Geospatial data mining in volunteer data: how natural conditions might increase the risk of tick bites and Lyme disease?

I. Garcia-Mart í[1], R. Zurita-Milla [1], S. Bennema[2], M. G. Harms [2],
C. C. van den Wijngaard [2], A. Swart [2], A. J. H. van Vliet[3]

[1]Department of Geo-Information Processing, ITC
University of Twente
PO Box 217, 7500 AE, Enschede, the Netherlands
Email: i.garciamarti@utwente.nl

[2] Centre for Infectious Disease Control
Netherlands National Institute for Public Health and the Environment (RIVM)
Bilthoven, the Netherlands

[3]Environmental Systems Analysis Group
Wageningen University
Wageningen, the Netherlands

Abstract

There is evidence of an increase of tick bites and Lyme disease in the Netherlands since 1994. Scientists of different disciplines have demonstrated the tight bond between natural conditions and the abundance of ticks in forests. However, traditional statistical models are too restrictive to model tick dynamics, as they require a significant number of predictors. In this work, we combine a tick bite dataset collected by citizens between 2006-2012, satellite-derived vegetation indices and weather data to advance our understanding of the impact of environmental conditions on tick bites reports and assess the value of volunteered information to model tick bites dynamics. This data was analysed using clustering and frequent pattern mining algorithms. Results show that this approach seems promising to identify environmental conditions that may be linked to a higher risk to get a tick bite.

Keywords: Spatio-temporal analytics, data-driven science, VGI, public health, citizen science.

1. Introduction

Since 1994 the incidence of tick bites and Lyme disease in humans has increased dramatically in the Netherlands (Hofhuis et al. 2015). The cause of this increase remains unknown, although a link with growing tick populations in deciduous forests has been suggested (Sprong et al. 2012). Modelling the seasonal incidence of tick bites and Lyme disease, and identification of high risk areas is complex due to the wide range of factors intervening simultaneously. Previous efforts in the field of biology or environmental modelling have determined a tight relationship between wildlife, environment, weather and tick populations (Medlock et al. 2013). However, for forecasting ticks and Lyme disease, traditional statistical models are too restrictive as they typically require a significant number of explanatory variables, constraining their applicability (Fink et al. 2010).

To improve the monitoring of tick bites and Lyme disease, Wageningen University and the Dutch National Institute for Public Health and the Environment have been collecting tick bite reports through the platforms *Natuurkalender* (NK) and *Tekenradar* (TR) since 2006. Currently, this collection of tick bites contains nearly 35.000 observations contributed by volunteers. To our knowledge, it is the first Citizen Science project in the world to monitor the occurrence of tick bites.

In this ongoing study we continue the effort of modelling tick dynamics, however, we are using data mining algorithms to find clusters and frequent patterns in tick bites as reported by citizens. Data mining algorithms are promising in the process of modeling tick dynamics, as they focus on finding patterns and correlations without preconceived ideas on their manifestation. These algorithms are capable of working with multivariate datasets and finding high-dimensional samples similar between them. The aim of this study is to determine if using a combination of clustering and pattern mining algorithms, with remote sensing products and volunteer information provides enough information for a spatiotemporal analysis, capable of determining suitable conditions for the occurrence of tick bites and provide additional clues about tick dynamics.

2. Data

The tick bites dataset is a crowdsourced collection of point feature observations from the period 2006-2014. Each observation contains the tick bite date, the location of the tick bite, the (approximated) personal address of the volunteer, the type of environment they were in (forest, bushes) and the activity they were carrying out when they got the tick bite (camping, gardening).

Ticks are sensitive to environmental conditions, such as thickness of forest canopy or soil moisture at the litter level (Medlock et al. 2013). In the past, tick populations have been modeled by combining weather variables with satellite-derived vegetation indices. *Normalized Difference Vegetation Index* (NDVI) has typically been used beyond the vegetation scope as a proxy to tick populations (Estrada-Peña 2001). Ticks are also sensitive to *Enhanced Vegetation Index* (EVI) (Estrada-Peña et al. 2011) and recent studies suggest that *Normalized Difference Water Index (*NDWI), which measures the water content of vegetation, might outperform NDVI for tick populations modelling (Barrios González 2013). All these indices are available in the *Google Earth Engine*[1],[2] (GEE) platform. GEE is an image processing cloud platform for environmental analysis, which aggregates products coming from different sensors, like *Moderate-Resolution Imaging Spectroradiometer* (MODIS). For this study we used NDVI, EVI and NDWI indices for the 2006 – 2012 period derived from MCD43A4 MODIS Surface Reflectance composites at a 500m spatial resolution and 16-day temporal resolution.

Apart from these indices we included temperature data from 2006-2012 provided by the *Royal Netherlands Meteorological Institute* (KNMI) in this analysis, as temperature determines the start of the questing season or the survival probability through winter (Ogden et al. 2006). Weather data for 2013 and 2014 are not yet available and therefore this study was carried out for the 2006-2012 period. When these data become available the study will be extended.

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

[2] https://earthengine.google.org/#intro

3. Methods

The tick bites datasets from NK and TR were cleaned, standardized and merged together in a single collection of point features. Using GEE we obtained the three vegetation indices (NDVI, EVI and NDWI) corresponding to each of the tick bite observations. After this, a process of feature engineering was performed to obtain additional features for various temperature measures. We thus obtained 1) the minimum (TMIN) and maximum (TMAX) temperatures for the day of the tick bite, 2) the average (AVGK), minimum (AVGN) and (AVGX) weekly temperatures and 3) the accumulated temperature in the previous winter (ACCW), the accumulated temperature until the date of the tick bite (ACCD) and the sum of both accumulations (ACC).

Once the multidimensional table was built, two data mining techniques were applied in parallel to know which environmental conditions are more risky for humans for tick bites and evaluate the significance of the features used. First, we clustered the data using Self-Organizing Maps (SOM) (Kohonen 1982) to find out the different groups of similar observations hidden in data as well as to obtain their main characteristics. Then frequent pattern mining with the Apriori algorithm (Agrawal & Srikant 1994) was used to find frequent patterns and to relate them with the output of the clustering. This analysis was done in Python using: SOMpy[3] for the clustering and PyFIM[4] for the frequent patterns.

4. Preliminary results

The SOM produced 8 clusters that aggregate a total of 13.875 tick bites observations for the period 2006 – 2012. Table 1 depicts the centroid of each cluster.

Cluster	NDVI	EVI	NDWI	TMIN	TMAX	AVGN	AVGX	AVGK	ACCW	ACCD	ACC	#OBS
0	0.40	0.63	0.04	12	23	8	23	8	514	511	1026	2146
1	0.31	0.57	0.05	4	13	2	14	7	481	573	1055	1134
2	0.42	0.64	0.03	8	18	5	22	9	460	1253	1714	2057
3	0.43	0.66	0.03	12	22	10	30	11	424	784	1208	1003
4	0.47	0.68	0.04	12	22	8	27	10	535	719	1255	2742
5	0.35	0.59	0.05	5	15	3	22	10	540	975	1516	1216
6	0.45	0.68	0.03	11	20	8	22	7	539	756	1296	2623
7	0.38	0.61	0.04	9	20	5	22	9	446	1082	1529	954

Table 1. Centroids list for the clusters obtained with SOM.

There are several interesting features in Table 1; here we highlight two of them: 1) the low accumulation of previous winter temperature, ACCW, in *Clusters 1, 3* and *7* seems to have a negative impact on the occurrence of tick bites. This may suggest cold winters decrease the survival rate of ticks, causing a lower incidence of tick bites on the next year. 2) *Cluster 2* has the highest ACCD accumulation, meaning that these tick bites occurred in an advanced stage of the year, possibly in line with the late-summer peak activity for adult ticks. Fig. 1 shows the geographical projection of the clusters, which show a high degree of spatial randomness. However, there is a clear spatial clustering around forests (center of the map) and natural recreational areas along the coast.

[3] https://github.com/sevamoo/SOMPY
[4] http://www.borgelt.net/pyfim.html

Clustering techniques do not reveal the most frequent patterns occurring in the data. The application of the Apriori algorithm produced more than 3 million frequent patterns, ranging from two (most frequent) to eleven features (less frequent). For instance, there are 2817 observations with *TMAX = 20* and *NDWI = 0.1* and only 321 observations with *NDVI = 0.5*, *EVI = 0.8*, *NDWI = 0.1*, *ACCD = 440*.

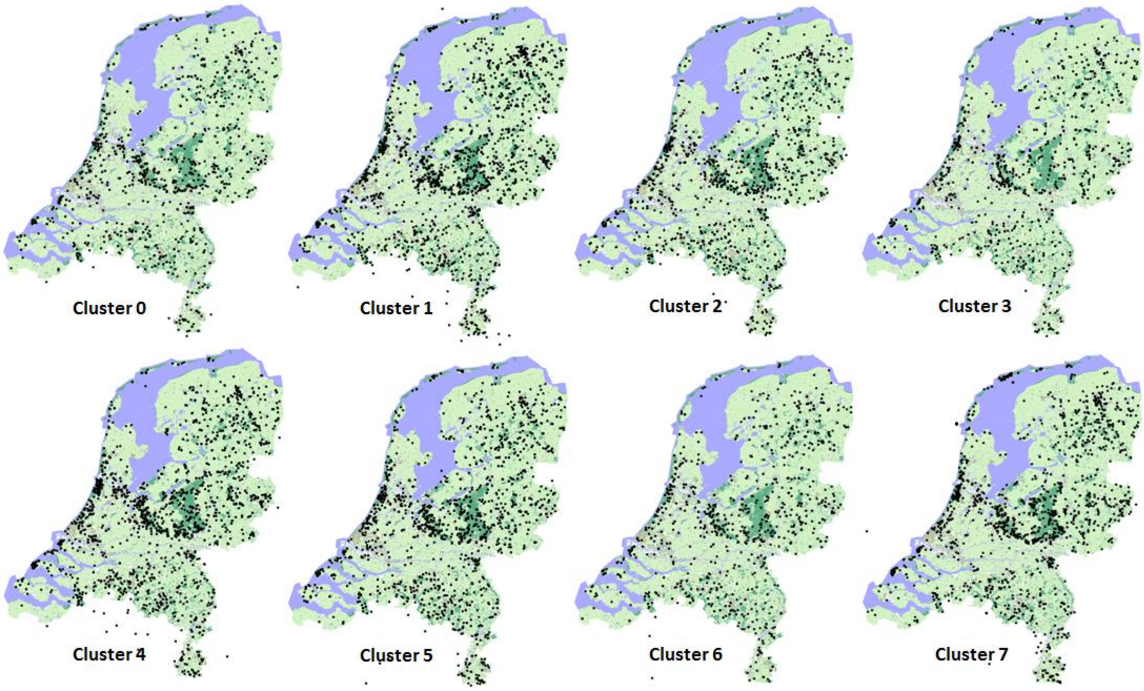

Figure 1. Geographic projection for the observations per cluster.
Clusters (0-7) are arranged left to right.

In summary, these preliminary results show that the application of clustering and frequent pattern mining seem promising to identify environmental conditions linked to a higher incidence of tick bites in humans. Mapping the spatio-temporal occurrence of these conditions could be used to guide further efforts to mitigate tick bites and to try reduce the incidence of Lyme disease. We studied this phenomena at a fine temporal scale (daily, weekly), however, the similarity of the results may suggest that further research should be performed at a coarser temporal scale (e.g. seasonal) as this might be a better way to characterize the locations where tick bites occurred.

5. Acknowledgements

The authors would like to thank KNMI for providing weather data. We also would like to thank Claudio Piccinini for providing the *arcpy* modules transforming temperature datasets and Dr. Emma Izquierdo-Verdiguier for her advice in data mining algorithms and remote sensing during the development of this experiment.

6. References

Agrawal, R & Srikant R, 1994. Fast Algorithms for Mining Association Rules. *Proceedings of the 20th VLDB Conference, 42, pp.487–499.*

Barrios González, JM, 2013. Spatio-temporal modelling of the epidemiology of Nephropathia Epidemica and Lyme Borreliosis. KU Leuven, Belgium

Estrada-Peña, A et al., 2011. Correlation of B. burgdorferi sensu lato prevalence in questing I. ricinus ticks with specific abiotic traits in the western palearctic. *Applied and environmental microbiology, 77(11), pp.3838–45*.

Estrada-Peña, A, 2001. Distribution, Abundance, and Habitat Preferences of I. ricinus in Northern Spain. *J. Medical Entomology, 38(3), pp.361–370*.

Fink, D et al., 2010. Spatiotemporal exploratory models for broad-scale survey data. *J. Ecological Applications, 20(8), pp.2131–2147*.

Hofhuis, A et al., 2015. Continuing increase of tick bites and Lyme disease between 1994 and 2009. *Ticks and tick-borne diseases 6, pp.69–74*.

Kohonen, T, 1982. Self-Organized Formation of Topologically Correct Feature Maps. *Biol. Cybern., 43, pp.59–69.*

Medlock, JM et al., 2013. Driving forces for changes in geographical distribution of I. ricinus ticks in Europe. *Parasites & vectors, 6, p.1.*

Ogden, NH et al., 2006. Climate change and the potential for range expansion of the Lyme disease vector I. scapularis in Canada. *International J. Parasitology, 36(1), pp.63–70.*

Sprong, H. et al., 2012. Circumstantial evidence for an increase in the total number and activity of Borrelia-infected I. ricinus in the Netherlands. *Parasites & vectors, 5(1), p.294.*