Zürcher Hochschule
für Angewandte Wissenschaften

zh
aw

Zurich University of Applied Sciences
School of Life Sciences and
Facility Management
Institute of Applied Simulation

Modelling tick risk from uncertain citizen science data

Exploring the benefits of irregular spatial tessellations

Master's thesis

by
Ueli Mauch

Master's degree course 2019
Applied Computational Life Sciences
April 2021

Supervisors:
Prof. Dr. Laube, Patrick
Prof. Dr. Ott, Thomas

Abstract

Ticks pose a risk to human health, as they can transmit dangerous diseases. As part of a citizen science project, tick observations are reported since 2015 to improve spatial tick risk modelling. This study aims to evaluate the possibility of using an alternative zoning based on irregular spaced tessellation to model tick risk. This zoning is compared with an existing risk model based on raster data. Furthermore, it is investigated how the uncertainty of tick observations impacts the risk model.

In this context spatial data of the three Cantons of Aargau, Bern and Zurich have been aggregated and processed, so that a zoning is created that describes the ground cover classes. The tick observations coming from the app *Zecke* serve as a basis for a Monte Carlo simulation. The output of this simulation is taken as a proxy for tick risk. Using a neural network this risk together with an input representing the human activity (exposure) has been used to predict tick hazard for the three cantons. The output is a predicted risk and predicted hazard for the study areas. These values have been visualized as a map for the raster and vector models and compared afterwards. Moreover, $1'000$ points have been randomly sampled within every canton to calculate Spearman's rank correlation coefficient. To evaluate the uncertainty estimation Monte Carlo simulations have been conducted. Every observation has been randomly placed around its reported location $1'500$ times. This has been repeated for different uncertainty radii. Afterwards, the simulated points have been joined with the zoning to calculate the fraction of points per ground cover class.

Based on the workflows presented in this work, it is possible to model tick risk using a vector zoning. The results showed a weak correlation between the raster and vector model by comparing the hazard ($0.18 - 0.36$) and a slightly stronger one by comparing the risk ($0.44 - 0.76$). The discrepancies, which can be seen based on the scatter plots, show that polygons extending over large areas are a weakness of this zoning. Thereby, summarizing underlying values using zonal statistics can lead to problems with large polygons, as their values are strongly generalised. Despite that, using vector data provides the possibility to include even tiny objects. However, the generated predictions are just simplified models, in particular because there is no validation about the human exposure that was used and the reported ticks only account for a fraction of the tick population. The results for the uncertainty estimation indicate that Monte Carlo simulations are useful to investigate the uncertainty factor in risk modelling. It has been shown that uncertainty can have a big impact on the risk modelling. On the one hand, when choosing a high uncertainty radius, a large fraction of the simulated points can be found in agricultural areas. On the other hand, another large fraction is located inside the building class. This indicates that ticks are not reported into the app until the user is at home and has removed it from the body.

The results suggest that the approach to model risk can also be conducted using a vector zoning. If an upper limit of the area for single polygons is addressed, the approach can be seen as a serious alternative to a raster zoning. Furthermore, it has been noted that the uncertainty has a large impact on tick risk modelling. However, the idea behind citizen science projects is full of potential. If a focus is put on minimising the uncertainty in the data, there certainly will follow other projects that will provide knowledge back to the population.

Zusammenfassung

Zecken stellen eine Gefahr für die Menschen dar, weil sie gefährliche Krankheiten übertragen können. Im Rahmen eines Citizen Science Projektes werden seit 2015 Zeckenbeobachtungen erfasst, um daraus die Vorhersage des räumlichen Zeckenrisikos zu verbessern. In dieser Arbeit wurde untersucht, wie eine alternative Zonierung anhand einer unregelmässigen Raumeinteilung für die Risikomodellierung genutzt werden kann. Diese wurde mit dem bestehenden Modell verglichen, welches auf Rasterdaten basiert. Zudem wurde untersucht, wie sich die Unsicherheit der Zeckenmeldungen auf die Modellierung auswirkt.

Zur Durchführung dieser Arbeit wurden räumliche Daten der Kantone Aargau, Bern und Zürich zusammengetragen und eine Zonierung erstellt, welche die Bodennutzungs- und Bodenbedeckungsklassen beschreiben. Die Beobachtungen aus der App *Zecke* dienten als Grundlage für Monte Carlo Simulationen. Deren Ergebnisse wurde als Annäherung für das Zeckenrisiko angenommen. Mittels neuronalem Netzwerk wurde aus diesem Risiko (risk) und einem Modell der menschlichen Aktivität (exposure) die Zeckengefahr für die drei Kantone berechnet. Als Output ergibt sich für alle Flächen ein vorausgesagtes Risiko und eine vorausgesagte Gefahr (hazard). Dieser wurde für die Vektor- und Rasterzonierung als Karten visualisiert und verglichen. Zusätzlich wurden pro Kanton 1′000 zufällige Samplepunkte verteilt und daraus die Spearman-Korrelation berechnet. Für die Abschätzung der Unsicherheit wurden Monte Carlo Simulationen durchgeführt. Jede Beobachtung wurde um ihren erfassten Standort 1′500 Mal simuliert. Dies wurde für verschiedene Unsicherheitsradien wiederholt. Die simulierten Punkte wurden dann mit der erstellten Vektorzonierung verknüpft und für jede Klasse berechnet, wie gross ihr Anteil an Punkten ist.

Basierend auf den in der Arbeit vorgeschlagenen Arbeitsabläufen kann eine Vektorzonierung verwendet werden, um die Zeckengefahr zu modellieren. Dabei zeigt sich eine schwache Korrelation der vorausgesagten Gefahr (hazard) im Vergleich mit dem Rastermodell (0.18 – 0.36) und eine leicht höhere Korrelation (0.44 – 0.76) beim Vergleich des vorausgesagten Risikos (risk). Die Diskrepanzen, welche anhand der erstellten Streudiagramme der Samplepunkte ersichtlich sind, zeigen Schwächen der Zonierung, wenn sich einzelne Polygone über grosse Flächen erstrecken. Das Zusammenfassen von darunterliegenden Werten mittels zonaler Statistiken führt dazu, dass der Ansatz für eher grosse Polygone problematisch ist, weil diese stark generalisiert werden. Dem gegenüber bietet die Verwendung von Vektordaten die Möglichkeit, auch sehr kleine Objekte zu erfassen. Allerdings sind die erzeugten Voraussagen nur ein vereinfachtes Modell, insbesondere weil für die menschliche Aktivität keine Validierung vorliegt und die gemeldeten Zeckenpunkte nur einen Teil der Zeckenvorkommen abdecken. Die Resultate der Unsicherheitsabschätzung zeigen, dass Monte Carlo Simulationen genutzt werden können, um die Ungenauigkeit in Citizen Science Daten zu untersuchen. Es ist ersichtlich, dass die Ungenauigkeit einen grossen Einfluss auf die Risikomodellierung hat. Einerseits landen bei der Wahl eines grossen Unsicherheitsradius ein grosser Anteil der Punkte in landwirtschaftlichen Flächen, andererseits liegt ein hoher Anteil von Zeckenpunkten in der Bodennutzungsklasse Gebäude vor. Dies kann daran liegen, dass die Zecken erst zu Hause auf dem Körper entdeckt und in der App erfasst werden.

Es wurde gezeigt, dass der Ansatz der Risikomodellierung auch mit einer Vektorzonierung funktioniert. Wenn bei der Zusammenstellung der Zonierung beachtet wird, dass die Polygone nicht zu gross sein sollten, kann der Ansatz als ernsthafte Alternative zu Rasterzonierungen in Betracht gezogen werden. Weiter wurde dargelegt, dass die Unsicherheit einen grossen Einfluss auf die Modellierung des Zeckenrisikos hat.
Die Idee hinter Citizen Science Projekten hat grosses Potenzial. Wird dabei ein Fokus auf die Minimierung der Datenungenauigkeit gelegt, werden in Zukunft bestimmt weitere Projekte folgen, die der Bevölkerung wichtige Erkenntnisse liefern.

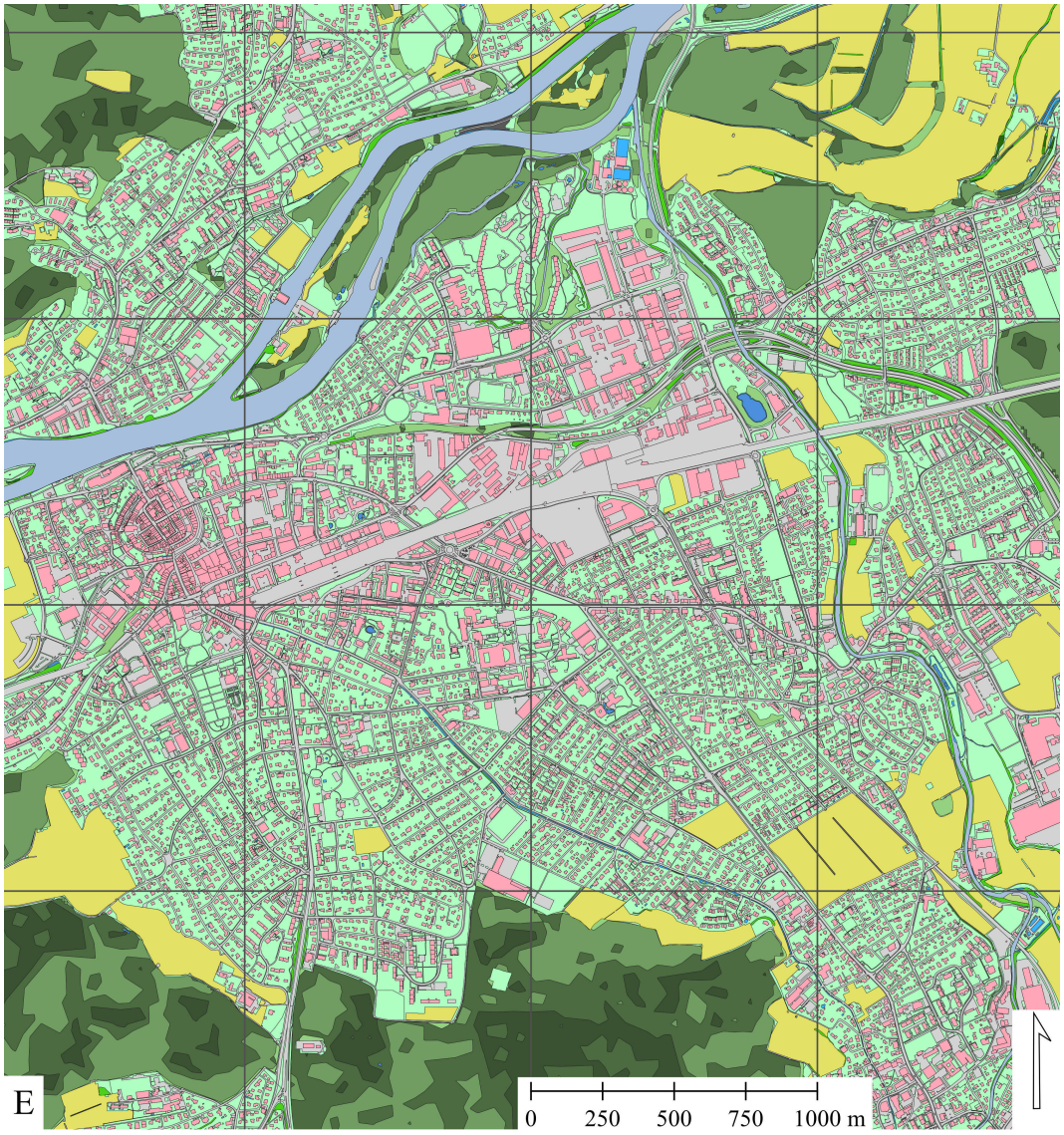

Figure 6 This map illustrates the final vector zoning for the same extent as in the minimaps of Figure 5. The number correspond to the class cover codes, with their labels in brackets.

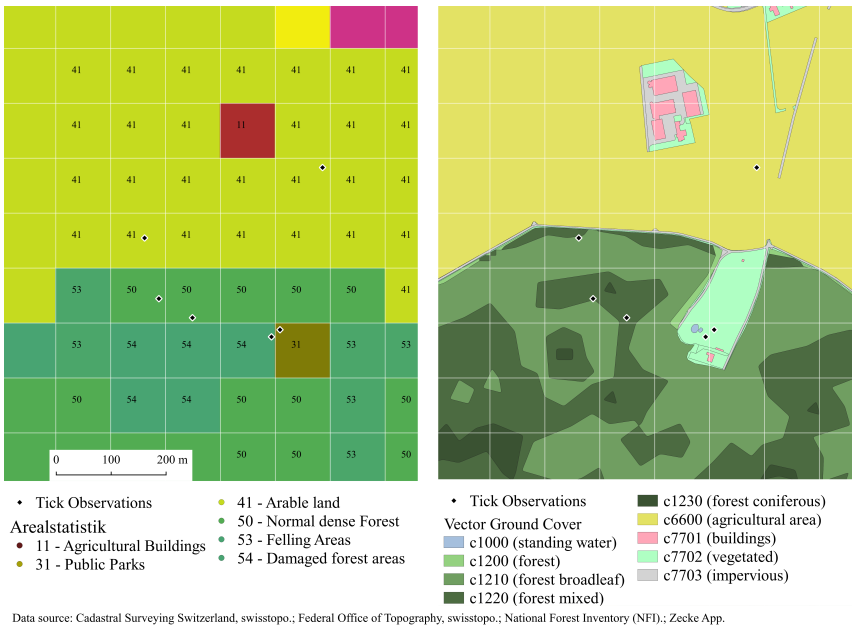

Figure 7 Comparison of the raster model and the vector model. The same area is shown in the maps once with the raster (left) and once with the vector model. Tick observations are indicated in black. The grey grid overlay serves the orientation and has the same extent as the Arealstatistik raster, which is used as a base map for the raster model.

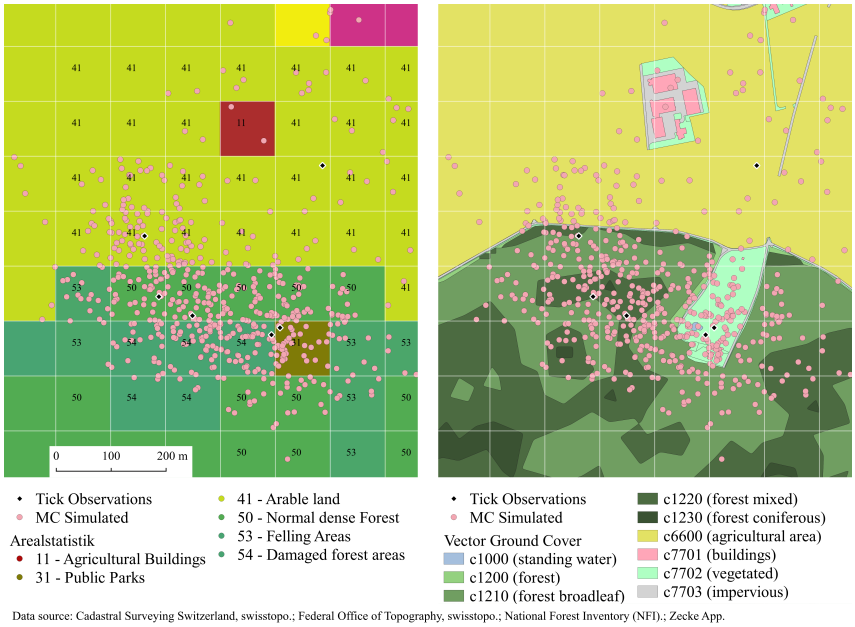

Figure 8 Same map as in Figure 7 but with simulated ticks (150 runs for illustration purposes). This serves as a demonstration that both, the granularity of map features as well as the semantic distinction, have an influence on where the simulated points are placed.

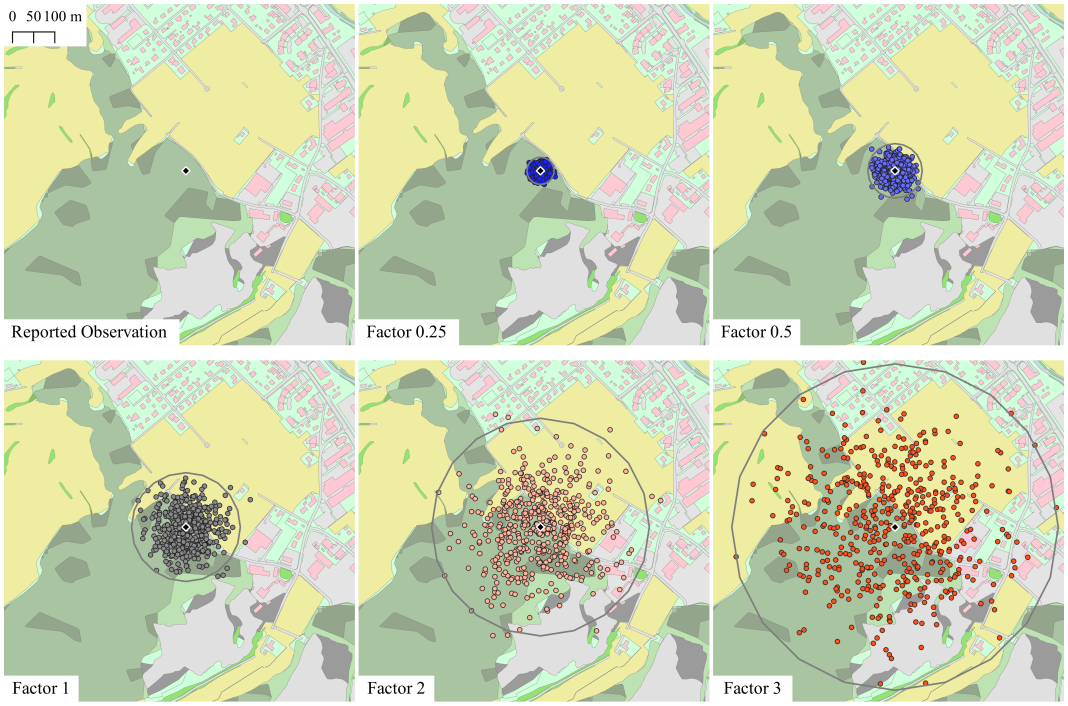

Figure 9 Demonstration of the Monte Carlo simulation using one observation with 500 iterations for each of the 5 factors, whereas factor 1 corresponds to the original accuracy reported by the app/user (for this particular point 257 m).

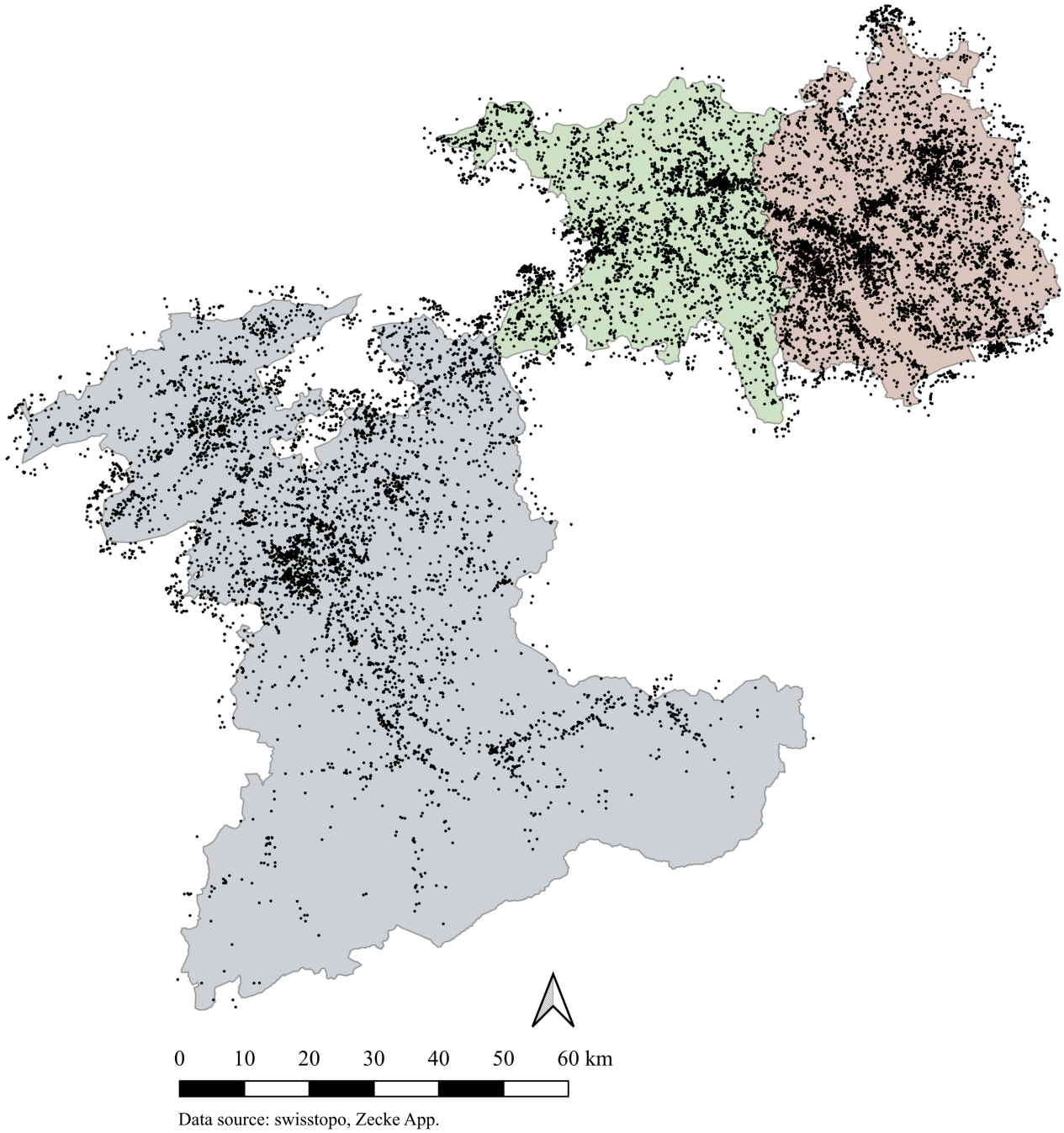

Figure 11 Depicted in the map are the cantons as well as the tick points that have been buffered. For this reason the points extend over the borders of the cantons. Cantons from left to right: Bern, Aargau, Zurich.

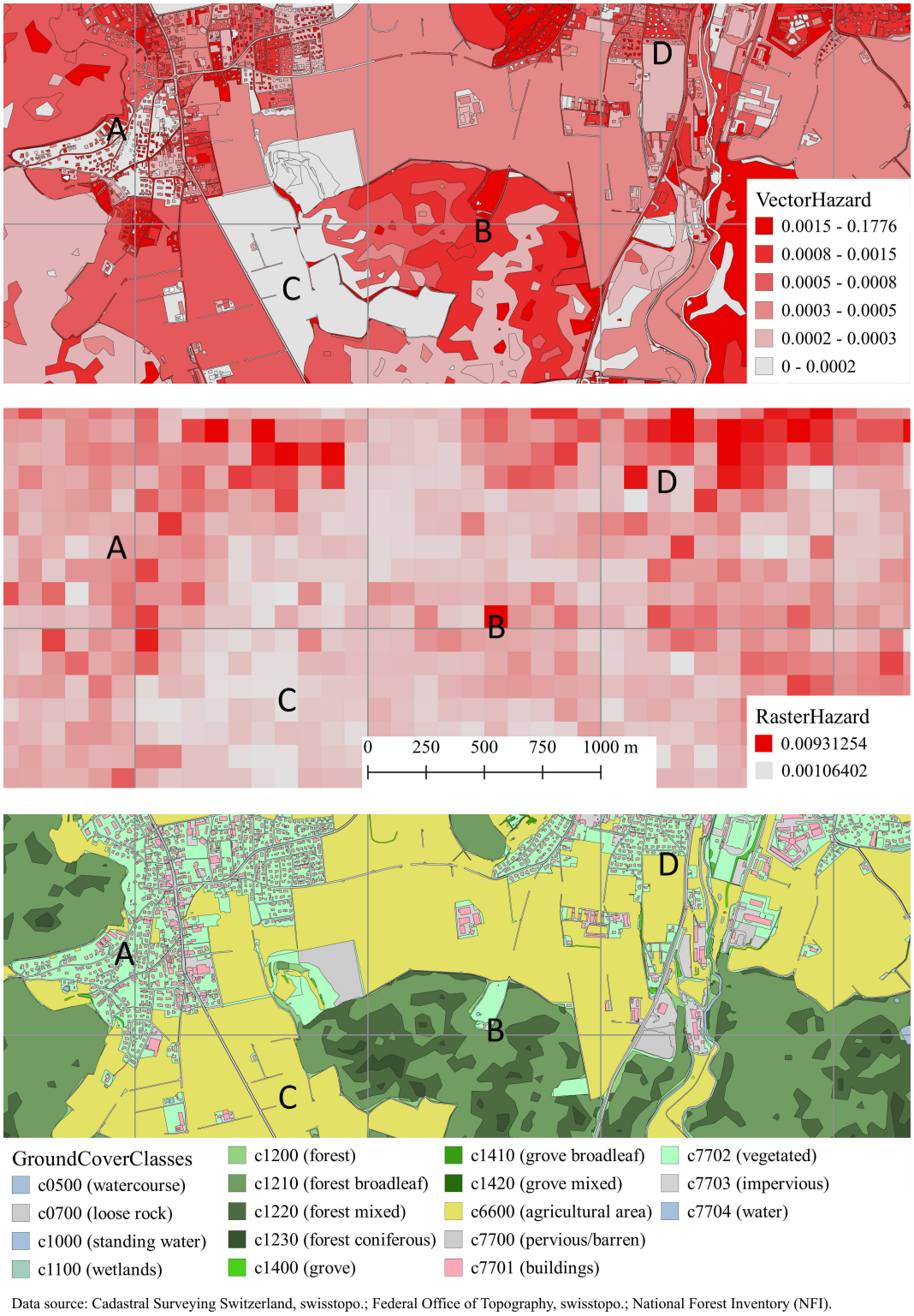

Figure 23 This figure illustrates differences in the modelled hazard between the vector and raster models for the region of Staufen, Canton of Aargau. The top map shows the vector zoning, followed by the raster zoning and the assembled ground cover map on the bottom.

8 Conclusion

This thesis shows a proof of concept how tick risk can be modelled based on vector data. The methods described can be useful for the project team, if the tick risk has to be modelled for the country of Switzerland using vector data. Thereby, it would be interesting to identify possible differences between the remaining cantons, especially since the cantons already investigated are among those with high contributions. The question arises, whether a model performs differently when applied to the complete land coverage.

What remains is the problem of exposure, that was plainly assumed without validation in this thesis. Despite that there are a few options to assess and validate the human activities for this purpose. First, there is the possibility to record the human motion within the *Zecke* app itself. Of course, this works only with consent of the users and implies adequate information campaigns beforehand. In addition, there has to be a simple option to opt out. Secondly, there is the option to buy telecommunication data and use it for modelling purposes. Something similar has been done to assess the mobility during the COVID-19 pandemic (Persson, Parie & Feuerriegel, 2021).
Another aspect is the fact that only true positives can be modelled by using the app data. That means that there is no feedback if no tick has been observed. This problem could be addressed if the app has permission to gather location data in the background. A simple push message could be sent, if a user is located within a certain distance to forests/parks, to send a reminder about the tick risk. This could also provide means to assess the mobility.

Furthermore, the uncertainty should be mentioned. As the results show, the Monte Carlo simulations provide a way to assess the uncertainty and make quantitative statements about tick distribution over land cover classes. This is a central topic that should be addressed even further. Instead of assessing the uncertainty, the reduction thereof should be put into focus. Two suggestions can be made. On the one hand, data science methods and tools can be used to filter the collected data but exclude the ones with high uncertainty. The hope is to achieve that the remaining fraction is positioned accurately. On the other hand, an option that might be more promising, is to inform and include the user as frequently as necessary. For example, the user has to be informed if a tick is reported within a class that is not a typical tick habitat. Follow-up questions can be used to gather additional data about the locations that are more likely the origin for that tick. Of great importance is, to bring to mind that not the location where the tick has been removed from the body is relevant. Rising awareness could help to integrate valuable points into to risk modelling and thus reduce the noise coming from false reported locations.

The solutions proposed to reduce the uncertainty in citizen science data should be feasible without intruding the users' privacy. Obviously, these projects are dependent on strong integration with the people who collect and contribute data. Therefore, the best way is probably to inform user what data is collected and get consent, that these contributions are essential to improve predictions. In the end, users can benefit from these better predictions. Finally, further research is done by *Fighting bites with bytes* in the direction of prediction tick risk, by considering the temporal dimension. Moreover, the integration of near real-time meteorological data will be used to model accurate daily predictions. It will become clear whether the final model will be implemented using a raster or a vector approach.