelling and mapping tick dynamics using volunteered observations. *International Journal of Health Geographics*, *16*(1), 41. https://doi.org/10.1186/s12942-017-0114-8

Gelderman, C. (2011). *Feiten en cijfersoOver de vrijetijdssector in Overijssel 2011*. https://doi.org/10.1007/978-90-313-6623-1_2

Kok, J. (2019). *Simulating tick bites in the Netherlands using agent-based modelling*. (March). Retrieved from http://dspace.library.uu.nl/bitstream/handle/1874/394866/Final_thesis_Jasmijn_Kok.pdf?sequence=1&isAllowed=y

Lyme disease RIVM. (2019). Retrieved from https://www.rivm.nl/ziekte-van-lyme

Randolph, S. E., & Storey, K. (1999). Impact of Microclimate on Immature Tick-Rodent Host Interactions (Acari: Ixodidae): Implications for Parasite Transmission. *Journal of Medical Entomology*, *36*(6), 741–748. https://doi.org/10.1093/JMEDENT/36.6.741

APPENDIX

The whole project was conducted by the collaboration of all the group members. The conceptualization of research problems, objectives and the logic flows of this project were discussed and established by the members together. Then, the project was divided into different tasksets for which each member was responsible. Regular meetings were held for communication and feedbacks and each one provided help whenever required.

The detailed contribution of each group member is as follows:

Anam Akhtar – Data exploration and cleaning using SQL, ML pre-processing & Overall analysis.

Eqi Luo - Data exploration using Jupyter Notebook, ML visualizations & ML pre-processing.

Rhea Singh Chib - Data exploration and cleaning using PyCharm, UML & ML pre-processing.

Zijing Wu – Tick bites data, human population data exploration and preparation; ABM model programming; ABM description, UML.

The task of report writing was divided among the group members equally.