Centre for Geo-Information

Thesis Report GIRS-2015-07

Spatial patterns of mosquitoes in the Netherlands: Finding mosquito hotspots based on crowdsourced data

Linda Klein

1-4-2015

WAGENINGEN UR
For quality of life

Spatial patterns of mosquitoes in the Netherlands: Finding mosquito hotspots based on crowdsourced data

Linda Klein[1]
910830-440130

Supervisors: Ron van Lammeren[1], Sander Koenraadt[2], Chantal Vogels[2]

[1]Laboratory of Geo-Information Science and Remote Sensing
[2]Laboratory of Entomology

A thesis submitted in partial fulfilment of the degree of Master of Science
at Wageningen University and Research Centre,
The Netherlands

1-4-2015
Wageningen, The Netherlands

Thesis code number: GRS-80436
Thesis Report: GIRS-2015-07
Wageningen University and Research Centre
Laboratory of Geo-Information Science and Remote Sensing

Acknowledgements

I would like to thank my supervisors, Ron van Lammeren (Laboratory of Geo-Information Science and Remote Sensing), Sander Koenraadt, and Chantal Vogels (Laboratory of Entomology) for assisting and supporting me with my thesis, and always being available for questions. The regularly scheduled meetings in which we had interesting discussions about both GIS and mosquitoes were of great help for me in finishing this thesis, and helped me to stay enthusiastic about my topic.

I also thank my fellow MGI students in the thesis room for keeping the fun-factor in working on my thesis. Without the great atmosphere we had in the thesis room, finishing my thesis would definitely have been much harder.

Abstract

Spatial patterns of mosquitoes in the Netherlands are largely unknown. Knowledge on these patterns is desired, since the risk of transmission of viruses transmitted by these mosquitoes, such as West Nile Virus, in Europe is increasing. The data that became available from the *Muggenradar* project, a project in which mosquitoes were collected in a crowdsourced way during January and February, and August and September 2014, can be used to analyse the spatial patterns of mosquitoes in the Netherlands. The main objective of this study was to find mosquito hotspots in the Netherlands in both winter and summer, based on crowdsourced data. It was found that more mosquito reports were located in urban areas. Four environmental factors were therefore tested for their relationship with mosquito presence in two urban study areas: Amsterdam and Rotterdam. The tested factors were proximity to water, proximity to deciduous forest and trees, building construction year, and population density. Although significance was found between the mosquito reports and proximity to water, proximity to deciduous forest, and population density in the 2014 January and February data, these patterns were not consistent. Hotspot maps created out of the found relationships did not result in accurate maps. The inconsistence in the results could be explained by the low number of *Muggenradar* mosquito reports, especially in the 2014 August and September data. Another possible explanation is the fact that more environmental factors that were not tested are predictive for mosquito presence. All reported mosquitoes were collected indoors. Local factors, like small water bodies on private properties, could therefore possibly be more predictive than the tested environmental factors.

Keywords: Culicidae, Spatial ecology, geographic information systems, GIS, crowd sourcing, citizen science

Table of Contents

1 Introduction 1
1.1 Context and Background 1
1.2 Problem Definition 2
1.3 Research Objective & Research Questions 3
1.4 Reading outline 3
2 Related work 4
2.1 Crowdsourcing in Ecology 4
2.2 Mosquito Habitats 4
2.3 Spatial Pattern Analysis 5
2.4 Conclusions 5
3 Methodology 7
3.1 Approach 7
3.2 Data 7
3.3 Pre-processing 9
3.4 Spatial Distribution Analysis 11
3.5 Demographic Characteristics 12
3.6 Environmental Predictors for Mosquito Presence 13
3.7 Mosquito Presence Hotspots 19
3.8 Software 20
4 Results 21
4.1 Spatial Distribution Analysis 21
4.2 Demographic Characteristics 25
4.3 Environmental Predictors for Mosquito Presence 26
4.4 Mosquito Presence Hotspots 34
5 Discussion & Conclusions 39
5.1 Discussion 39
5.2 Main conclusions 43

List of Figures

1 The approach of this study, in which the part in the red circle shows the exploration of the crowdsourced data, and the blue circle shows the spatial pattern analysis. 7
2 Flowchart of the pre-processing of the Muggenradar data 10
3 Mosquito reports originating from Amsterdam. The January and February reports are shown in blue, and August and September reports in red. 13
4 Mosquito reports originating from the Rotterdam region. The January and February reports are shown in blue, and August and September reports in red. 14
5 Proximity to water for Rotterdam. This is not the final input raster for the analysis, since that one consists of the mean proximity to water per PC5 area. 15
6 Proximity to deciduous forest and trees for Rotterdam. This is not the final input raster for the analysis, since that one consists of the mean distance to deciduous forest per PC5 area. 16
7 Building construction year raster for Rotterdam, showing the median construction years for each PC5 area. 17
8 The population density (inhabitants per km²) for Rotterdam. 17
9 Ripley's K estimate for January and February (left), and September and August (right) point patterns. Kobs(r) is the observed value of K(r) for the data pattern, Ktheo(r) is the theoretical value of K(r) for a simulated CSR, Khi(r) is the upper pointwise envelope of K(r) from simulations, and Klo(r) the lower pointwise envelope of K(r) from simulations. 21
10 Kernel density maps of the January and February 2014 reports (left) and August and September 2014 reports (right) 22
11 Correlation between expected mosquito reports per km² and addresses per km², Pearson's correlation coefficient (r) and its p-value. The red background colour represents urbanisation classes as used by CBS, in which the lightest colour represents 'not urbanised' (0-500 addresses per km2), and the darkest 'strongly urbanised' (2500 or more addresses per km2). 23
12 Co-occurrence matrices of all reported genera in January/February (left) and August/September (right) 24
13 Scatter plot of the correlation between presence reports per km² and absence reports per km² for both January and February (top) and August and September (bottom), Pearson's correlation coefficient (r) and its p-value 25
14 Line graphs showing the Chi Square test's p-value for all tested class breaks for proximity to water in Amsterdam in Jan/Feb (top), and Aug/Sep (bottom). The white line represents the p-value. At break values where the white line exceeds the dotted line at 0.05, the p-value is significant. The blue line represents the number of reports that are made at a distance from 0 metres to the distance of the break value. The red line represents the count of the mean of 20 simulations for each break value. 27
15 Line graphs showing the Chi Square test's p-value for all tested class breaks for proximity to water in Rotterdam in Jan/Feb (top), and Aug/Sep (bottom). The white line represents the p-value. At break values where the white line exceeds the dotted line at 0.05, the p-value is significant. The blue line represents the number of reports that are made at a distance from 0 metres to the distance of the break value. The red line represents the count of the mean of 20 simulations for each break value. 28

16 Line graphs showing the Chi Square test's p-value for all tested class breaks for proximity to deciduous forest/trees for Amsterdam in Jan/Feb (top), and Aug/Sep (bottom). The white line represents the p-value. At break values where the white line exceeds the dotted line at 0.05, the p-value is significant. The blue line represents the number of reports that are made at a distance from 0 metres to the distance of the break value. The red line represents the count of the mean of 20 simulations for each break value. 29
17 Line graphs showing the Chi Square test's p-value for all tested class breaks for proximity to deciduous forest/trees for Amsterdam in Jan/Feb (top), and Aug/Sep (bottom). The white line represents the p-value. At break values where the white line exceeds the dotted line at 0.05, the p-value is significant. The blue line represents the number of reports that are made at a distance from 0 metres to the distance of the break value. The red line represents the count of the mean of 20 simulations for each break value. 30
18 Line graphs showing the Chi Square test's p-value for all tested class breaks for inhabitants per km^2 for Amsterdam in Jan/Feb (top), and Aug/Sep (bottom). The white line represents the p-value. At break values where the white line exceeds the dotted line at 0.05, the p-value is significant. The blue line represents the number of reports that are made from areas with 0 inhabitants per km^2 to the break value of number of inhabitants per km^2. The red line represents the count of the mean of 20 simulations for each break value. 32
19 Line graphs showing the Chi Square test's p-value for all tested class breaks for inhabitants per km^2 for Rotterdam in Jan/Feb (top), and Aug/Sep (bottom). The white line represents the p-value. At break values where the white line exceeds the dotted line at 0.05, the p-value is significant. The blue line represents the number of reports that are made from areas with 0 inhabitants per km^2 to the break value of number of inhabitants per km^2. The red line represents the count of the mean of 20 simulations for each break value. 33
20 Hotspot map of Amsterdam, showing the locations where mosquito presence during the winter months is most likely. Muggenradar mosquito reports that were made in January and February 2014 are shown as black point symbols. 34
21 Hotspot map of Amsterdam showing the locations where mosquito presence during the summer months is most likely. Muggenradar mosquito reports that were made in August and September 2015 are shown as black point symbols. 35
22 Hotspot map of Rotterdam, showing the locations where mosquito presence during the winter months is most likely. Muggenradar mosquito reports that were made in January and February 2014 are shown as black point symbols. 35
23 Hotspot map of Rotterdam, showing the locations where mosquito presence during the summer months is most likely. Muggenradar mosquito reports that were made in August and September 2014 are shown as black point symbols. 36
24 Validation hotspot maps of Utrecht, based on the January and February parameters of Amsterdam (a) and Rotterdam (b), showing the locations where mosquito presence during the winter months is most likely based on the parameters of these cities. Muggenradar mosquito reports that were made in January and February 2014 are shown as black point symbols. 37

25 Validation hotspot maps of Utrecht, based on the August and September parameters of Amsterdam (a) and Rotterdam (b), showing the locations where mosquito presence during the summer months is most likely. Muggenradar mosquito reports that were made in August and September 2014 are shown as black point symbols. 38

List of Tables

1 Summary of the Muggenradar data (based on raw data without pre-processing) 8
2 Summary of the identified mosquitoes up to genus level of the Muggenradar data (based on raw data without pre-processing) 8
3 Differences, advantages, and disadvantages of CO_2 traps and crowdsourced mosquito collection 9
4 Secondary data used 9
5 Number of mosquito reports per study area 14
6 Environmental factors, their source and the used attribute fields 15
7 Probability table of the co-occurrence analysis between the different genera (January/February 2014) 24
8 Probability table of the co-occurrence analysis between the different genera (August/September 2014) 24
9 Correlation coefficients (r) and their significance (p) for each of the tested demographic variables with the mosquito reports. Significant positive correlations are shown in green, while significant negative correlations are shown in red 26
10 Overview table showing the strongest odds ratio and corresponding break value for each study area and Muggenradar round for the proximity of water 28
11 Overview table showing the strongest odds ratio and corresponding break value for each study area and Muggenradar round for the proximity to deciduous forest and trees 30
12 The results of the Chi Square test for building construction year. The first two columns show the number of reports coming from a building with a certain construction year, while the following two columns show the number of reports that is expected to be reported from these buildings, based on 20 simulations of random PC5 centroids. The last three columns show the Chi Square result, the corresponding significance (p-value) and the odds ratio. 31
13 Overview table showing the strongest odds ratio and corresponding break value for each study area and Muggenradar round for population density 33
14 The number of reports made in expected low presence areas (where OR < 1), expected high presence areas (OR > 1), and the relative count for Amsterdam and Rotterdam. 36
15 The number of reports made in expected low presence areas (where OR < 1), expected high presence areas (OR > 1), and the relative count for the validation area Utrecht, using the parameters of both Amsterdam and Rotterdam. 38
16 Examples of citizen science projects and their characteristics 43

1 Introduction

1.1 Context and Background

Mosquitoes are found to be a large nuisance in everyday life in the Netherlands, because of their buzzing and biting. The most abundant mosquitoes in the Netherlands are *Culex* and *Culiseta* mosquitoes. *Anopheles* and *Aedes* genera are less abundant, but can also be found.

Although mosquitoes in the Netherlands are generally known for their nuisance in our everyday lives, they are also a vector for the transmittance of many pathogens. The *Anopheles plumbeus* species, also present in the Netherlands, is a minor vector for human malaria. It is able to transmit *Plasmodium falciparum*, a parasite responsible for the tropical and most deadly form of malaria (Schaffner et al., 2012). Some cases of malaria have been reported in Europe, and two cases caused by *Anopheles plumbeus* occurred in a German hospital (Krüger et al., 2001). Although these cases call for continued malaria surveillance, the chance of malaria incidences such as this one are extremely rare (Krüger et al., 2001).

Aedes mosquitoes are sometimes imported via tyres and Lucky bamboo from East-Asian countries (Takumi et al., 2009; Scholte et al., 2010). This genus, that also includes the well-known Asian tiger mosquito (*Aedes albopictus*), is known as a vector for the tropical dengue- and chikungunya viruses. The public health risks related to *Aedes* mosquitoes are high in southern Europe. In the Netherlands, however, the genus will probably not survive the cold temperatures during winter (Scholte et al., 2010).

More concerning is the risk of West Nile Virus (WNV). The most important vectors for spreading this virus in Europe are *Culex* mosquitoes, which are highly abundant in the Netherlands (Reusken et al., 2010). The main hosts of WNV are birds: both resident, non-migratory birds and migratory birds may contribute to the dispersal of WNV (Hayes et al., 2005). Mosquitoes that take their blood meal from WNV infected birds can spread the virus to humans and other mammals.

Several cases of West Nile fever have been reported over the last few years in Europe. A study of Tran et al. (2014) shows a large increase in the number of reported cases in eastern and central Europe since 2010. Most cases were reported in countries in eastern Europe, although some cases have been reported in southern European countries, such as Spain, Italy, and Greece since 2008 (Tran et al. 2014). The only reported cases of West Nile fever in the Netherlands were all imported from other countries, such as Canada (Prick et al., 2003). No direct infections of West Nile fever are reported in the Netherlands yet (Reusken et al., 2011). The fact that WNV is able to overwinter in *Culex* mosquitoes (Nasci et al., 2001; Reisen et al., 2006), together with the increasing number of WNV cases in Europe shows that it could become a risk in the future.

This potential WNV risk emphasizes the need for mosquito presence monitoring. In order to increase the effectivity and efficiency of interventions in case of potential WNV threats in the Netherlands, insight in the spatial patterns of mosquitoes is desired. Studies in other countries found that environmental factors can be predictive for the presence of mosquitoes (DeGroote et al., 2007; Deichmeister & Telang, 2011; Diuk-Wasser et al., 2006; Gleiser & Zalazar, 2010; Hongoh et al., 2012; Tran et al., 2014; Valiakos et al., 2014). All of these studies used an approach based on geographic information systems (GIS). GIS-based approaches have become widely used by professionals in research, also in the field of mosquito mapping, to provide a better understanding of spatial phenomena and their relationships (Becker et al., 2010). Almost all of the studies named have used CO_2 traps to collect mosquitoes. This method gives insight in mosquito presence and abundance, but does not give information about nuisance to humans. Furthermore, all studies are conducted during the summer, and spatial patterns of mosquitoes in winter are still unknown.

Environmental factors found to be predictive for mosquito presence in a specific country are not always applicable to the Netherlands, however, because of different climatic conditions and different types of landscapes. The spatial patterns of mosquitoes in the Netherlands are therefore still largely unknown.

A recent project by Wageningen University and Research Centre (Wageningen UR) can provide the input for a study on the spatial patterns of mosquitoes in the Netherlands. During January and February 2014, the Laboratory of Entomology, in collaboration with the Environmental Systems Analysis group started a project called *Muggenradar*. The project was repeated during August and September of 2014. Mosquitoes were collected through a crowdsourced approach. Instead of the more traditional use of CO_2 traps, Dutch citizens were asked to fill in an online questionnaire regarding mosquito presence and nuisance. Respondents were also requested to collect dead mosquitoes and to send these in an envelope to Wageningen UR.

The use of crowdsourcing, also called citizen science, is increasingly used in ecology to collect species on a broad geographic scale (Crall et al., 2010; Dickinson et al., 2010). The use of crowdsourcing as a tool for mosquito collection, however, is new. It is therefore not clear if the quality and quantity of this data is good enough to study the spatial patterns of mosquitoes. Using the *Muggenradar* dataset for this purpose will thus give insight in how useful citizen science data is for GIS-based analyses. Based on this, recommendations can be done for future ecological citizen science projects.

Using the *Muggenradar* dataset to explore spatial patterns of mosquitoes in the Netherlands could give insight in which areas are exposed to higher mosquito presence than other areas. Having knowledge on these areas could help the prevention of potential WNV incidences in the future, since interventions can take place at the right location. Therefore, the main objective of this study is to find hotspots of mosquito presence in the Netherlands in both summer and winter, based on crowdsourced data.

1.2 Problem Definition

An early warning regarding the potential incidence of WNV in the Netherlands needs an insight in the spatial pattern of mosquitoes. Such a study may help to increase the effectivity and efficiency of interventions in case of potential WNV threats. Spatial patterns of mosquitoes and areas with high mosquito abundance have been studied before in other countries, such as the US (DeGroote et al., 2007; Deichmeister & Telang, 2011; Diuk-Wasser et al., 2006). In the Netherlands, a GIS-based study on the spatial patterns by sampled mosquitoes has never been done before. One of the reasons may be the laborious and expensive data collection methods, like the use of CO_2 traps to collect mosquitoes. Two studies conducted in Europe (Tran et al., 2014; Valiakos et al., 2014) did use WNV reports instead of mosquito collection to define WNV risk areas. To collect mosquitoes by crowdsourcing and using these crowdsourced data to study spatial patterns of mosquitoes have never been done before, although crowdsourcing is used increasingly in ecology studies (Silvertown, 2009). The *Muggenradar* project could for the first time provide crowdsourced input for a study on the spatial patterns of mosquitoes in the Netherlands.

However studying spatial patterns of mosquitoes in the Netherlands based on locations of crowdsourced mosquito reports puts question marks regarding the size and quality of sampled data in relation to the study of spatial patterns. The outcomes of this study can therefore also be used to do recommendations for future citizen science projects in the field of ecology.

1.3 Research Objective & Research Questions

The research objective of this study is to find hotspots of mosquito presence in the Netherlands in both summer and winter, based on crowdsourced data.

Four research questions are formulated to reach the objective of this study.

1. How are the observations of mosquitoes in the *Muggenradar* dataset distributed over the Netherlands?
2. Which demographic factors are characteristic for the areas from which mosquito reports are sent?
3. What are environmental predictors for the presence of mosquitoes in the Netherlands in summer and winter?
4. Which areas in the Netherlands are potential hotspots for the presence of mosquitoes in summer and winter?

1.4 Reading outline

First, Chapter 2 gives an overview of the related work regarding crowdsourcing in ecology, mosquito habitats, and spatial pattern studies. This small literature review is thereafter used to formulate hypotheses regarding the four research questions. These hypotheses are presented in Chapter 3, which explains the methodology of this study, and describes the data that is used. Chapter 4 presents the results of the study. Eventually, Chapter 5 discusses these results, and relates them to those of other studies. This chapter also acknowledges the limitations of this study, and gives suggestions for further research, on the subject of both spatial pattern studies and crowdsourcing.

2 Related work

2.1 Crowdsourcing in Ecology

Crowdsourcing, also often called citizen science because of the involvement of citizens, is increasingly used in ecology (Silvertown, 2009). The unique benefit of citizen science in ecology is its ability to collect data on a broad geographic scale at small costs. By enabling ecologists to move from a local scale to the scale of ecosystems, citizen science accounts for growth in the field of geographical ecology (Dickinson et al., 2010).

Gardiner et al. (2012) distinguishes between two types of citizen science: direct citizen science, and verified citizen science. Direct citizen science uses the crowdsourced data without verification, resulting in a cost effective data collection method. Verified citizen science uses trained experts to verify every report made. Direct citizen science often results in more data, but tends to overestimate species richness and diversity. Verified citizen science, however, has a higher accuracy but is more laborious. The ability to gather a large number of samples by using verified citizen science may compensate for the reduced accuracy compared to traditional science methods (Gardiner et al., 2012; Miller-Rushing et al., 2012).

Brotons et al. (2004) emphasizes that having both presence and absence data can improve the overall accuracy of the results of a study. Next to presence, crowdsourcing projects should thus also have the option to report absence. However, Sequeira et al. (2014) mentions that respondents are often hesitant in reporting absence, resulting in a strong bias towards presence-only data.

In order to assure the quality of crowdsourced data, Silvertown (2009) defined a number of challenges, of which the first one is that the data must be validated by experts (verified citizen science). Second, the data collection method must be well designed and standardised. Third, in order to do a targeted research, hypotheses must be formulated before starting the data collection. Fourth, the citizen scientists, or volunteers, must receive feedback on their responses as a reward for their participation.

2.2 Mosquito Habitats

Becker et al. (2010) names urban areas as a suitable habitat for many mosquito species, since there is an abundant availability of blood meals and a wide range of water bodies. Mosquitoes need water for breeding purposes: they lay their eggs in these water bodies. Examples of mosquito breeding sites in urban areas are construction sites, water storage containers such as rain barrels, and drainage systems and sewage and waste-water processing (e.g. gutters and ditches). Cemeteries and urban sanitation are also suitable sites for mosquito breeding, since they contain many small water reservoirs such as flower cases, drink cans, plant pots and tyres (Becker et al., 2010).

Anopheles larvae are usually found in natural water bodies, such as semi-permanent ponds, pools, puddles or ditches (Becker et al., 2010). They can also be found in wetlands or floodplains. A few species show exceptions: *Anopheles plumbeus* breeds typically in tree holes of deciduous trees. *Aedes* mosquitoes are found in various habitats, but most often they occur at the edges of semi-permanent pools, floodplains, wetlands or bogs. The *Aedes albopictus* (Asian tiger mosquito) species forms an exception, since it breeds in places as tree holes, vehicle tyres or broken glass bottles. This species, although frequently imported with second-hand car tyres and bamboo, is still considered 'not established' in the Netherlands (Takumi et al., 2009). *Culex pipiens pipiens* is able to inhabit almost every kind of water source, but they frequently occur in man-made water bodies. *Culex pipiens* biotype *molestus* is known to occur more frequently in human environments, including dark and moist cellars of large buildings in urban areas. The *Culiseta* genus breeds

mostly in open, unshaded natural water bodies. Contrary to *Anopheles, Aedes* and *Culex*, this genus is more resistant to colder temperatures (Becker et al., 2010).

2.3 Spatial Pattern Analysis

Several GIS-based studies found relationships between mosquito genera and environmental factors. Most of these studies are conducted in the United States. DeGroote et al. (2007) studied spatiotemporal patterns of adult mosquitoes in Iowa from May to August 2003. They mainly focused on the *Aedes* species, and found positive correlations with the landscape parameters wetlands, soil hydrological properties, deciduous forest and climatic factors such as temperature and precipitation. Deichmeister & Telang (2011) focused on the abundance of WNV vector species *Culex pipiens*, *Culex restuans* and *Aedes albopictus* in Henrico County, Virginia. They found that temperature along with low precipitation are strong predictors for vector abundance. A similar study is conducted by Diuk-Wasser et al. (2006) in Connecticut. They found that non-forested areas are predictive for the abundance of *Culex pipiens*. Distance to wetlands was found to be predictive for *Culiseta melanura*.

Tran et al. (2014) and Valiakos et al. (2014) did GIS-based studies on WNV abundance in Europe. Both studies used reported cases of WNV instead of collected mosquitoes to identify risk areas. The research of Tran et al. (2014) focussed on environmental predictors of West Nile fever in Europe. Environmental factors such as Normalized Difference Vegetation Index (NDVI), Modified Normalized Difference Water Index (MNDWI), population, birds' migratory routes and presence of wetlands were used. They found that anomalies of temperature in July, MNDWI in early June, the presence of wetlands, a location under birds' migratory routes and WNV outbreaks in previous years are predictive for WNV in Europe. The research's focus was mainly on the *Culex* genus, since this species is a principal WNV vector. Abundance of birds can be seen as a predictive factor, but only for ornithophilic species; species that take their blood meal from birds instead of mammals, such as *Culex pipiens pipiens* and *Culiseta morsitans* (Becker et al., 2010). *Anopheles* and *Aedes* mosquitoes are mammophilic, meaning that they take their blood meal from mammals (Becker et al., 2010). They will therefore not be attracted by the abundance of birds.

Valiakos et al. (2014) used wild bird surveillance and human WNV case data together with spatial analysis to predict the spatial distribution of WNV in Greece. Factors as low altitude and proximity to water were found to be important predictors of WNV in wild birds and humans.

2.4 Conclusions

Although the accuracy of the data is generally not as good as the accuracy retrieved with traditional science methods, it can be seen as good enough, especially when comparing it to direct citizen science methods. The low amount of costs related to crowdsourced studies could make up for the loss of accuracy. Crowdsourcing is not per definition better than traditional collection methods. It has the ability to monitor ecological phenomena on a broad geographic scale, whereas traditional collection methods can gather more accurate data on a smaller geographic scale. Crowdsourcing is therefore a good method to analyse the spatial patterns of mosquitoes on country level.

It is found that most mosquito genera appear close to water bodies (floodplains, wetlands, and bogs) and close to deciduous or broad leaved forests. Temperature and precipitation also accounts for the presence of mosquitoes, but the *Muggenradar* project only provides data from two moments in time, making the analysis of mosquito patterns in relation to these time-related factors not possible.

Hypotheses for this study are formulated based on this related work. These hypotheses can be found in the methodology section (Chapter 3).

3 Methodology

3.1 Approach

The methodology of this study can roughly be divided into two parts. The first part (red circle in Figure 1) focused on the exploration of the crowdsourced data. This part includes research question 1 and 2, which explains the spatial distribution of mosquito reports, and determines demographic factors that are of importance for citizens that report mosquitoes.

The second part (blue circle in Figure 1), including research question 3 and 4, focused on the spatial pattern analysis, by finding environmental predictors for mosquito presence. These environmental predictors were thereafter used to make a first determination of mosquito presence hotspots in the winter and summer seasons.

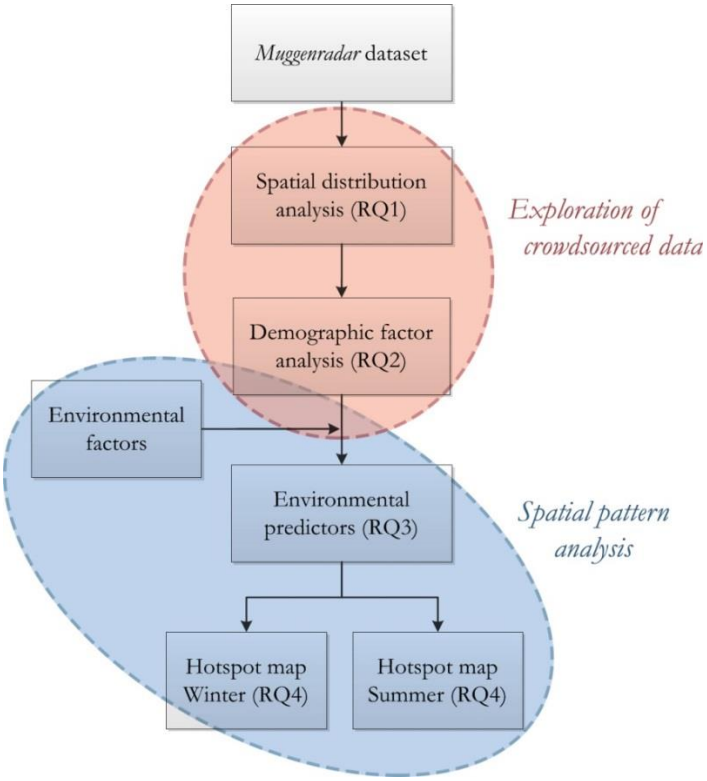


Figure 1: The approach of this study, in which the part in the red circle shows the exploration of the crowdsourced data, and the blue circle shows the spatial pattern analysis.

3.2 Data

This study made use of two types of data, with the *Muggenradar* project data as primary, and environmental demographic data and geo-data from external sources as secondary data.

3.2.1 Muggenradar dataset

The *Muggenradar* project started during a period of five weeks in January and February of 2014. Citizens of the Netherlands were asked to report mosquito presence and possible nuisance by filling in a questionnaire on the website (www.muggenradar.nl). The questionnaire contained questions whether or not mosquitoes were present inside the house, the experienced nuisance, and the type of nuisance. Citizens were also asked to fill in their full postal code (4 letters and 2 numbers, PC6 level). This information can be used to link every observation to a geographic

location. At the end of the questionnaire, respondents were requested to send a dead mosquito to the Laboratory of Entomology. A unique code made it possible to link the questionnaire to the mosquito that was sent. When no mosquitoes were seen, this could also be reported on the website. This can be seen as the absence data. At the Laboratory of Entomology, every mosquito was identified based on morphology to the genus level. The number of mosquitoes, and whether they were blood-fed or not was also recorded. Insects that do not belong to the *Culicidae* family were then filtered out.

Table 1: Summary of the Muggenradar data (based on raw data without pre-processing)

	January/February 2014		**August/September 2014**	
	Absolute	Relative (%)	Absolute	Relative (%)
Total response	3625	100	855	100
Reported presence	3305	91	590	69
Reported absence	319	9	143	17
Reports per post	2724	75	577	67
Identified as Culicidae	1563	43	472	55

Table 2: Summary of the identified mosquitoes up to genus level of the Muggenradar data (based on raw data without pre-processing)

	January/February 2014		**August/September 2014**	
	Absolute	Relative (%)	Absolute	Relative (%)
Culex	919	59	391	79
Culiseta	527	34	30	6
Anopheles	104	7	51	10
Aedes	0	0	24	5

During the summer of 2014, in August and September, the *Muggenradar* project was repeated during a period of two weeks, using the same format.

Table 1 shows a short summary of the *Muggenradar* data, and Table 2 shows a summary of the identified mosquitoes.

As became clear from the related work (Chapter 2), crowdsourced data differs from data collected using traditional science methods. Traditionally, CO_2 traps are used to collect mosquitoes. The *Muggenradar* project used a crowdsourced approach, resulting in different data. These differences, but also advantages and disadvantages of crowdsourced mosquito collection are shown in Table 3.

Table 3: Differences, advantages, and disadvantages of CO_2 traps and crowdsourced mosquito collection

CO_2 traps	**Crowdsourced mosquito collection**
Outdoor mosquito collection	(Mostly) Indoor mosquito collection
Relatively less presence locations	More (random) presence locations
Includes more objective absence locations	Includes only a small amount of (subjective) absence locations
Better insight of the number of mosquitoes at a specific location	Limited insight in the number of mosquitoes at a specific location
Less extra information can be collected	More extensive information can be collected via an online questionnaire
CO_2 used to attract mosquitoes	No specific attractor used
Higher accuracy of catch location (coordinates)	Lower accuracy of catch location (PC6-level)
Large differences in performance of different trap types (Hiwat et al., 2011)	Performance depends on communication strategy
Reports from both uninhabited and inhabited areas	Reports are only made from inhabited areas
High costs	Low costs (Crall et al., 2010; Dickinson et al., 2010)
Hard to cover a broad geographic scale	Possible to cover a broad geographic scale (Crall et al., 2010; Dickinson et al., 2010)

3.2.2 Other datasets

Next to the *Muggenradar* data, additional datasets are used (Table 4). The population data is mainly used for the analysis of demographic factors (RQ2). PC5, neighbourhood, building, water, and terrain data is used for the environmental factor analysis, and the hotspot analysis (RQ3 and 4).

Table 4: Secondary data used

Dataset	**Source**	**Year**
PC5 areas	Bridgis (2014)	2014
Population per PC6 area	CBS (2010)	2010
Neighbourhoods	CBS (2012)	2012
Buildings	BAG (Building Administration)	2012
Water bodies	Kadaster (Top10NL)	2014
Terrain	Kadaster (Top10NL)	2014

3.3 Pre-processing

The *Muggenradar* dataset as obtained from the Laboratory of Entomology cannot be used directly for spatial pattern analysis. It contains information that is redundant or not necessary for this study, and although the postal code is known for each report, the geographical location in coordinates is not known. It is therefore necessary to pre-process the data to a format that can be used for spatial analysis.

A flowchart of the pre-processing of the data can be seen in Figure 2. The results will be saved to comma-separated values (CSV) files, a format that can be read by many different software packages, and only uses a small storage space.

First, all reports will be linked to geographical coordinates based on their PC6 area. Since the study area is the Netherlands, the Dutch Rijksdriehoekstelsel coordinate system will be used. The geo-coding service of the Dutch Nationaal Georegister, which takes addresses or postal codes as

input, and gives coordinates as output, will be used for this purpose. Thereafter, the *Culicidae* field in the data needs to be edited, in order to change the values ranging from 0 to 3 (0 = No *Culicidae*, 1 = *Culicidae*, 2 = Unidentifiable, 3 = empty envelope) to only 0 (No *Culicidae*) and 1 (*Culicidae*). Although there is a probability that reports with values 2 and 3 could possibly be mosquitoes, they are considered as not being one. Third, the '#N/A' value that is used in the original *Muggenradar* dataset will be replaced by 'Null', since the string '#N/A' might give errors in specific software packages. Thereafter, duplicates in the data should be removed. In some cases, respondents submitted the online questionnaire more than once while only sending one envelope. In other cases, respondents sent the questionnaire multiple times with different reports. The reports will in both cases be combined into one record. In order to check if the reports were made by the same respondent, the report needs to be made with the same postal code and e-mail address.

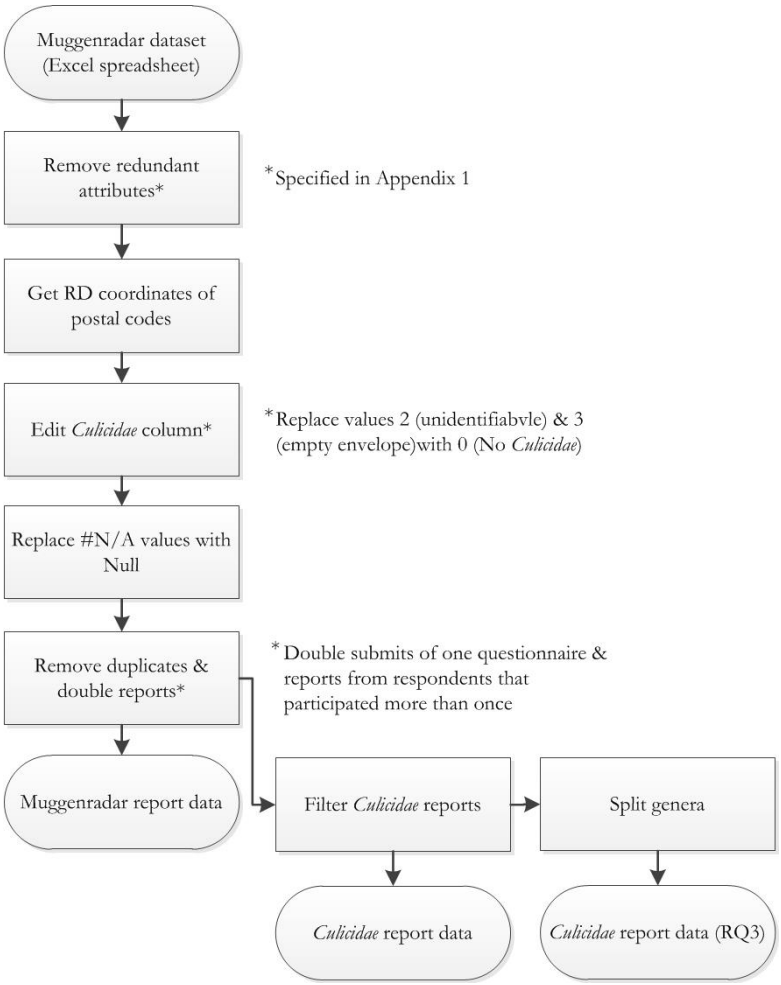


Figure 2: Flowchart of the pre-processing of the Muggenradar data

All steps will be performed twice, once for the January and February dataset, and once for the August and September dataset. The duplicate-removed datasets form the two primary files of the *Muggenradar* report data. Next to these files, four other files will be created. The primary files will be filtered for only *Culicidae* reports, forming the '*Culicidae* report data'. As a pre-processing step for RQ3, the different genera will be split up into separate records. If respondents sent more than one mosquito of a different genera, both genera were included in the original *Muggenradar* data (e.g. '*Culex* & *Culiseta*').

The pre-processing of the secondary data will be explained per research question.

3.4 Spatial Distribution Analysis

Hypotheses are formulated based on previous studies on crowdsourcing and mosquito habitats (Chapter 2), and the differences between crowdsourcing and traditional collection methods (Table 3).

- More mosquito observations are reported from areas with a higher population density.
- *Culex* mosquito reports are segregated from *Culiseta, Anopheles* and *Aedes* reports.
- Mosquito presence reports are segregated from the absence reports.

In order to test these hypotheses, the spatial distribution of the reported observations is analysed.

3.4.1 Ripley's K Function for Spatial Clustering

First, the amount of clustering in the data is tested using Ripley's K function, a method that is increasingly used in ecology to analyse the distribution of point patterns (Haase, 1995). The method tests if the input data is significantly clustered or dispersed, by calculating the number of neighbours of a point within a range up to a given radius for the input data. In this case, the maximum range is set to 10,000 m. This is compared to the number of points within the same radius range when using a Poisson distribution, a point pattern where complete spatial randomness (CSR) is assumed (Haase, 1995). To test if the amount of clustering or dispersion is statistically significant, envelopes are computed by using 50 simulations of CSR. The method works for study areas of every shape, although an edge correction should be applied when using complex study areas such as the Netherlands. To keep computation times reasonable, the border of the Netherlands is simplified for the edge correction. The Kest function of the R package spatstat is used to compute Ripley's K function (Baddeley & Turner, 2005).

3.4.2 Kernel Density Estimation for Hotspot Localization

Ripley's K function checks the amount of clustering in the data, but does not return information on where the clustering appears. Therefore, a kernel density estimation (KDE) was applied on the *Muggenradar* data. KDE is an interpolation method that is appropriate for individual point features (Levine, 2005). It is an effective tool to identify hotspots within point patterns by creating a continuous surface that defines the point density for an area (Bithell, 1990). An area with a higher point density can be interpreted as a mosquito report hotspot. The two computed KDE rasters use a bandwidth of 4000 m, which is a useful bandwidth to inspect the raster visually to identify hotspots. The rasters have a cell size of 50 metres.

3.4.3 Clustering versus Urban Area: Pearson's Correlation Coefficient

To check the first hypothesis of research question 1 (more mosquito observations are reported from areas with a higher population density), the correlation between mosquito reports and urbanisation was calculated using Pearson's correlation coefficient. For this purpose, additional KDE rasters were computed. A bandwidth of 4000 m was useful for visual inspection, since the hotspots are clearly visible when using a large bandwidth. For analysis of the mosquito report density, however, the bandwidth should be chosen more rationally. Therefore, a bandwidth that is similar to mosquito flight ranges is used.

The most abundant mosquito species, *Culex* mosquitoes, have the smallest flight range of the reported genera, with a range of about 500 up to 1000 m (Cui et al., 2013; Greenberg et al., 2012; Service, 1997; Reeves et al., 1948; Tsuda et al., 2008). The other mosquito genera are known to have larger flight ranges: *Anopheles* mosquitoes fly up to 3 km (Becker et al., 2010), and *Culiseta*

mosquitoes can fly distances up to 5 to 12 km (Howard et al., 1989). For the KDE bandwidth, only appetential flights, which are self-steered flights with feeding or egg-laying as main goal (Service, 1997), were taken into account. A bandwidth of 1500 m is used, which is slightly more than the flight range of *Culex* mosquitoes, but it compensates for the longer ranges of the other genera.

The raster values of the KDE are extracted to the report point features, and these point values form the dependent variable input for Pearson's correlation coefficient function. The Dutch Centraal Bureau voor de Statistiek (CBS) provides data on the number of addresses per km^2 for each PC6 area. This number represents the urbanisation of a certain area, and is used as the independent variable.

3.4.4 Spatial Segregation of Genera

A probabilistic model by Veech (2013), as implemented in the R package cooccur by Griffith et al. (2014) was used to analyse the co-occurrence patterns of the different genera. Presence-absence matrices are created for the January and February, and August and September data. These matrices contain the different genera as columns, and sites as rows, and is filled with binary values (1 = genus is present, 0 = genus is absent). PC4 areas, which are larger than PC6 areas, are used as sites to make sure that every site contains enough reports for analysis.

The co-occurrence function calculates the observed and expected frequencies of co-occurrence between each genera pair. This expected frequency is based on a random and independent distribution of each genus. The function returns the probabilities that a more extreme value of co-occurrence (higher or lower) could be obtained by chance.

3.4.5 Presence and Absence Reports

The third hypothesis was tested by using correlation between presence and absence reports. Absence data is defined as a report in which the respondent did not see a mosquito (column MOSQ_SEEN = No), whereas a presence report is defined as reports in which the respondent did see one or more mosquitoes (MOSQ_SEEN = Yes).

To compute the correlation between presence and absence reports, a KDE raster is created out of the report point features. Again, a bandwidth of 1500 m is used, in combination with a courser cell size of 500 m. Using a smaller cell size would result in too long computation times when computing the correlation. The values of both KDE rasters are copied into arrays, serving as the input variables for the Pearson's correlation function. If a cell in both the presence and absence data had a no data value (-9999) or a value of 0 (no reports per km^2), the cells were masked to keep the size of the arrays manageable.

3.5 Demographic Characteristics

To find out how the respondents of the *Muggenradar* project can be demographically characterised, the mosquito reports were correlated with demographic factors. CBS demographic data per PC6 area is linked to the mosquito report data. These demographic factors are used as the independent variable for Pearson's correlation. Again, the mosquito report densities as extracted from the KDE rasters are used as dependent variable. A script is used to correlate each demographic factor with the dependent variable (mosquito report density) and to export the scatterplots, r, and p-values.

3.6 Environmental Predictors for Mosquito Presence

The following hypotheses are formulated, based on the literature study (Chapter 2):

- A closer proximity to stagnant water will result in more mosquito reports.
- A closer proximity to deciduous forest or deciduous trees will result in more mosquito reports.

The results of the spatial distribution analysis (RQ1, Chapter 4.1) made clear that more mosquito reports are made from more urbanised areas. The literature review already showed that *Culex* mosquitoes, the most reported genus, appear often in dark and moist cellars of large buildings in urban areas (Becker et al., 2010). The Dutch government (Rijksoverheid) uses 1976 as a threshold value to distinguish between low-energy (build before 1976), and high-energy housing (build after 1976) (http://www.rijksoverheid.nl/onderwerpen/huurwoning/puntensysteem-huurwoning/puntensysteem-en-energielabel). This classification is based on a variety of factors, including isolation. This study assumes that low-energy housing, and thus less isolated houses, are more accessible for mosquitoes, and thus more reports are expected from these older houses. Two additional hypotheses were formulated:

- More mosquito reports are made from buildings built before 1976
- A higher human population density will result in more mosquito reports.

A flowchart of the methodology of this research question can be seen in Appendix 3.

3.6.1 Study Area Selection

Because more mosquito reports were made from the more urbanised areas, the environmental factor analysis will be conducted on two urban study areas. The areas correspond to the areas with the highest values in the kernel density maps (Chapter 4.1.2). Selected study areas are Amsterdam, and the region around Rotterdam, as can be seen in Figure 3 and Figure 4. All of these areas had a high value in the kernel density rasters. A summary of the study areas can be seen in Table 5.

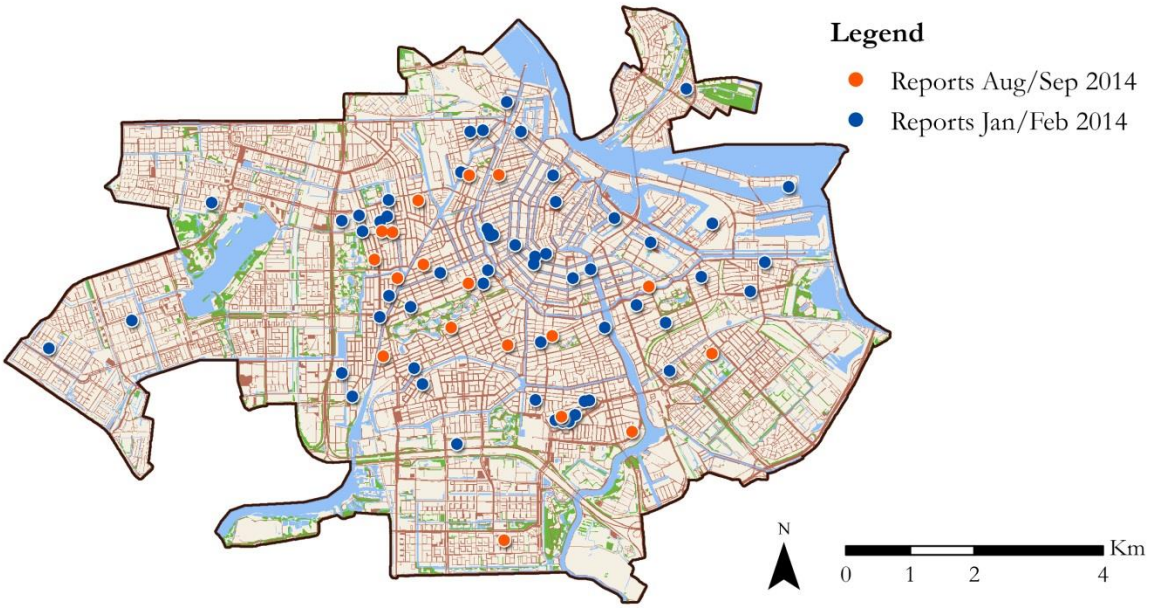


Figure 3: Mosquito reports originating from Amsterdam. The January and February reports are shown in blue, and August and September reports in red.

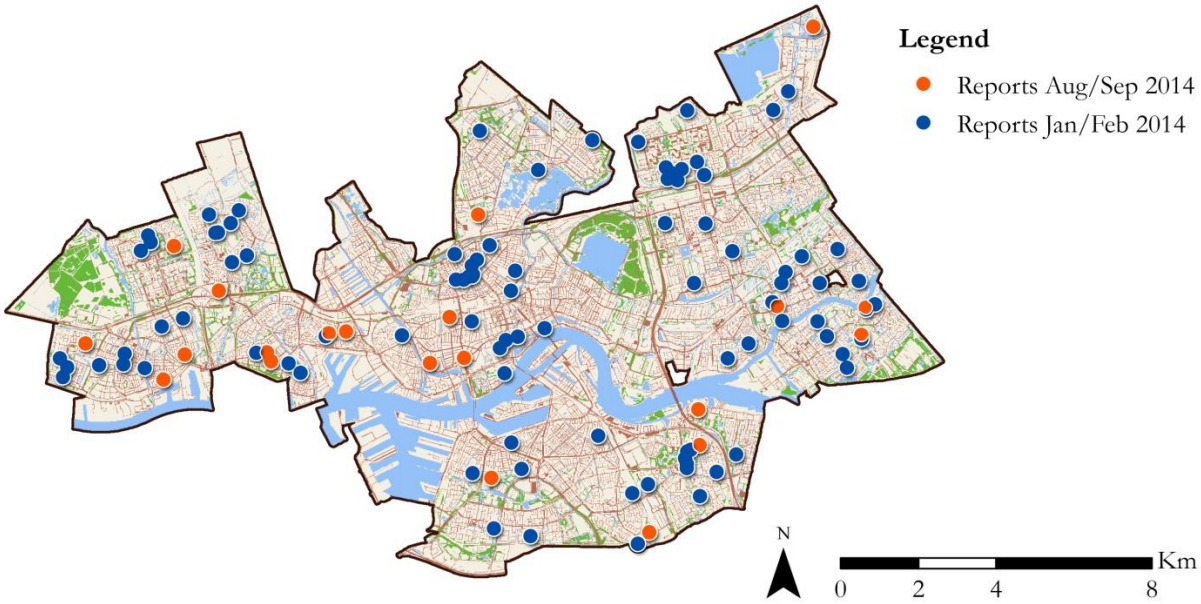


Figure 4: Mosquito reports originating from the Rotterdam region. The January and February reports are shown in blue, and August and September reports in red.

Table 5: Number of mosquito reports per study area

	Amsterdam	Rotterdam
Surface area (km²)	71.5	165.5
Jan/Feb 2014	60	95
Aug/Sep 2014	18	21
Total reports	78	116

The study areas were selected based on their level of urbanisation, by using CBS neighbourhood data. Only strongly urbanised neighbourhoods were taken into account, since the first research question showed that more reports were made from areas with a high address density. Because the address density in Amsterdam is generally higher than the density in Rotterdam, a threshold of 2500 addresses per km^2 (OAD > 2500) is used here. For Rotterdam, a threshold of 1000 addresses per km^2 is used. In order to have enough reports in each study area, additional municipalities are added to the study areas, until the number of reports in each study area was satisfying. Amsterdam already counted enough reports, and thus only consists of the neighbourhoods with more than 2500 addresses per km^2 within the municipality of Amsterdam. The study area of Rotterdam is extended with the highly urbanised neighbourhoods of Schiedam, Vlaardingen, Capelle aan den IJssel, and Krimpen aan den IJssel. A flowchart of the selection of study areas can be seen in Appendix 4.

3.6.2 Environmental Factors

The environmental factors that are used to test for their relation with the mosquito reports can be seen in Table 6.

Table 6: Environmental factors, their source and the used attribute fields

Environmental factor	Source	Attribute(s)
Building construction year	BAG (Administration of buildings)	Building construction year
Proximity to water	Top10NL Water	-
Proximity to vegetation	Top10NL Terrain	Deciduous forest
Population density	CBS Population data per PC6 area	Inhabitants
	Bridgis PC6 areas	Surface area

All datasets are provided as vector formats. In order to use the environmental factors for further analysis, these datasets should be converted to a raster format. Several pre-processing steps had to be performed to create these input raster datasets. A flowchart of these steps can be seen in Appendix 5. All rasters have a cell size of 5 metres: enough to show detail in the environmental datasets, but high enough to keep computation times reasonable.

Proximity to Water

The Top10NL Water dataset is used to compute the proximity to water for each raster cell. It would be preferred to use only stagnant water for the environmental factor analysis. Data on the locations of stagnant water, however, is not available. Moreover, waterways may provide larval breeding habitat in pools on margins of these waterways (Gleiser & Zalazar, 2009). All water bodies are therefore taken into account for the analysis.

First, the water bodies in the Top10NL dataset are converted to a raster format. Thereafter, the Euclidean distance of each raster cell to the closest water body is calculated. Since the location of each report is only known up to PC6 level, the mean proximity to water is calculated for each PC5 area. It would be preferred to use PC6 areas to keep the level of detail higher, but the most recent available PC6 dataset dates from 2006. Postal code areas in the Netherlands are updated regularly, so using a more recent dataset is desired. Therefore, the 2014 PC5 dataset provided by Bridgis is used, data from the same year as the *Muggenradar* project data. The raster with the mean proximity to water for each PC5 area will form the final input for the environmental factor analysis. An example of the resulting raster for Rotterdam can be seen in Figure 5.

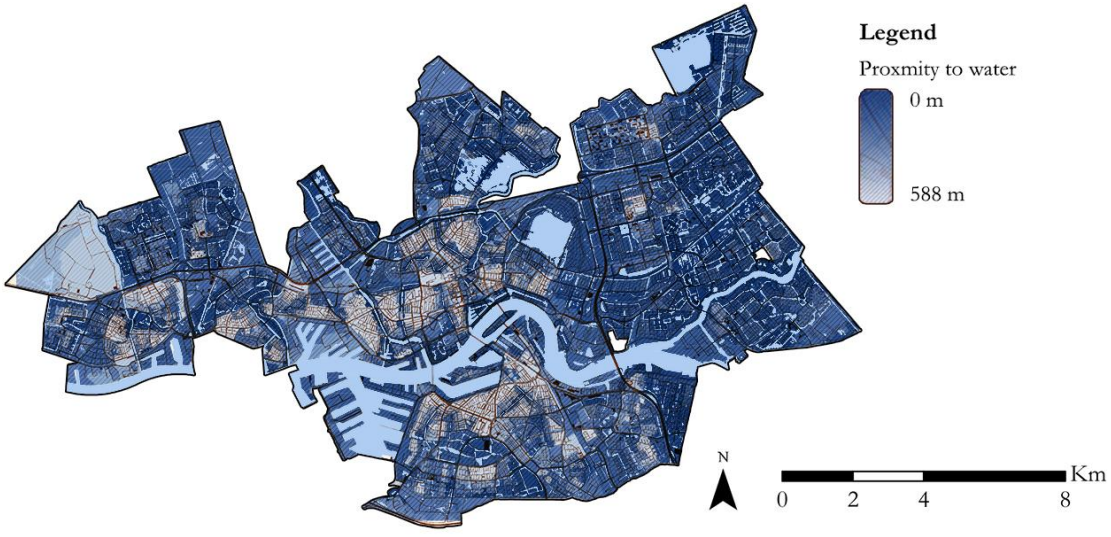


Figure 5: Proximity to water for Rotterdam. This is not the final input raster for the analysis, since that one consists of the mean proximity to water per PC5 area.

Proximity to Deciduous Forest and Trees

The proximity to deciduous forest and trees is calculated in the same way as the proximity to water. The locations of deciduous forest and trees are derived from the Top10NL terrain dataset.

The mean proximity to deciduous forest has been computed per PC5 area, for the reason that a recent PC6 area dataset is not available. An example of the output raster for Rotterdam can be seen in Figure 6.

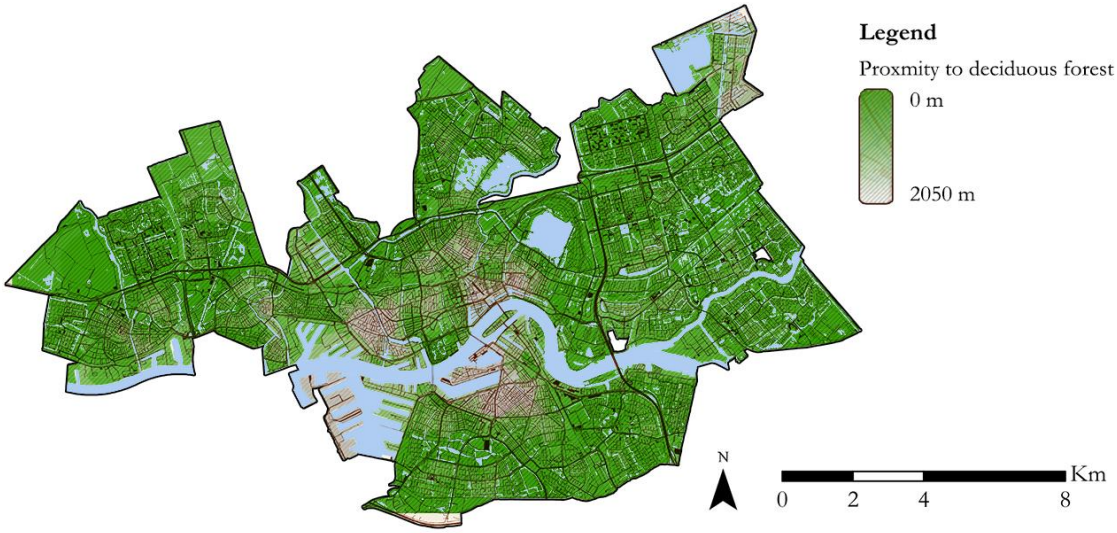


Figure 6: Proximity to deciduous forest and trees for Rotterdam. This is not the final input raster for the analysis, since that one consists of the mean distance to deciduous forest per PC5 area.

Building construction year

In order to convert the BAG dataset of buildings to a useful raster dataset, all building features were first converted to a raster, in which the raster values represent the construction years. This raster has only values for cells that contain a building. All other cells have a specified no data value (-9999). A continuous raster in which all cells have values is needed for further analysis. Each report is made from a specific PC6 location. The exact building from which the report is made is therefore not known. For this reason, the median of the construction year of buildings is computed for each PC5 area. Again, the more detailed PC6 areas were not used because of the outdated data.

An example of the output raster for Rotterdam can be seen in Figure 7.

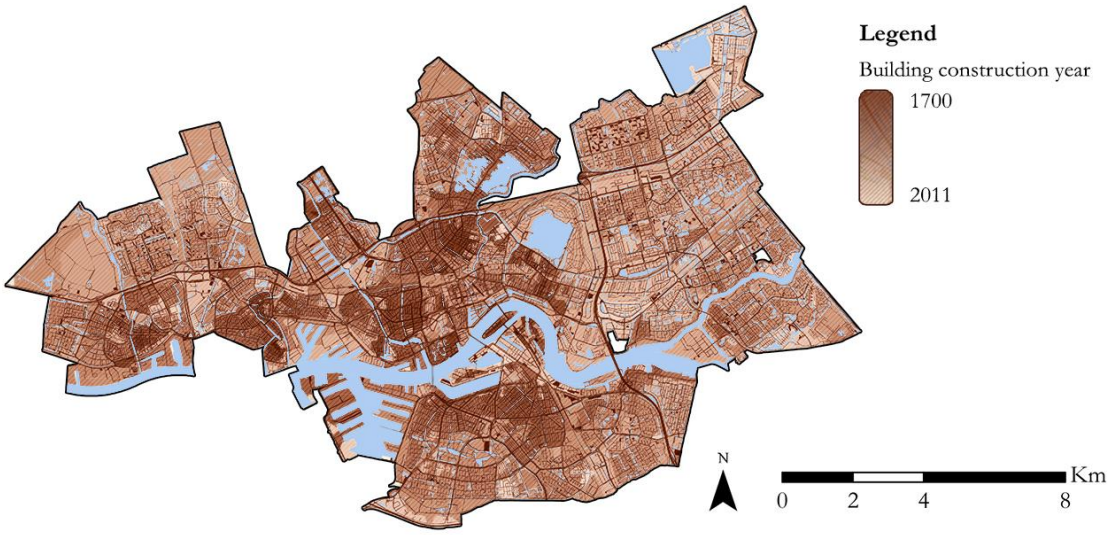


Figure 7: Building construction year raster for Rotterdam, showing the median construction years for each PC5 area.

Population Density

It became clear from the results of research question 1 that population density influences the amount of mosquito reports, when looking at the Netherlands. It is however not clear to what extent the amount of reports is influenced, and if the pattern is also visible within urban areas. Population density, described as the number of inhabitants per square kilometre, will also be used to find a relationship with the mosquito reports.

The number of inhabitants per square kilometre is computed for each PC5 area. Since population data for PC5 areas is not available, the number of inhabitants per PC6 area is first derived from CBS population data. Thereafter, the number of inhabitants of all PC6 areas belonging to one PC5 area are summed, in order to find the number of inhabitants per PC5 area. To get the number of inhabitants per square kilometre, this sum of inhabitants is divided by the surface area of each PC5 area in square kilometres.

An example of the output raster for Rotterdam can be seen in Figure 8.

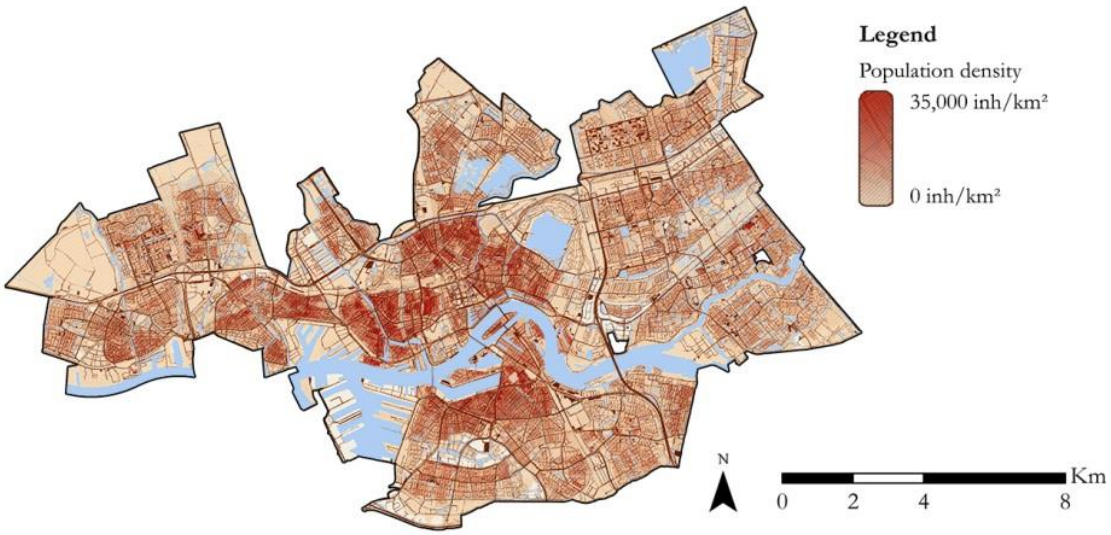


Figure 8: The population density (inhabitants per km²) for Rotterdam.

3.6.3 Environmental Factor Analysis

For each study area, season, and environmental factor, the relationship between the environmental factor and the number of mosquito reports is analysed.

Proximity to Water, Proximity to Deciduous Forest and Trees, and Population Density

To find the relation of mosquito reports with the proximity to water, proximity to deciduous forest and trees, and population density, the following method is used. First, the values of the environmental factor rasters are extracted at the locations where reports were made, resulting in a table that shows the proximity to water, proximity to deciduous forest, population density, and the building construction year for each mosquito report location.

In order to find out if the *Muggenradar* reports show a different pattern than a set of random reports, simulations were used. Separate simulations were generated for each study area and each season, with the number of randomly simulated point locations being equal to the number of *Muggenradar* reports in that certain study area. The random reports are simulated over the existing PC5 centroids, to avoid that random reports are created in uninhabited areas. Simulating random points all over the study areas, without taking the distribution of inhabitants into account would result in many points being placed in areas with a low population density. This would not represent a realistic simulation of the *Muggenradar* mosquito reports.

Therefore, the mosquito reports are divided into two classes, with the initial break value between these classes being 10, independent from what the unit of this value is. For example, all reports within a distance of 10 metres from water belong to the first class, whereas reports at a larger distance from water belong to the second class. Thereafter, the break value is gradually increased until the maximum value of the environmental factor is reached; for example, until the maximum distance to water at which a report was made is reached. The break value is increased in 80 steps. For each break value, the number of reports that fall within the first class and within the second class are counted. At the maximum break value, all mosquito reports fall within the first class. At the same time, the number of random reports that fall within each of the two classes is counted for each of the 20 simulations. Thereafter, the average of these 20 counts of the simulations is computed, and used as the expected number of reports in each class.

For each break value, a Chi Square test with one degree of freedom is done using the observed (reported) counts and the expected (simulated) counts in both classes as input. When the p-value of the Chi Square test is significant ($p < 0.05$), it is assumed that the *Muggenradar* reports differ significantly from the 20 random simulations. Odds ratios are then computed to find out how much higher the chance of getting a report in the first class is compared to a random point distribution over the two classes.

For each break value where the Chi Square test is significant, the odds ratios is calculated. These odds ratios are numbers between 0 and infinity, in which the values 0 to 1 represent a lower chance of finding a mosquito report than expected, while values of 1 and higher represent a higher chance of finding a mosquito report than expected. For the entire range of tested break values, the strongest odds ratio will be saved (odds ratio is closest to 0, or the highest value above 1) for each study area and each *Muggenradar* round. The corresponding break value will be used as value to distinguish between areas with low ($OR < 1$) and high ($OR > 1$) mosquito presence. A flowchart of these steps can be seen in Appendix 3.

Building Construction Year

Since a substantiated class break value for building construction years is given, an iterative approach to find the best break value is not needed. Using the break value 1976, the number of mosquito reports coming from PC6 areas with an older building construction year are computed, as well as the number of mosquito reports from areas with a newer construction year. These numbers are used as the observed input. In the same way, the randomly simulated reports of the 20 simulations are classified. The mean class counts of these simulations will be used as expected input for a Chi Square test with one degree of freedom. Thereafter, the odds ratios for both study areas and both *Muggenradar* rounds are computed.

3.7 Mosquito Presence Hotspots

The odds ratios of significant environmental factors that were found in research question 3 will be used as input to generate a hotspot map of three Dutch cities: Amsterdam, Rotterdam and Utrecht. Utrecht was used to validate the parameters, and thus the odds ratios, used for hotspot maps of Amsterdam and Rotterdam.

3.7.1 Creating the Hotspot Maps

A flowchart of the methodology for the hotspot map creation can be seen in Appendix 6. The odds ratios as computed in research question 3 are used as input. Out of these odds ratios, an odds ratio map is created for each of the environmental factors and for each *Muggenradar* round. Each environmental factor raster is divided into two classes, using the break value as defined in research question 3 as class break. For the environmental factors proximity to water, proximity to deciduous forest, and population density, the break value where the odds ratio is strongest will be used, as described in the methodology of research question 3.

These odds ratios represent the chance of finding mosquito reports in the first class, compared to finding reports in the second class. The inverse odds ratio (1/OR) was computed to find the chance of finding mosquito reports in the second class. A raster cell belonging to the first class (raster value < break value) gets the odds ratio assigned as its value, while cells belonging to the second class (raster value >= break value) get the inverse odds ratio (1/OR) assigned as cell value. The output consists of eight odds ratio rasters (water, deciduous forest and trees, building construction year, and number of inhabitants per km², for both January/February and August/September), in which cells can have two values: the odds ratio or inverse odds ratio. For each Muggenradar round, the four odds ratio rasters are multiplied in order to combine all environmental predictors, and eventually to find hotspots of mosquito presence within the January/February reports and the August/September reports. At places where the multiplied odds ratio is higher than 1, the mosquito presence is expected to be higher than when using a random distribution of reports. A multiplied odds ratio lower than 1 means that the expected mosquito presence is lower than when using random simulations.

To check if the hotspot maps are accurate, the number of reports within expected high presence areas are counted, as well as the number of reports in expected low presence areas. If a majority of the number of reports falls within high presence areas, and the hotspot map is based on enough environmental factors, the hotspot map can be seen as accurate.

3.7.2 Validation with Utrecht

The hotspot maps of Amsterdam and Rotterdam could show different results, since the odds ratio maps that were used as input for the hotspot maps are derived from the significance in the

Chi Square tests. These significance can differ for both study areas, and for each *Muggenradar* round. In order to find out if the parameters used for the hotspot maps of Amsterdam and Rotterdam are also valid for the mosquito reports of Utrecht, a total of four hotspot maps will be created for Utrecht. Two maps (January/February & August/September) are based on the odds ratios and break values of Amsterdam, while the other two maps are based on the parameters of Rotterdam.

Thereafter, the number of mosquitoes that are reported within high presence areas (raster values > 1) will be counted for each of the hotspot maps (Rotterdam, Amsterdam, Utrecht using Amsterdam's parameters, and Utrecht using Rotterdam's parameters). If the hotspot maps are correct, a majority of the reports is expected to be reported from high presence areas for all study areas.

3.8 Software

3.8.1 Analysis

The Python programming language is used for the analysis of the *Muggenradar* data. Additional Python libraries are used to make the analysis easier. The *pandas* library (McKinney, 2010) is used for quick and easy handling of CSV files. *Numpy* (Oliphant, 2007), a package for scientific computing with Python, is used to handle large array objects. *SciPy* (Oliphant, 2007) is used to access statistical methods such as Pearson's correlation. *Matplotlib* (Hunter, 2007) is used to plot figures and graphs. For spatial analysis, OSGEO's GDAL and OGR libraries (GDAL, 2014), and Esri's ArcPy library are used. Since *SciPy* and *Spatstat* did not cover all needed statistical methods, there was a need for additional statistical methods. The software package R (R Core Team, 2013) covers more statistical methods, and therefore the Rpy2 Python library is used to access R packages from Python code. Two R packages are used: s*patstat* (Baddeley & Turner, 2005) and *cooccur* (Griffith et al., 2014).

3.8.2 Visualisation

Python and R are very suitable when it comes to fast processing of data. For quick and easy visualisation of the data output, however, these programs are less suitable. Esri ArcMap is therefore used to view and examine the data, and to create maps. The Python library *Matplotlib* is used to create plots.

4 Results

4.1 Spatial Distribution Analysis

4.1.1 Ripley's K Function for Spatial Clustering

The *Muggenradar* dataset shows significant clustering over the whole range of 0 to 10,000 metres in both January and February, and August and September reports. The output of Ripley's K function can be seen in Figure 9, where the black line represents the observed value of K(r), and the dotted red line the simulated value of K(r) for a simulated point pattern with CSR. The grey area shows the envelope which represents the outer boundaries of the 50 simulations of CSR. If the observed value is higher than the upper envelope, there is significant clustering in the dataset.

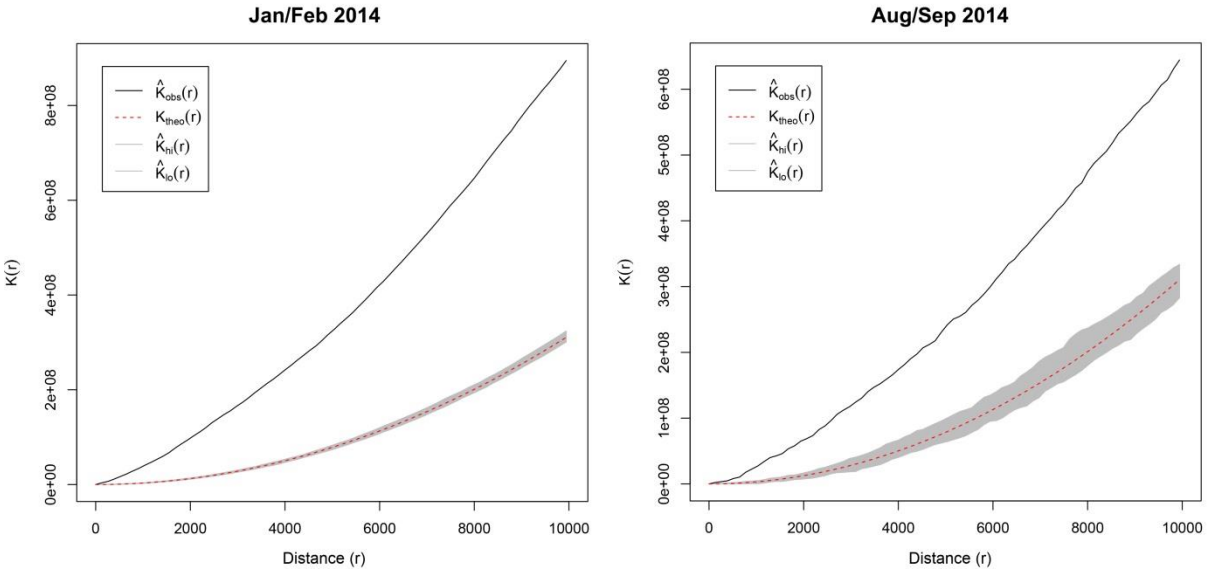


Figure 9: Ripley's K estimate for January and February (left), and September and August (right) point patterns. Kobs(r) is the observed value of K(r) for the data pattern, Ktheo(r) is the theoretical value of K(r) for a simulated CSR, Khi(r) is the upper pointwise envelope of K(r) from simulations, and Klo(r) the lower pointwise envelope of K(r) from simulations.

4.1.2 Kernel Density Estimation for Hotspot Localisation

Figure 10 shows that the total amount of reports was higher in January and February, resulting in higher report density values. In both January and February, and August and September, clusters are visible in the regions of Amsterdam, Rotterdam, Utrecht, Hilversum, and Wageningen. In August and September, additional clusters are located in the Randstad area.

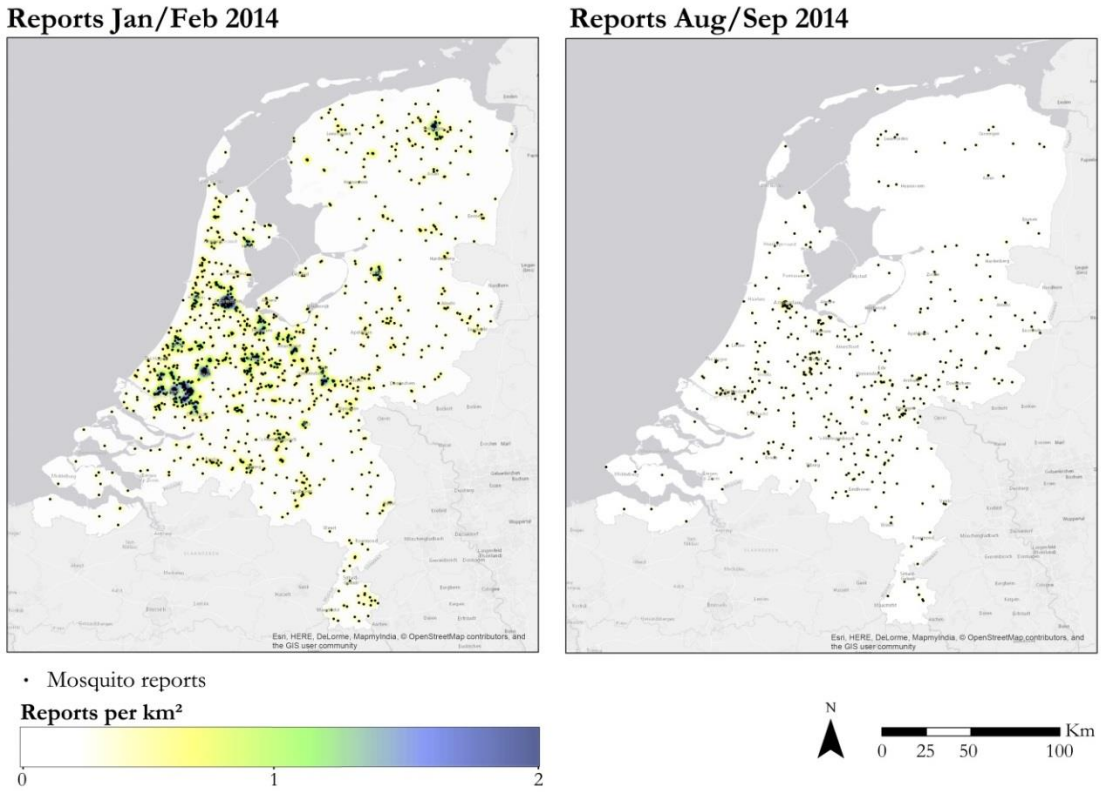


Figure 10: Kernel density maps of the January and February 2014 reports (left) and August and September 2014 reports (right)

4.1.3 Clustering versus Urban Area: Pearson's Correlation Coefficient

There is a significant positive correlation between the mosquito reports and the number of addresses per km^2, meaning that more mosquitoes are reported from more urbanised areas, for both January and February, and August and September 2014 (Figure 11). The first hypothesis can therefore be accepted.

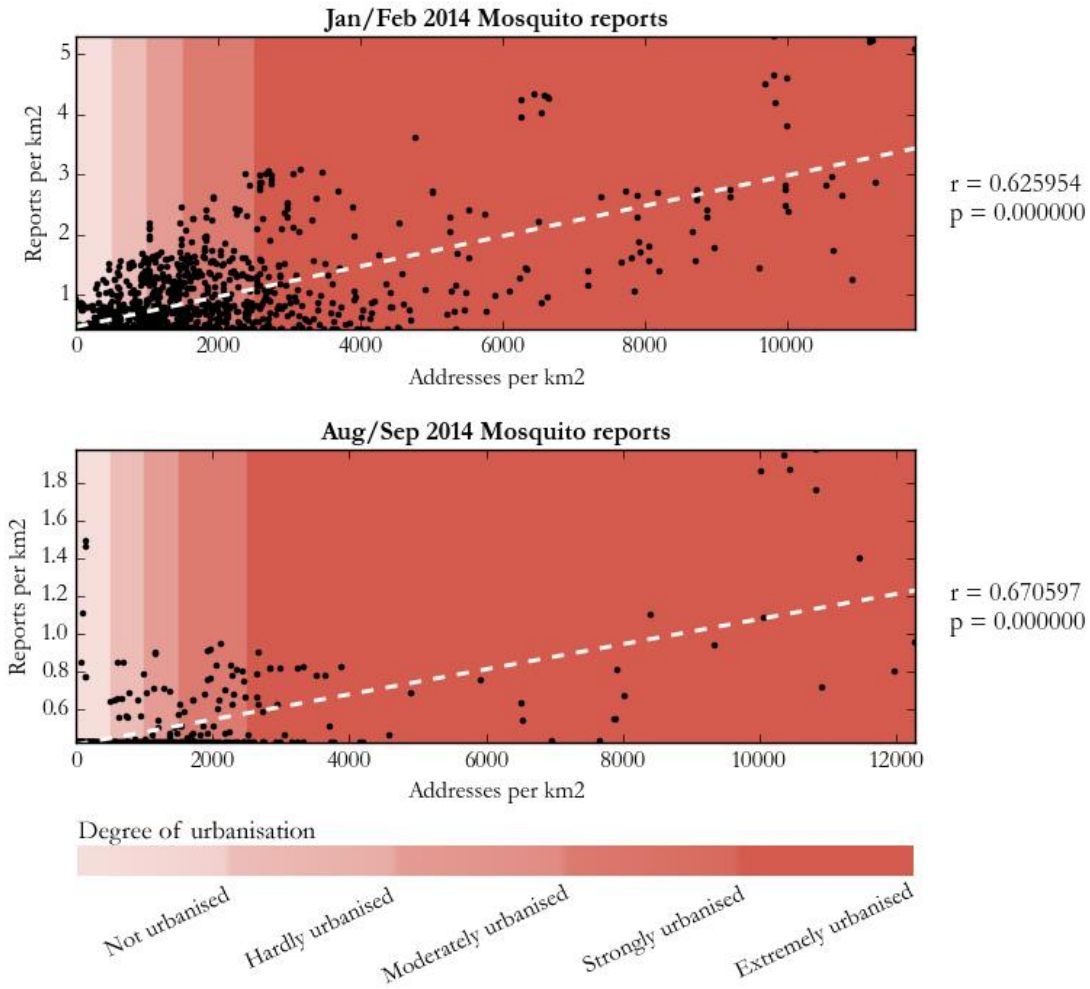


Figure 11: Correlation between expected mosquito reports per km² and addresses per km², Pearson's correlation coefficient (r) and its p-value. The red background colour represents urbanisation classes as used by CBS, in which the lightest colour represents 'not urbanised' (0-500 addresses per km2), and the darkest 'strongly urbanised' (2500 or more addresses per km2).

4.1.4 Spatial Segregation of Genera

Figure 12 shows the co-occurrence matrices of all reported genera in January and February, and August and September. Table 7 and Table 8 show the probability tables of co-occurrences for respectively January and February, and August and September. In January and February, all genera pairs have a negative co-occurrence, meaning that according to the co-occurrence analysis, the genera are significantly segregated. In August and September, *Culex* mosquitoes have a negative co-occurrence with all other genera. *Culiseta*, *Anopheles,* and *Aedes* mosquitoes have a random co-occurrence, meaning that there is no significant clustering or segregation.

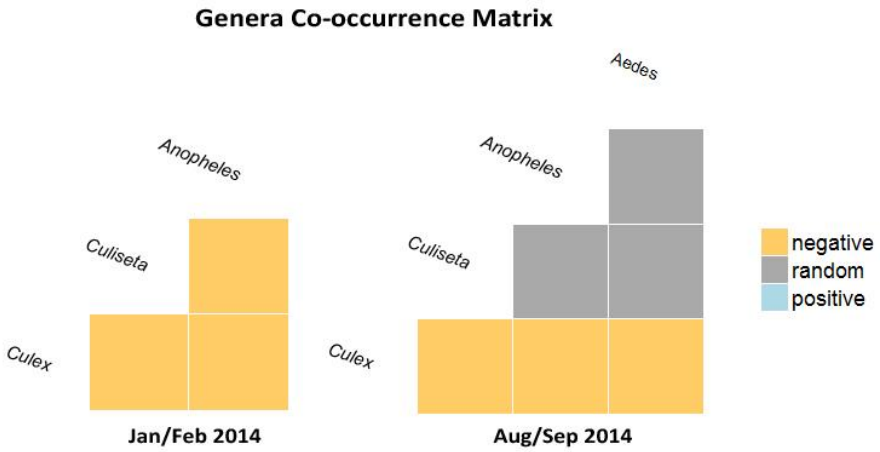


Figure 12: Co-occurrence matrices of all reported genera in January/February (left) and August/September (right)

Table 7: Probability table of the co-occurrence analysis between the different genera (January/February 2014)

Genera 1	**Genera 2**	Sites with G1	Sites with G2	Obs. co-occur	Prob. co-occur	Exp. co-occur	Prob. exp < obs	Prob. exp > obs
Culex	***Culiseta***	601	403	129	0.288	264.1	0	1
Culex	***Anopheles***	601	89	39	0.064	58.3	0.00005	1
Culiseta	***Anopheles***	403	89	22	0.043	39.1	0.00006	0.99998

Table 8: Probability table of the co-occurrence analysis between the different genera (August/September 2014)

Genera 1	**Genera 2**	**Sites with G1**	**Sites with G2**	**Obs. co-occur**	**Prob. co-occur**	**Exp. co-occur**	**Prob. exp < obs**	**Prob. exp > obs**
Culex	***Culiseta***	330	27	14	0.06	23.1	0.00001	1
Culex	***Anopheles***	330	45	13	0.1	38.6	0	1
Culex	***Aedes***	330	20	7	0.045	17.1	0	1
Culiseta	***Anopheles***	27	45	3	0.008	3.2	0.61014	0.6346
Culiseta	***Aedes***	27	20	1	0.004	1.4	0.58275	0.77525
Anopheles	***Aedes***	45	20	1	0.006	2.3	0.2959	0.92221

4.1.5 Presence and Absence Reports

No strong positive or negative correlation between presence and absence reports was found (Figure 13), meaning that there is a random pattern between the two variables. The p-value of the August and September reports is significant, but the positive r-value is rather low. A negative correlation was expected, which would represent the fact that mosquito presence reports are spatially segregated from the absence reports. The third hypothesis should thus be rejected.

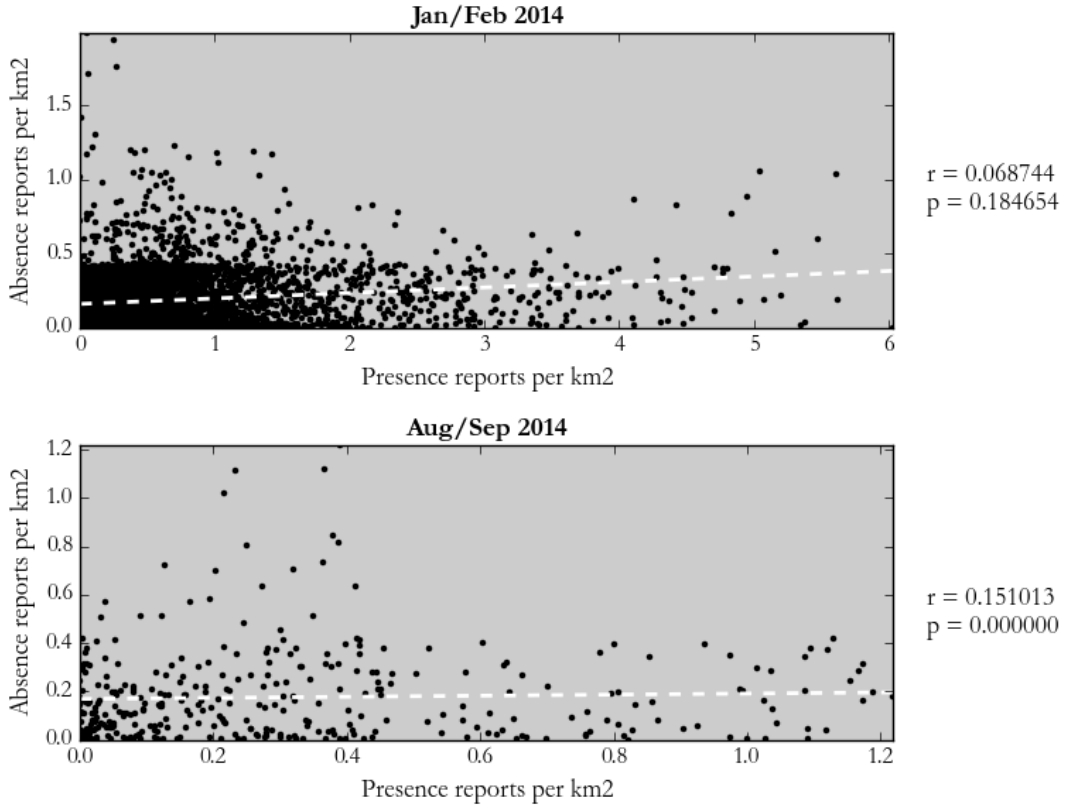


Figure 13: Scatter plot of the correlation between presence reports per km² and absence reports per km² for both January and February (top) and August and September (bottom), Pearson's correlation coefficient (r) and its p-value

4.2 Demographic Characteristics

The correlation results can be seen in Table 9. Scatter plots of the significant demographic factors versus the mosquito report densities can be found in Appendix 7. As can be seen, none of the demographic factors has a strong correlation with the reports, although some factors have a significant positive or negative correlation.

The percentages of 0 to 14 year olds and people older than 75 in a PC6 area have a negative correlation with the report density, meaning that an area with more children or more elderly people will result in less reports. An area with many 25 to 44 year olds, however, will result in more mosquito reports. The percentage of non-western immigrants has a weak positive correlation with the number of mosquito reports. The percentage of single households has a positive correlation, while the percentage of households with two parents has a negative correlation.

From Table 9 can be concluded that many mosquitoes are likely reported by 25 to 44 year olds without children, although the correlations are not that strong.

Table 9: Correlation coefficients (r) and their significance (p) for each of the tested demographic variables with the mosquito reports. Significant positive correlations are shown in green, while significant negative correlations are shown in red

	Jan/Feb 2014		Aug/Sep 2014	
Demographic variable	r	p	r	p
Number of Males	-0.036	0.482	0.026	0.360
Number of Females	-0.056	0.276	0.039	0.169
Percentage 0-14 y/o	-0.080	0.129	-0.105	0.000
Percentage 15-24 y/o	-0.067	0.201	0.030	0.292
Percentage 25-44 y/o	0.106	0.042	0.194	0.000
Percentage 45-64 y/o	0.009	0.862	-0.137	0.000
Percentage 65-74 y/o	-0.028	0.598	-0.070	0.013
Percentage 75+ y/o	-0.003	0.947	-0.048	0.091
Non-western immigrants	0.177	0.001	0.224	0.000
Average household size	-0.214	0.000	-0.205	0.000
Percentage single households	0.187	0.000	0.217	0.000
Percentage households with one parent	-0.011	0.837	0.084	0.003
Percentage multi-person households (without children)	-0.031	0.554	-0.148	0.000
Percentage households with two parents	-0.207	0.000	-0.183	0.000
Housing stock	-0.030	0.598	0.063	0.036
House value	-0.193	0.019	-0.115	0.008

4.3 Environmental Predictors for Mosquito Presence

4.3.1 Proximity to Water

The results of the iterative Chi Square tests to analyse the effect of water on the presence of mosquitoes can be seen in Figure 14 (Amsterdam) and Figure 15 (Rotterdam). The Chi Square tests for Amsterdam in January and February show significance at break points with a large proximity to water. The strongest odds ratio (0.190) can be found at 345 metres from water (Table 10), meaning that significantly more mosquito reports are made at a larger proximity to water.

The tests for August and September show no significance. This could be caused by the low amount of reports, resulting in a pattern that does not differ significantly from random simulations.

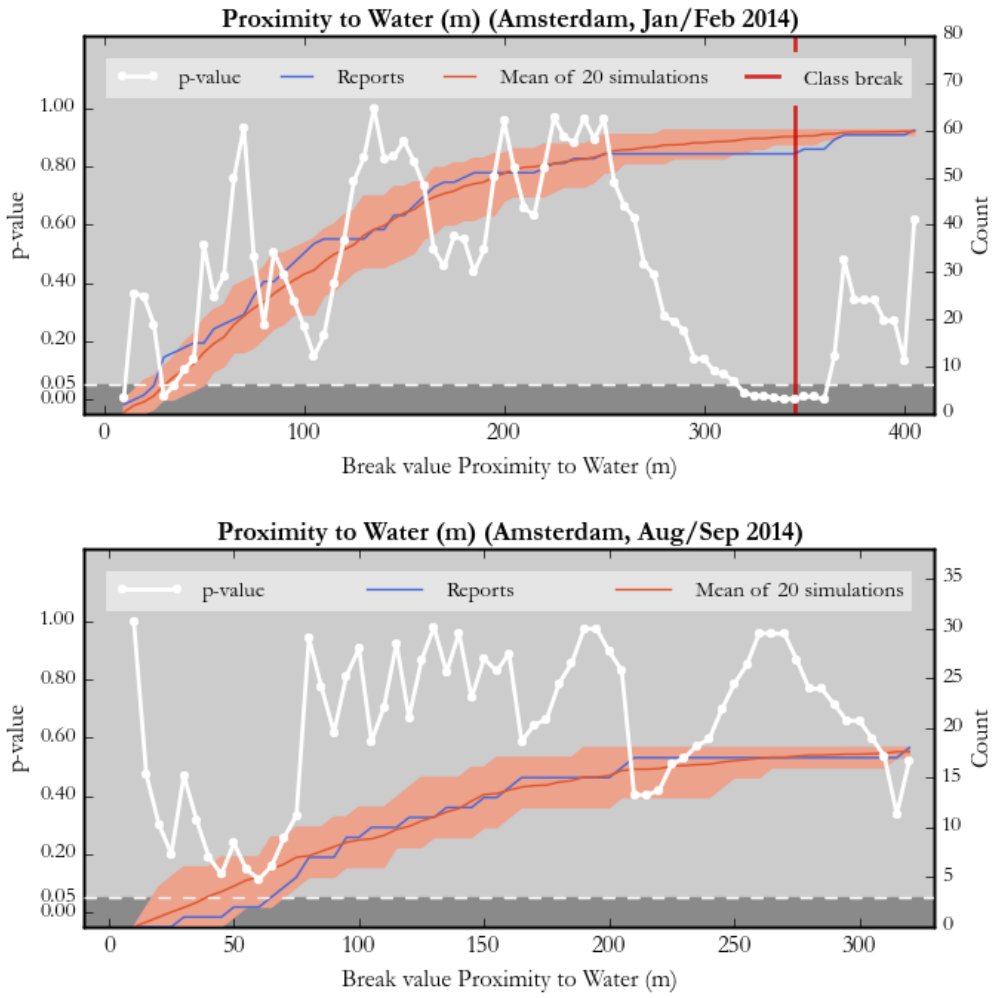


Figure 14: Line graphs showing the Chi Square test's p-value for all tested class breaks for proximity to water in Amsterdam in Jan/Feb (top), and Aug/Sep (bottom). The white line represents the p-value. At break values where the white line exceeds the dotted line at 0.05, the p-value is significant. The blue line represents the number of reports that are made at a distance from 0 metres to the distance of the break value. The red line represents the count of the mean of 20 simulations for each break value.

The tests for Rotterdam (Figure 15) show some significance when the break value is placed between approximately 100 and 160 metres. When using this break value, the number of reports is significantly higher than the expected number of reports (the mean of the simulations). The blue report line stays above the red line for all break values, meaning that in Rotterdam during January and February 2014, more mosquito reports are made at a closer proximity to water, than one would expect based on a set of random report simulations. The strongest odds ratio can be found at 112 metres from water. The odds ratio of getting a mosquito report within a distance of 112 metres from water is 1.851 (Table 10), meaning that less mosquito reports are made at a proximity of more than 112 metres to water.

The Chi Square tests for August and September however, show no significance. This could again be caused by the low number of reports during this round of the *Muggenradar* project.

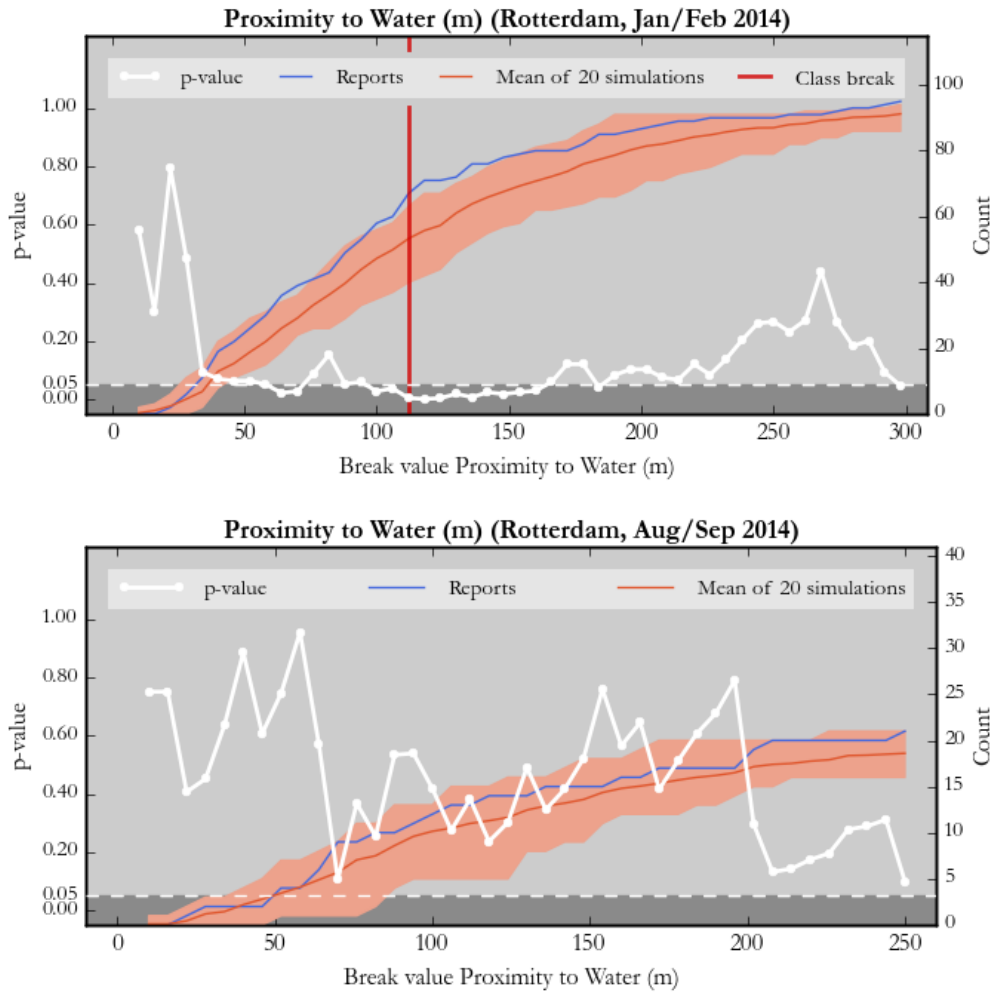


Figure 15: Line graphs showing the Chi Square test's p-value for all tested class breaks for proximity to water in Rotterdam in Jan/Feb (top), and Aug/Sep (bottom). The white line represents the p-value. At break values where the white line exceeds the dotted line at 0.05, the p-value is significant. The blue line represents the number of reports that are made at a distance from 0 metres to the distance of the break value. The red line represents the count of the mean of 20 simulations for each break value.

Table 10: Overview table showing the strongest odds ratio and corresponding break value for each study area and Muggenradar round for the proximity of water

	Odds ratio	Break value
Amsterdam (Jan/Feb)	0.190	345
Amsterdam (Aug/Sep)	1.000	No significance
Rotterdam (Jan/Feb)	1.851	112
Rotterdam (Aug/Sep)	1.000	No significance

4.3.2 Proximity to Deciduous Forest and Trees

The Chi Square test's p-value results for the proximity to deciduous forest and trees can be seen in Figure 16 (Amsterdam) and Figure 17 (Rotterdam). The tests for Amsterdam's January and February reports show significance at approximately 200 to 600 metres, where the number of expected reports based on simulations is larger than the number of observed reports. This means that the number of mosquito reports at a closer proximity to deciduous forest is lower than one

would expect when taking random report locations into account. The most significant break value is 486 metres from deciduous forest or trees. The odds ratio of finding a mosquito report within this range is 0.271 (Table 11), so more mosquitoes are reported at a proximity to deciduous forest larger than 486 metres.

The result of the Chi Square tests for the August and September reports of Amsterdam (Figure 16) show no significance at all, meaning that the reports do not differ significantly from a random pattern.

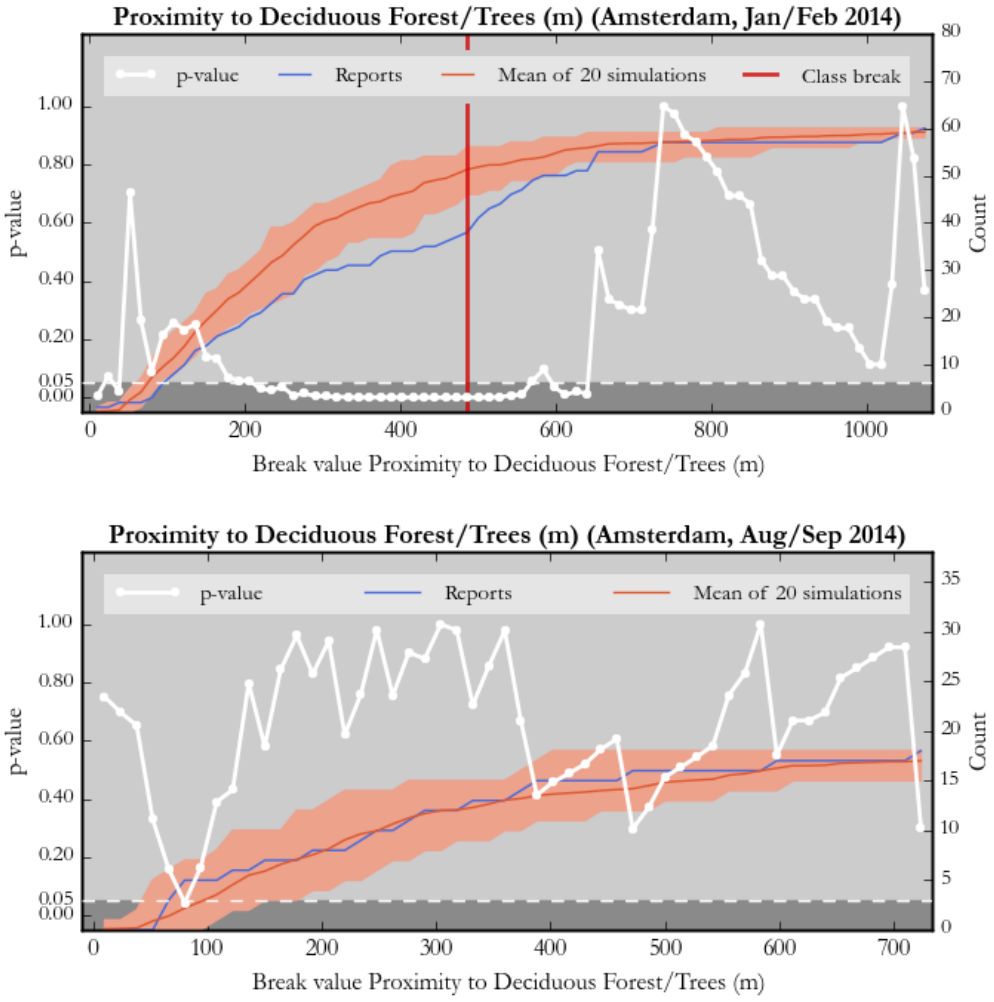


Figure 16: Line graphs showing the Chi Square test's p-value for all tested class breaks for proximity to deciduous forest/trees for Amsterdam in Jan/Feb (top), and Aug/Sep (bottom). The white line represents the p-value. At break values where the white line exceeds the dotted line at 0.05, the p-value is significant. The blue line represents the number of reports that are made at a distance from 0 metres to the distance of the break value. The red line represents the count of the mean of 20 simulations for each break value.

The results of the Chi Square tests in Rotterdam (Figure 17) of the January and February data show a clear significance from 60 to 650 metres, where the number of reports is higher than the number of expected reports based on simulations. This would mean that a closer proximity to deciduous forest would result in more mosquito reports than when looking at a random simulation of points, which is in line with the hypothesis. The strongest odds ratio can be found at 105 metres. The corresponding odds ratio for finding a mosquito report within this distance is 3.381 (Table 11).

The results of the August and September reports of Rotterdam show significance at 314 metres from deciduous forest. Finding a mosquito report within this distance is 4.071 times more likely

than finding a report at a distance further away from deciduous forest (Table 11). This is again in line with the hypothesis.

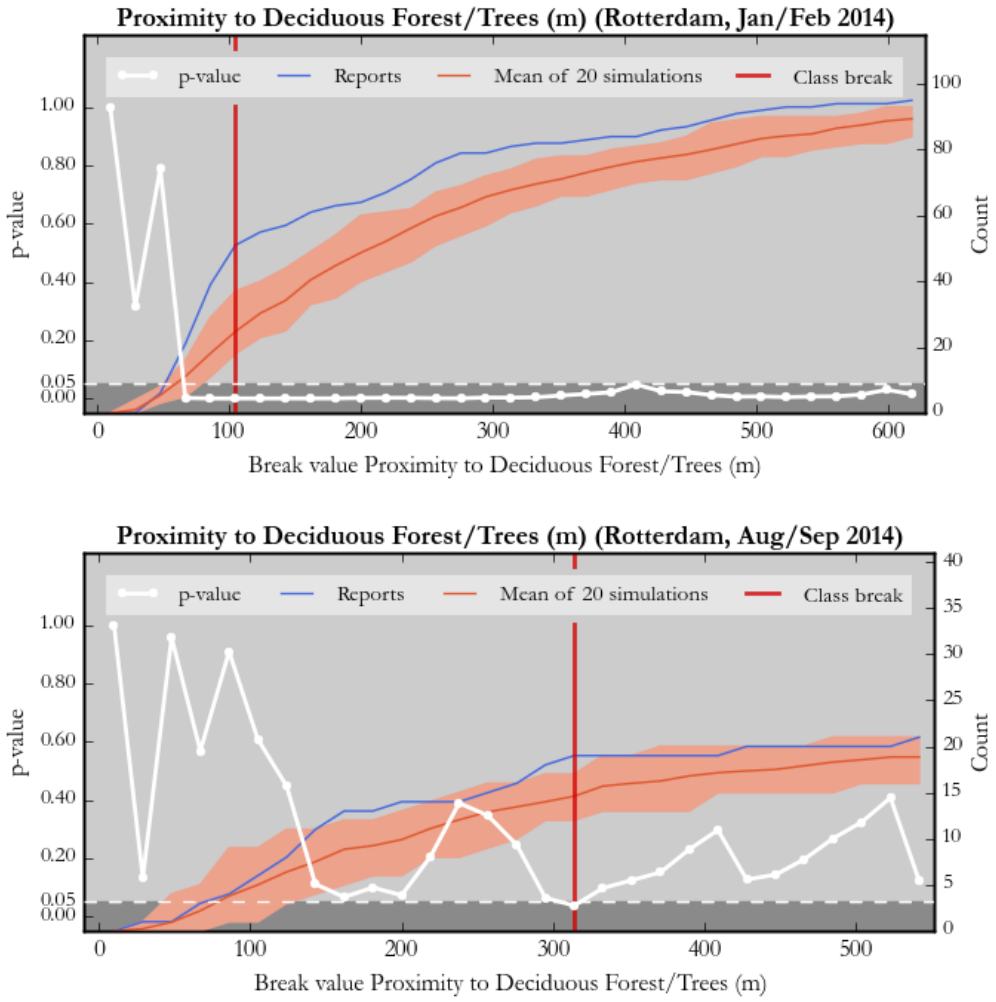


Figure 17: Line graphs showing the Chi Square test's p-value for all tested class breaks for proximity to deciduous forest/trees for Amsterdam in Jan/Feb (top), and Aug/Sep (bottom). The white line represents the p-value. At break values where the white line exceeds the dotted line at 0.05, the p-value is significant. The blue line represents the number of reports that are made at a distance from 0 metres to the distance of the break value. The red line represents the count of the mean of 20 simulations for each break value.

The Amsterdam results show a different pattern than expected: areas close to deciduous forest are less likely to have mosquito reports. The hypothesis can therefore not be accepted for the Amsterdam reports. In Rotterdam, however, the odds ratio for reports at a close proximity to deciduous forest is higher than 1 for both January and February, and August and September reports, meaning that more mosquito reports are expected at a small proximity to deciduous forest. The hypothesis can therefore be accepted based on the Rotterdam data.

Table 11: Overview table showing the strongest odds ratio and corresponding break value for each study area and Muggenradar round for the proximity to deciduous forest and trees

	Odds ratio	Break value
Amsterdam (Jan/Feb)	0.271	486
Amsterdam (Aug/Sep)	1.000	No significance
Rotterdam (Jan/Feb)	3.381	105
Rotterdam (Aug/Sep)	4.071	314

4.3.3 Building Construction Year

The output of the Chi Square test for building construction year can be seen in Table 12. Although the number of reports made from buildings older than 1976 is consistently larger than the expected reports in this class, the p-values are in all cases not significant. Therefore, it can be concluded that building construction years have no effect on the likelihood of reporting a mosquito to the *Muggenradar* project. The hypothesis will therefore be rejected.

Table 12: The results of the Chi Square test for building construction year. The first two columns show the number of reports coming from a building with a certain construction year, while the following two columns show the number of reports that is expected to be reported from these buildings, based on 20 simulations of random PC5 centroids. The last three columns show the Chi Square result, the corresponding significance (p-value) and the odds ratio.

	Obs. Buildings < 1976	**Obs. Buildings >= 1976**	**Exp. buildings <1976**	**Exp. buildings >= 1976**	**X^2**	**p**	**Odds ratio**
Amsterdam (Jan/Feb)	46	14	45	15	0.027	0.868	1.022
Amsterdam (Aug/Sep)	17	1	14	4	3.035	0.081	5.231
Rotterdam (Jan/Feb)	70	25	68	27	0.118	0.731	1.071
Rotterdam (Aug/Sep)	18	3	15	6	1.868	0.172	2.000

4.3.4 Population Density

The results for the iterative Chi Square tests for population density can be seen in Figure 18 (Amsterdam) and Figure 19 (Rotterdam). The January and February reports of Amsterdam show significance, mainly when using the lower population density values as break value. The break value with the strongest odds ratio is 4,509 inhabitants per km^2, with an odds ratio of 0.256. It is thus more likely to find mosquito reports at the more densely populated areas, which is in line with the formulated hypothesis.

The August and September reports of Amsterdam does only differ significantly from the random point simulations when placing the break value between 24,000 and 28,000 inhabitants per km^2. Using these break values, the amount of reports in less densely populated areas is lower than expected based on the random simulations. More reports are expected in the more densely populated areas. The strongest break value is 27,004, where the odds ratio is 0.125. It is more likely to find a mosquito report in the more densely populated areas.

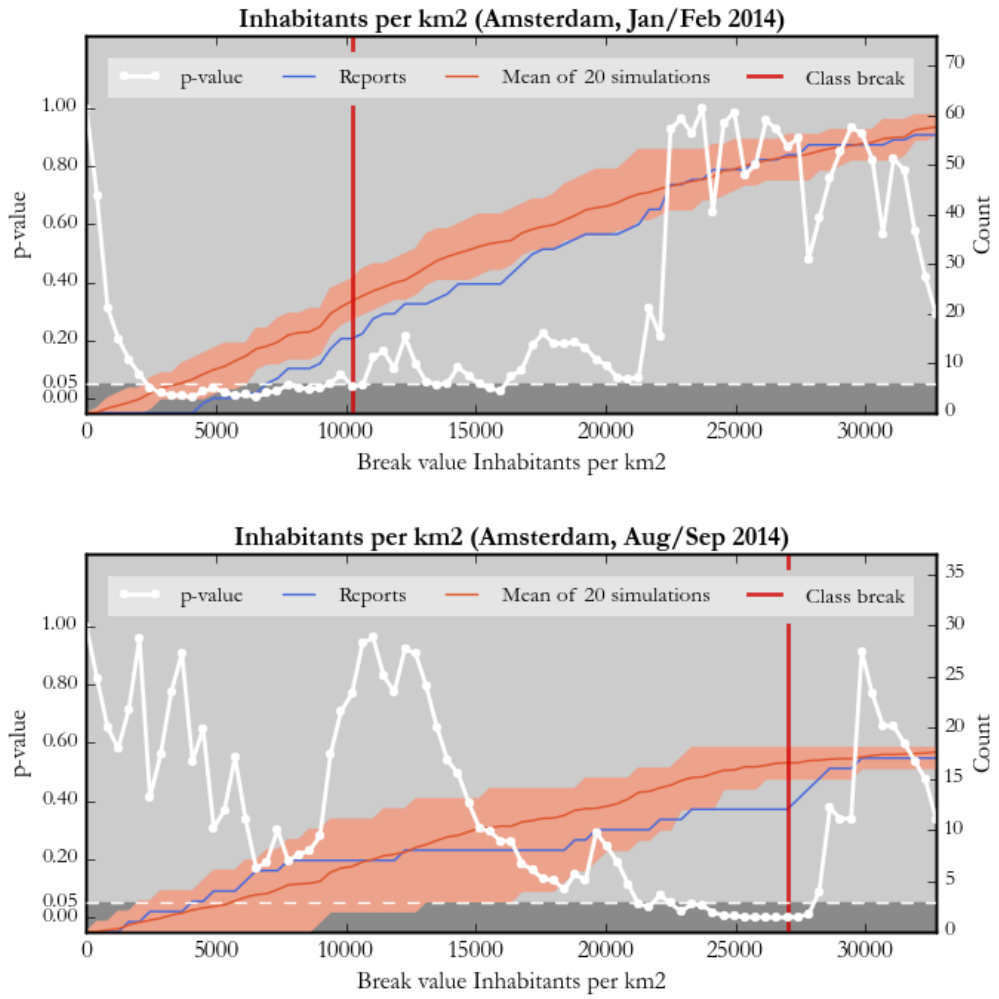


Figure 18: Line graphs showing the Chi Square test's p-value for all tested class breaks for inhabitants per km² for Amsterdam in Jan/Feb (top), and Aug/Sep (bottom). The white line represents the p-value. At break values where the white line exceeds the dotted line at 0.05, the p-value is significant. The blue line represents the number of reports that are made from areas with 0 inhabitants per km² to the break value of number of inhabitants per km². The red line represents the count of the mean of 20 simulations for each break value.

The January and February reports of Rotterdam show significance over almost the whole range of break values. The blue line exceeds the red line, meaning that the number of reports is higher than expected based on random simulations for almost all values. The strongest break value is 9,008 inhabitants per km². The odds ratio for reports made at less densely populated areas is 2.424, meaning that more reports are made at these less densely populated areas.

The August and September reports for Rotterdam show some significance at certain break values. However, one of the used criteria for a break value to be strong enough is that the number of reports in the class up to the break value should be outside the minimum and maximum boundaries of the random simulations (i.e. outside the bright red envelope shown in Figure 19). Since the p-value is never under the 0.05 where the report count in the first class is outside the outer boundaries of the simulations, no valid break point was found.

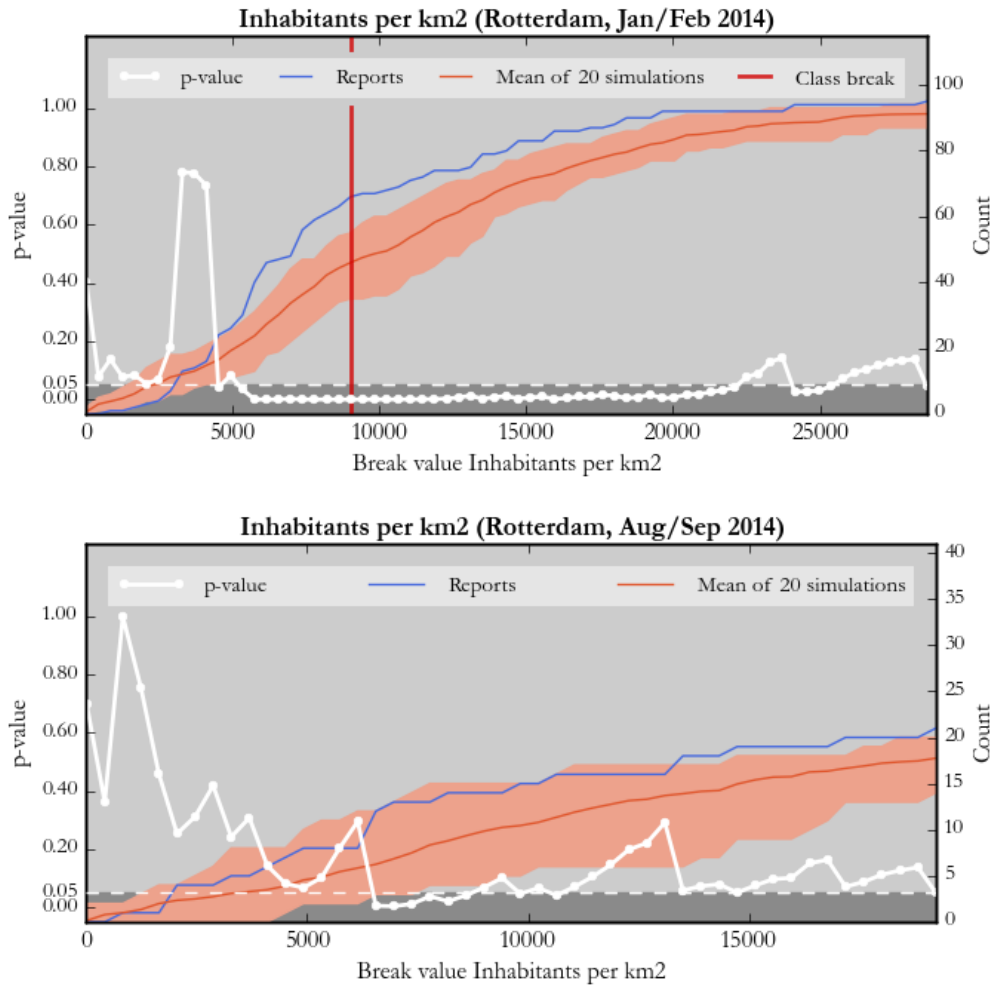


Figure 19: Line graphs showing the Chi Square test's p-value for all tested class breaks for inhabitants per km² for Rotterdam in Jan/Feb (top), and Aug/Sep (bottom). The white line represents the p-value. At break values where the white line exceeds the dotted line at 0.05, the p-value is significant. The blue line represents the number of reports that are made from areas with 0 inhabitants per km² to the break value of number of inhabitants per km². The red line represents the count of the mean of 20 simulations for each break value.

Three out of the four situations show clear significance. In Amsterdam, more reports are made at the more densely populated areas, which is in line with the hypothesis. In Rotterdam however, the opposite pattern is visible in the January and February data: more reports are made at the less densely populated areas, meaning that the hypothesis cannot be accepted for Rotterdam.

Table 13: Overview table showing the strongest odds ratio and corresponding break value for each study area and Muggenradar round for population density

	Odds ratio	Break value
Amsterdam (Jan/Feb)	0.561	10,235
Amsterdam (Aug/Sep)	0.125	27,004
Rotterdam (Jan/Feb)	2.424	9,008
Rotterdam (Aug/Sep)	1.000	No significance

4.4 Mosquito Presence Hotspots

4.4.1 Mosquito presence hotspot maps

The hotspot maps can be seen in Figure 20 to Figure 23. The colours in the hotspot map range from bright green to bright red, in which green represents an expected low mosquito presence (an odds ratio close to 0), and red represents an expected high mosquito presence (an odds ratio higher than 1). Bright yellow colours represent an odds ratio of 1. Table 14 shows the count of the number of reports in these low presence areas, and the count of reports in high presence areas.

The resulting hotspot maps for Amsterdam's January and February reports can be seen in Figure 20. As a result of the environmental factor analysis for the January and February reports of Amsterdam, densely populated areas at a large distance from water and deciduous forest have the highest chance of mosquito presence in this area. Only 21.7% of the mosquito reports are located in the expected high presence areas. The majority of the reports is thus located at raster cells where the odds ratio is lower than 1.

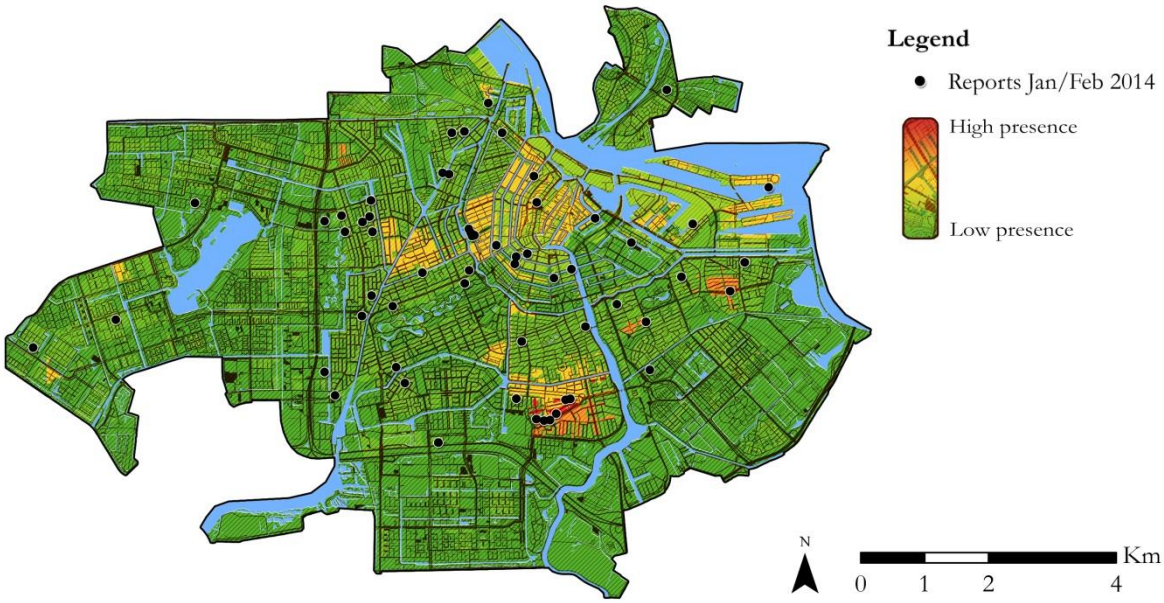


Figure 20: Hotspot map of Amsterdam, showing the locations where mosquito presence during the winter months is most likely. Muggenradar mosquito reports that were made in January and February 2014 are shown as black point symbols.

Figure 21 shows the hotspot map for Amsterdam, based on the August and September reports. Since population density was the only environmental factor that showed significance from random point simulations, the hotspot map is solely based on this factor. Only the strongly urbanised areas, where the population density is higher than 27,004 inhabitants per km^2, show a high chance of mosquito presence. Only 33.3% of all August and September reports are located in these highly urbanised areas.

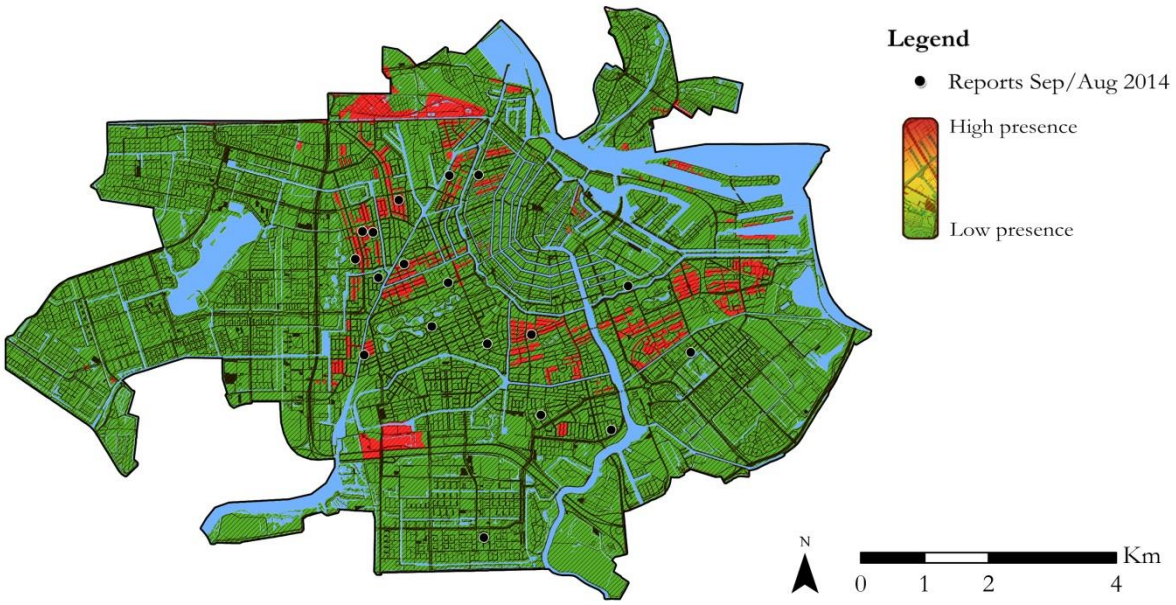


Figure 21: Hotspot map of Amsterdam showing the locations where mosquito presence during the summer months is most likely. Muggenradar mosquito reports that were made in August and September 2015 are shown as black point symbols.

Figure 22 show the hotspot map for Rotterdam, based on the winter reports of January and February 2014. The areas where the mosquito presence is highest are characterised by a short proximity to water and deciduous forest (respectively smaller than 112 and 105 metres), and a population density of less than 9,008 inhabitants per km^2. As can be seen in the hotspot map, the city centre of Rotterdam is resulting in an area where the mosquito presence is expected to be low. The areas with a high mosquito presence are the areas with a large proximity to water and deciduous forest, and a high population density. 66.3% of all reports made are located in areas where the expected mosquito presence is high, meaning a multiplied odds ratio that exceeds 1.

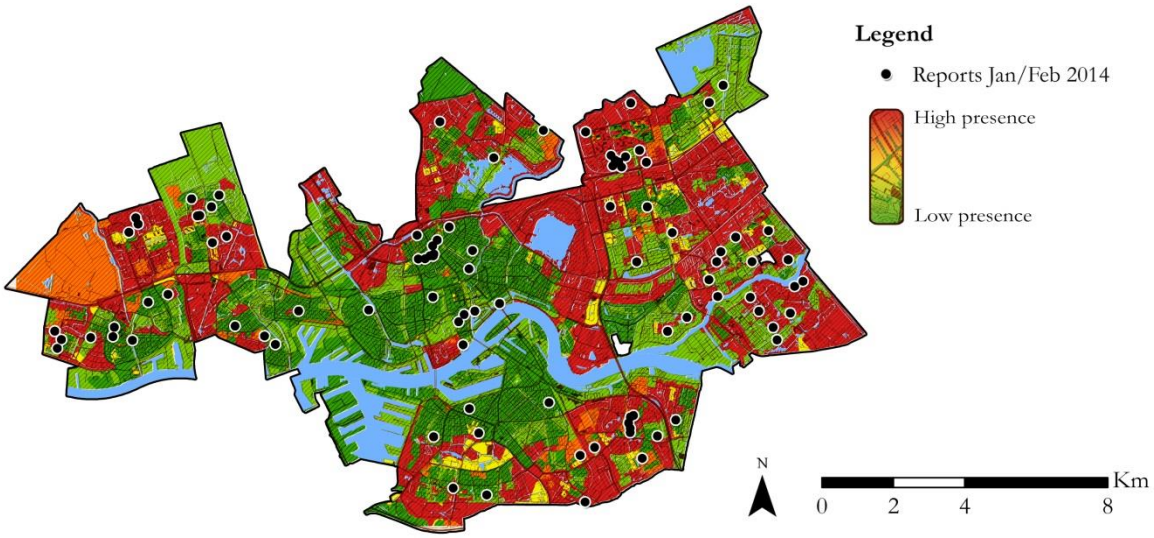


Figure 22: Hotspot map of Rotterdam, showing the locations where mosquito presence during the winter months is most likely. Muggenradar mosquito reports that were made in January and February 2014 are shown as black point symbols.

The hotspot map of Rotterdam's August and September reports can be seen in Figure 23. Only the environmental factor proximity to deciduous forest was found to be predictive. The other environmental factors did not show any significance, meaning that the August and September reports of Rotterdam do not differ enough from a random point pattern. The hotspot map is therefore only based on the proximity to deciduous forest: raster cells at a small proximity to deciduous forest (maximum 314 metres) have a higher chance of mosquito presence. Since only a small part of the study area is located at a larger proximity than 314 metres, a majority of the reports, 90.5%, is located within areas with a high chance of mosquito presence.

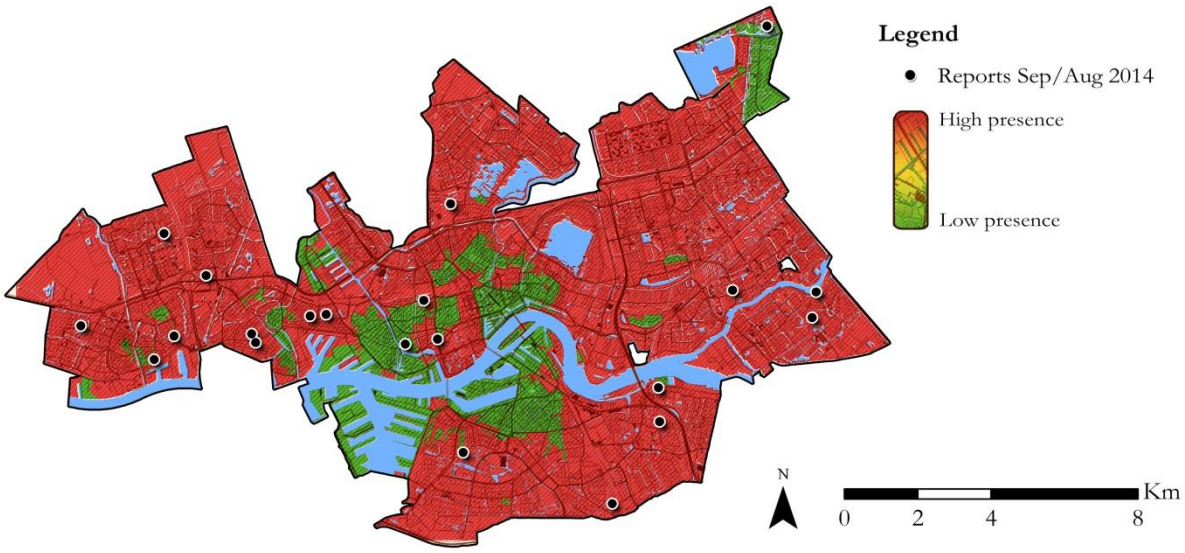


Figure 23: Hotspot map of Rotterdam, showing the locations where mosquito presence during the summer months is most likely. Muggenradar mosquito reports that were made in August and September 2014 are shown as black point symbols.

Table 14: The number of reports made in expected low presence areas (where OR < 1), expected high presence areas (OR > 1), and the relative count for Amsterdam and Rotterdam.

	OR < 1	OR > 1	OR < 1 (%)	OR > 1 (%)
Amsterdam (January/February)	47	13	78.3	21.7
Amsterdam (August/September)	12	6	66.7	33.3
Rotterdam (January/February)	32	63	33.7	66.3
Rotterdam (August/September)	2	19	9.5	90.5

Both of the hotspot maps for Amsterdam have a rather low accuracy. The Rotterdam hotspot map based on the January and February reports shows a small majority of reports within the expected high presence areas, and can therefore be seen as accurate. Although the Rotterdam hotspot map based on the August and September reports shows a majority of reports in high presence areas, this map cannot be seen as accurate. Only one environmental predictor (proximity to deciduous forest) was found to be significant in the environmental factor analysis, resulting in a hotspot map that is solely based on this factor.

4.4.2 Validation using mosquito reports Utrecht

The results of the validation of the odds ratios for January and February can be seen in Figure 24. Figure 24a shows the hotspot map for Utrecht using the odds ratios of Amsterdam. Table 15 shows the number of reports counted in respectively low and high presence areas.

Only 5.4% of all January and February reports is located within high presence areas, meaning that the result of the environmental factor analysis of Amsterdam in winter cannot be extrapolated to other areas.

Figure 24b shows the hotspot map for Utrecht using the odds ratios of Rotterdam. This map has a higher accuracy: 64.9% of the January and February reports of Utrecht are located within high presence areas.

Figure 24: Validation hotspot maps of Utrecht, based on the January and February parameters of Amsterdam (a) and Rotterdam (b), showing the locations where mosquito presence during the winter months is most likely based on the parameters of these cities. Muggenradar mosquito reports that were made in January and February 2014 are shown as black point symbols.

The hotspot maps for Utrecht, based on the August and September parameters of Amsterdam and Rotterdam can be seen in Figure 25. The corresponding counts can be seen in Table 15. From the results of the summer hotspot maps for Amsterdam and Rotterdam became already clear that those maps were not accurate enough. Since the reports for Amsterdam and Rotterdam did not differ enough from a random point pattern, the hotspot maps are only based on one environmental factor.

When using the same break values and odds ratios as the ones that were used for Amsterdam for the validation area Utrecht, large parts of the area show up in green (Figure 25a), meaning an expected low mosquito presence. This effect is similar to what can be seen in the hotspot map of Amsterdam, where a large part of Amsterdam shows up as a low presence area. 100% of the Utrecht reports are located within low presence areas, which can be explained by the fact that almost the whole area is classified as low presence area.

Using Rotterdam's odds ratios for the validation area Utrecht shows an opposite pattern: large parts of the area are classified as high presence areas (Figure 25b), based on this one significant environmental factor (proximity to deciduous forest). 94,7% of all August and September reports

of Utrecht are reported in these high presence areas, meaning that only one report is located in a low presence area.

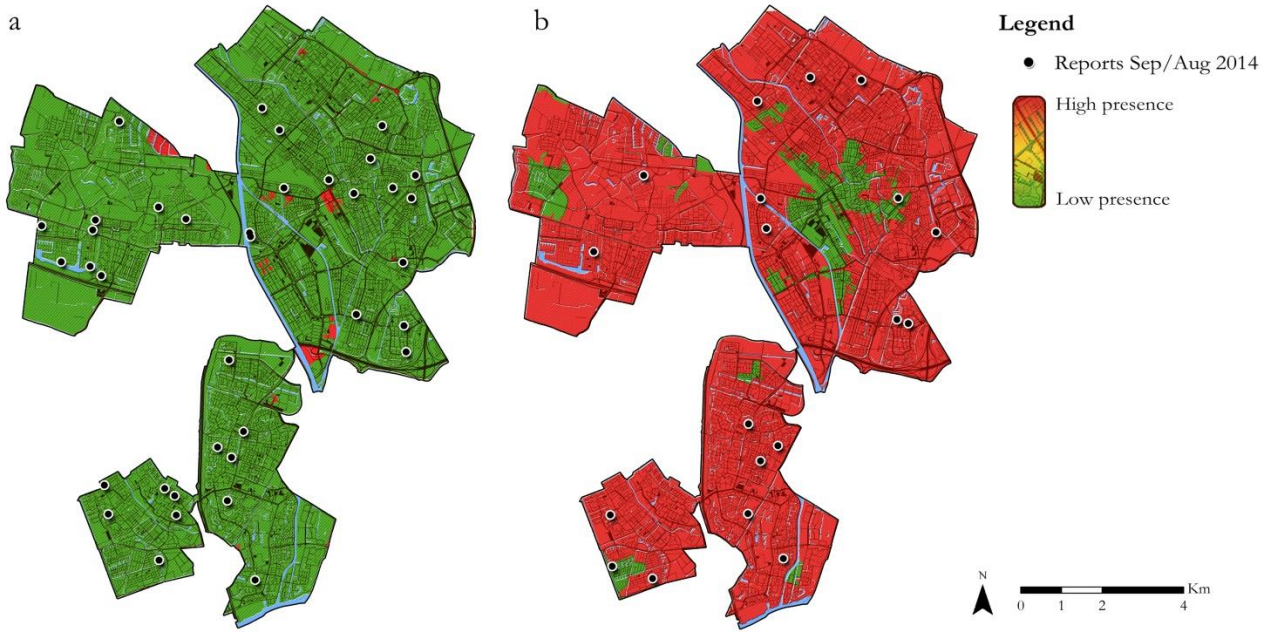


Figure 25: Validation hotspot maps of Utrecht, based on the August and September parameters of Amsterdam (a) and Rotterdam (b), showing the locations where mosquito presence during the summer months is most likely. Muggenradar mosquito reports that were made in August and September 2014 are shown as black point symbols.

Table 15: The number of reports made in expected low presence areas (where OR < 1), expected high presence areas (OR > 1), and the relative count for the validation area Utrecht, using the parameters of both Amsterdam and Rotterdam.

	OR < 1	**OR > 1**	**OR < 1 (%)**	**OR > 1 (%)**
Utrecht winter (Amsterdam parameters)	35	2	94.6	5.4
Utrecht winter (Rotterdam parameters)	13	24	35.1	64.9
Utrecht summer (Amsterdam parameters)	19	0	100.0	0.0
Utrecht summer (Rotterdam parameters)	1	18	5.3	94.7

5 Discussion & Conclusions

This study found that clustering of *Muggenradar* mosquito reports appeared in the more urbanised areas. A relationship was found between the number of reports made, and the urbanisation of the report locations. A spatial pattern analysis was done in two urban areas in the Netherlands, to find out which environmental factors are predictive for mosquito presence. Distance to water, distance to deciduous forest, and population density were found to be predictive for mosquito presence in Amsterdam and Rotterdam, based on the January and February data. The significance that was found, however, was not consistent, and not always in line with the formulated hypotheses. In Amsterdam for example, a large distance to water was related to more mosquito reports, whereas a small distance to water was related to more mosquito reports in Rotterdam, based on the January and February reports. A small proximity to water was expected to be predictive. Building construction years were not found to be significant. The August and September data showed less significance, or no significance for all environmental factors. Hotspot maps based on the significant environmental factors were in general rather inaccurate.

5.1 Discussion

Since this study made use of crowdsourced data, the used *Muggenradar* dataset was first explored. Clustering was found compared to a spatially random pattern, and occurred mainly in the more urbanised areas, which was in line with the hypothesis. Thereafter, this study analysed the spatial co-occurrence of the different genera.

Based on literature, one would expect that *Culex* mosquitoes appear spatially segregated from the other reported genera, because of their different habitat preferences (Becker et al., 2010). The results from the August and September reports made clear that this was indeed the case. Ulrich & Gotelli (2007), however, state that the results of co-occurrence analyses are strongly dependent on the index that is used. Presence-absence matrices that are disordered or high-filled are associated with a negative co-occurrence of genera, meaning segregation of the genera. The genus that is reported the most during August and September (*Culex*), has a significant negative co-occurrence with the other genera. This could be explained by the presence-absence matrices being highly filled with this mosquito genus, whereas the other genera are less present in the matrices. The results of the co-occurrence analysis of the January and February reports could possibly be explained in the same way: Both *Culex* and *Culiseta* mosquitoes are reported more often than *Anopheles* mosquitoes, resulting in a matrix highly filled with these two genera. The results show segregation between these genera and the third reported genus, *Anopheles* mosquitoes. It can be concluded that this method was not the right method to use with the *Muggenradar* dataset, in which the number of reports was rather low for *Aedes* and *Anopheles* mosquitoes. With a higher number of reports, especially for the *Aedes* and *Anopheles* mosquitoes, the outcome of the analysis could be more useful.

The reported absence in the *Muggenradar* dataset was found to be not useful. If mosquito presence is dependent on broad-scale environmental factors, one would expect to find absence reports at locations where no presence was reported. No correlation was found between the two types of reports, meaning that a random pattern is occurring. The number of absence reports in the *Muggenradar* dataset was rather low, which could have caused this result. Sequeira et al. (2014) mentions that citizen scientists are often rather hesitant in reporting absence, resulting in a strong bias towards more presence data.

Absence reports are highly important when studying the spatial patterns of mosquitoes. Although it is possible to derive environmental predictors from presence-only data, it is preferred to have both presence and absence data available (Brotons et al., 2004). The low number of absence reports in the *Muggenradar* dataset also obstructed a better validation of the hotspot maps:

validating the low-presence areas could be easily done with absence reports. When using presence-only data, the number of useful methods or statistical tests for an environmental factor analysis is also rather low. Having a sufficient amount of absence reports would open opportunities to use different methods or statistical tests, since both 0 (absence) and 1 (presence) data would then be available.

A demographic factor analysis tried to create an image of the respondents of the *Muggenradar* project. Although significant correlations were found, the correlations were not that strong. Having insight in the demographic groups that are active in reporting mosquito presence or nuisance could improve the communication strategy for possible future *Muggenradar* rounds, by triggering different demographic groups to contribute to the project. In an ideal situation, citizens from all demographic groups would be represented in a crowdsourced project.

The second part of this study focused on the spatial pattern analysis. Spatial patterns of mosquitoes have been studied before, but most of these studies used traditional methods such as CO_2 traps to collect mosquitoes (DeGroote et al., 2007; Deichmeister & Telang, 2011; Diuk-Wasser et al., 2006), whereas this study used crowdsourced data. After finding that *Muggenradar* mosquito reports were mainly located in urban areas, the study's focus shifted towards a more urban approach during this spatial pattern analysis. Contrary, most studies focused on a more broad study area that includes both urban and rural areas.

The environmental factors that were tested in other studies therefore differ from the environmental factors used in this study. Also, most studies were conducted in other countries, where other mosquito species are abundant, and other landscape characteristics could play a role in the distribution of mosquitoes. This study was the first to study the spatial patterns of mosquitoes in the Netherlands based on crowdsourced data, and therefore has an explorative character. It showed, however, that crowdsourced data can be of great value to study spatial patterns of mosquitoes, especially in urban areas. Having knowledge on the distribution of mosquitoes in urban areas is important, since a high human population density is found to be a risk factor for WNV in Europe (Tran et al., 2014).

Some results of the environmental factor analysis were in line with the hypotheses. The patterns were not consistent for all of the study areas and *Muggenradar* report rounds. The hypotheses for the environmental factor analysis could therefore in general not be accepted, based on the *Muggenradar* mosquito reports. The hypotheses were formulated based on the findings of comparable studies. Differences in the findings between this study and other studies can possibly be explained by the fact that this study focused on urban areas, whereas other studies focused on both the urban and rural areas.

Although the results of the environmental factor analysis were not always in line with the hypotheses, the retrieved break values and odds ratios were nonetheless used to generate hotspot maps. The fact that the spatial patterns were not as expected, does not mean that the reports do not have a pattern at all. The accuracy of these maps was generally not very high. In a number of maps, large areas showed up in red, meaning that high presence is expected. In most cases, this result is caused by low significance in the environmental factor analysis. The hotspot maps for the August and September data, for example, are solely based on one environmental factor, so one could argue over the value of these hotspot maps. When the only significance during the environmental factor analysis was found in the extremely high or low values, for example at an extremely small distance from water, or an extremely high distance from water, a majority of the hotspot map will show up as low or high presence. Large areas showing up as high presence areas can be seen as false-positives: the mosquito presence is expected to be high, whereas the actual mosquito presence is not that high. This overestimation of mosquito presence should be handled with care. The value of these hotspot maps is questionable. Absence data would be desired to

identify these false positives. Additionally, this absence data can be used to validate predicted low presence areas. Having a higher number of mosquito reports available for analysis could possibly result in more significance in the environmental factor analysis. This would eventually result in hotspot maps with a higher accuracy.

This study was not the first to focus on mosquito presence in urban areas. A study by Gleiser & Zalazar (2010) focused on the spatial patterns of mosquitoes in the city of Córdoba, Argentina. This study, however, did find that proximity to water and a higher NDVI are predictive for mosquito presence. The difference between the results of their study and this study could possibly be explained by the differences between the city of Córdoba, Argentina, and cities in the Netherlands. Another reason could lie within the fact that the study by Gleiser & Zalazar (2010) used CO_2 traps instead of crowdsourced mosquito collection. Mosquitoes caught indoors, by using crowdsourced mosquito collection, could be less dependent on the more broad-scale environmental factors such as proximity to water and deciduous forest, than mosquitoes caught outdoors by using CO_2 traps. Local factors could potentially be more predictive.

According to Becker et al. (2010), urban areas provide an ideal habitat for mosquitoes, with abundant availability of blood meals, and a wide range of water bodies. This study only took the larger, permanent water bodies that are included in the Kadaster Top10NL data into account. Local, more temporary water bodies such as flower cases, drink cans and plant pots could also form suitable breeding sites for mosquitoes (Becker et al., 2010). Medlock et al. (2012) names garden ponds as potential habitats for mosquitoes. All of these sites cannot be taken into account with a GIS-based analysis, since their appearance is highly temporary, or the sites are located at private properties where no data is available.

The fact that proximity to water was not found to be predictive for mosquito presence based on the *Muggenradar* reports is contradictory with the general perception of citizens. They tend to criticize the construction of water bodies in their neighbourhood, because these water bodies would attract mosquitoes and thus cause a higher mosquito nuisance in their neighbourhood.

Although the use of crowdsourcing as a method for mosquito collection is new, the concept of crowdsourcing or citizen science is used more often in ecology (Silvertown, 2009). The *Muggenradar* project fits well in this new trend of collecting data at a broad-scale and low costs. The accuracy, however, will never be as good as when using traditional collection methods (Gardiner et al., 2012).

Silvertown (2009) and Crall (2010) formulated a number of challenges to ensure the quality of citizen science projects. The *Muggenradar* project complies with most of these challenges. According to Crall et al. (2010), online data entry formats, such as the *Muggenradar* online questionnaire, are an important factor to limit the errors in the study. Using pen and paper would easily result in more mistakes when processing the reports in spreadsheets. When using online data entry formats, it is important to provide respondents an online instruction on how to recognize a mosquito, and on how to correctly send the mosquito to the Laboratory of Entomology. These instructions are available on the *Muggenradar* website. A large number of the reports that were made, however, was still identified as not being a mosquito. This sends out a signal that the provided instructions on the *Muggenradar* website are possibly not sufficient.

Another important factor to ensure the quality of citizen science projects is the use of expert validation (Crall et al., 2010), in order to filter the insects that are not correctly identified as mosquitoes by citizen scientists. An expert validation was carried out by the Laboratory of Entomology, by checking if an insect is a mosquito, and identifying all mosquitoes based on morphology up to the genus.

As a third challenge, Silvertown (2009) states that citizen scientists must receive feedback as a reward for their contribution to the project. This motivates citizen scientists to contribute to future projects. The *Muggenradar* respondents were all thanked for their contribution, and updated with information on the mosquito genus that they send via e-mail.

The fourth challenge argues that a citizen science project should be designed with a certain hypothesis or purpose in mind (Silvertown, 2009). The *Muggenradar* data was initially designed to monitor the mosquito presence and nuisance in winter. The project wanted to distinguish between the two species *Culex pipiens pipiens*, and *Culex pipiens molestus*. Therefore, mosquito samples had to be send to the Laboratory of Entomology for molecular tests. The collected data, however, could be used for a variety of purposes, including spatial pattern analyses.

Because the *Muggenradar* project complies with the quality standards for citizen science projects as formulated by Crall et al. (2010) and Silvertown (2009), it can be concluded that the quality of the data was probably not the problem. The quantity or number of reports was more of a limiting factor for this study. A number of recommendations can be done in order to increase the number of reports for a citizen science project. When designing a project specifically for a GIS-based study on the spatial patterns of mosquitoes in the Netherlands, the design of the project would have been different. Since a high number of reports is desired for a spatial pattern analysis, the focus of the project design would be on reducing the effort to report a mosquito.

Sending dead mosquitoes to the Laboratory of Entomology for identification and potential molecular tests would not be needed when using the data solely for a spatial pattern analysis. A good quality photograph of the mosquito would in most cases be sufficient to identify the mosquito up to the genus. Taking photographs would take less effort for respondents than sending a mosquito per post, which could eventually increase the total response of the project. Several other crowdsourcing projects already make use of an online questionnaire, combined with photographs to verify the reports and identify the insects (Table 16).

Another way to increase the number of reports, and to simplify the process of reporting, could be the use of a smartphone specific app. The UK Ladybird Survey and the Lost Ladybug Project already use such an app (Table 16). Citizens can then rapidly report mosquito nuisance via their smartphone. Additionally, the app could provide extra information on how to identify a mosquito correctly. Since most smartphones have a built-in camera and GPS function, respondents can easily send photographs of the mosquito, and include their exact geographic location in coordinates via the app. This would result in a report location with a higher accuracy than the PC6 level which is currently used in the *Muggenradar* project.

Additionally, when using photographs instead of sending a mosquito per post, the verification of reports will be less laborious. It will therefore become more feasible to increase the duration of the project. As can be seen in Table 16, many citizen science projects collect insects on a year-round basis, resulting in a high number of reports. the *Muggenradar* project made use of two specific data collection periods of respectively 5 and 2 weeks in January and February, and September and August. Increasing the duration of the project could increase the number of reports.

Another method that could possibly increase the number of reports of the *Muggenradar* project is the online publication of reports. The Firefly Watch (https://legacy.mos.org/fireflywatch/), for example, provides all its reports on an up-to-date online map, where citizens can view and explore the reports. A low number of reports in a citizen's neighbourhood could encourage one to contribute to the *Muggenradar* project.

Table 16: Examples of citizen science projects and their characteristics

Project	Type of citizen science	Description	Purpose	Sample size
UK Ladybird Survey (http://www.ladybird-survey.org/)	Verified	- Online questionnaire & smartphone app - Photo required	Record all UK ladybird species	100,000+ records since 2005
Lost Ladybug Project, North America (http://www.lostladybug.org/)	Verified	- Online questionnaire & smartphone app - Photo required	Monitor ladybird diversity of native and introduced species in North America	10,000 records since 2008
Moths Count, UK (http://www.mothscount.org/)	Verified	- Online questionnaire	Monitor moth species in order to avoid extinction of (more) moth species	16 million records since 2007
Firefly Watch, US (https://legacy.mos.org/firefly watch/)	Direct	- Online questionnaire	Learn about distribution of fireflies, and find the relation with human-made light	5,000+ records between 2008 and 2013

Since most *Muggenradar* mosquito reports are made indoors, broad-scale environmental factors as proximity to deciduous forest and proximity to water might be less relevant than in other spatial pattern studies. Simultaneously, the presence of local water bodies could be more predictive for indoor mosquito presence. A spatial pattern study based on a higher number of mosquito reports, preferably both presence and absence reports, is needed to make legitimate conclusions on the relationship between broad-scale environmental factors and indoor mosquito reports.

5.2 Main conclusions

Crowdsourcing is a great method that enables researchers to study the spatial patterns of mosquitoes on a broad geographic scale. Tran et al. (2014) found that human population density is a risk factor for WNV in Europe: areas with a higher human population density have a higher risk on WNV outbreaks. Even though the direct risk in the Netherlands is still low, having good knowledge of the spatial patterns of mosquitoes in urban areas is desired. By showing that more mosquito reports are made from the more urbanised areas, this study shows that crowdsourcing is an excellent tool to study the urban distribution of mosquitoes. Consistent relationships were not found between the proximity to water, proximity to deciduous forest, building construction year, and population density, and the *Muggenradar* mosquito reports. The objective of this study was to find mosquito presence hotspots in both summer and winter, based on crowdsourced data. The generated hotspot maps showed a rather low accuracy. More (environmental) factors, and possibly also more local factors, could be influencing the spatial patterns of mosquito presence in urban areas. It can be concluded that crowdsourced mosquito collection could be of great value for spatial pattern studies in urban areas, however, a higher number of mosquito reports is desired.

References

Baddeley A., Turner R. (2005). *spatstat: An R Package for Analyzing Spatial Point Patterns*. Journal of Statistical Software 12(6), 1-42. http://www.jstatsoft.org/v12/i06/

Becker N., Petric D., Zgomba M., Boase C., Madon M., Dahl C., Kaiser A. (2010). *Mosquitoes and their control (Vol. 2)*. Heidelberg: Springer.

Bithell J.F. (1990). *An application of density estimation to geographical epidemiology*. Statistics in Medicine, 9 (6), pp. 691-701.

Brotons, L., Thuiller, W., Araújo, M.B., Hirzel, A.H. (2004). *Presence-absence versus presence-only modelling methods for predicting bird habitat suitability*. Ecography, 27 (4), pp. 437-448.

Crall A.W., Newman G.J., Jarnevich C.S., Stohlgren T.J., Waller D.M., Graham J. (2010) *Improving and integrating data on invasive species collected by citizen scientists*. Biological Invasions, 12 (10), pp. 3419-3428.

Cui J., Li S., Zhao P., Zou F. (2013) *Flight capacity of adult Culex pipiens pallens (Diptera: Culicidae) in relation to gender and day-age*. Journal of Medical Entomology, 50 (5), pp. 1055-1058.

DeGroote J., Mercer D. R., Fisher H., Sugumaran R. (2007). *Spatiotemporal Investigation of Adult Mosquito (Diptera: Culicidae) Populations in an Eastern Iowa County, USA*. Journal of Medical Entomology 44(6):1139-1150.

Deichmeister J. M., Telang A. (2011). *Abundance of West Nile virus mosquito vectors in relation to climate and landscape variables*. J Vector Ecol. 2011 Jun;36(1):75-85. doi: 10.1111/j.1948-7134.2011.00143.x.

Dickinson J.L., Zuckerberg B., Bonter D.N. (2010). *Citizen science as an ecological research tool: Challenges and benefits*. Annual Review of Ecology, Evolution, and Systematics, 41, pp. 149-172.

Diuk-Wasser M.A., Brown H.E., Andreadis T.G., Fish D. (2006). *Modeling the spatial distribution of mosquito vectors for West Nile virus in Connecticut, USA*. Vector Borne Zoonotic Dis. 2006 Fall;6(3):283-95.

Gardiner M.M., Allee L.L., Brown P.M.J., Losey J.E., Roy H.E., Smyth R.R. (2012). *Lessons from lady beetles: Accuracy of monitoring data from US and UK citizen science programs*. Frontiers in Ecology and the Environment, 10 (9), pp. 471-476.

GDAL (2014) *GDAL - Geospatial Data Abstraction Library: Version 1.11.1*, Open Source Geospatial Foundation, http://gdal.osgeo.org

Gleiser R. M., Zalazar L. P. (2010). *Distribution of mosquitoes in relation to urban landscape characteristics*. Bulletin of Entomological Research, 100, pp 153-158. doi:10.1017/S0007485309006919.

Greenberg J. A., DiMenna M. A., Hanelt B., Hofkin B. V. (2012). *Analysis of post-blood meal flight distances in mosquitoes utilizing zoo animal blood meals*. Journal of Vector Ecology, 37(1), 83–89. doi:10.1111/j.1948-7134.2012.00203.x

Griffith D., Veech J., Charles M. (2014). *cooccur: Probabilistic Species Co-occurrence Analysis in R*

Haase P. (1995). *Spatial pattern analysis in ecology based on Ripley's K-function: Introduction and methods of edge correction*. Journal of Vegetation Science, 6 (4), pp. 575-582.

Hayes E.B., Komar N., Nasci R.S., Montgomery S.P., O'Leary D.R., Campbell G.L. (2005). *Epidemiology and transmission dynamics of West Nile virus disease*. Emerging Infectious Diseases, 11 (8), pp. 1167-1173

Hiwat H., Andriessen R., Rijk M. de, Koenraadt C.J.M., Takken W. (2011). *Carbon dioxide baited trap catches do not correlate with human landing collections of Anopheles aquasalis in Suriname*. Mem Inst Oswaldo Cruz, Rio de Janeiro, Vol. 106(3): 360-364

Hongoh V., Berrang-Ford L., Scott M. E., Lindsay L. R. (2012). *Expanding geographical distribution of the mosquito, Culex pipiens, in Canada under climate change*. Applied Geography, Volume 33, April 2012, Pages 53-62, ISSN 0143-6228, http://dx.doi.org/10.1016/j.apgeog.2011.05.015.

Howard J.J., White D.J., Muller S.L. (1989). *Mark-recapture studies on the Culiseta (Diptera: Culicidae) vectors of eastern equine encephalitis virus*. Journal of Medical Entomology, 26 (3), pp. 190-199.

Hunter J. D. (2007). *Matplotlib: A 2D Graphics Environment*, Computing in Science & Engineering, 9, pp. 90-95, DOI:10.1109/MCSE.2007.55

Krüger A., Rech A., Su X.-Z., Tannich E. (2001). *Short communication: Two cases of autochthonous Plasmodium falciparum malaria in Germany with evidence for local transmission by indigenous Anopheles plumbeus*. Tropical Medicine and International Health, 6 (12), pp. 983-985.

Levine N. (2005). *Chapter 3: Entering data into CrimeStat. In CrimeStat III: A spatial statistics program for the analysis of crime incident locations* (pp. 3.1–3.45). Houston, TX: Ned Levine & Associates. Retrieved at 23-9-2014 from http://www.icpsr.umich.edu/CrimeStat/files/CrimeStatChapter.3.pdf

McKinney W. (2010). *Data Structures for Statistical Computing in Python.* Proceedings of the 9th Python in Science Conference, 51-56

Medlock J. M., Hansford K. M., Anderson M., Mayho R., Snow K. R. (2012). *Mosquito nuisance and control in the UK–A questionnaire-based survey of local authorities*. Eur. Mosq. Bull, 30, pp. 15-29.

Miller-Rushing A., Primack R., Bonney R. (2012). *The history of public participation in ecological research*. Frontiers in Ecology and the Environment, 10 (6), pp. 285-290.

Nasci R.S., Savage H.M., White D.J., Miller J.R., Cropp B.C., Godsey M.S., Kerst A.J., Bennett P., Gottfried K., Lanciotti R.S. (2001). *West Nile virus in overwintering culex mosquitoes, New York City, 2000*. Emerging Infectious Diseases, 7 (4), pp. 742-744.

Oliphant T. E. (2007). *Python for Scientific Computing*, Computing in Science & Engineering, vol.9, no. 3, pp. 10-20, May/June 2007, doi:10.1109/MCSE.2007.58

Prick J.J., Kuipers S., Kuipers H.D., Vliegen J.H., van Doornum G.J. (2003). *Another case of West Nile fever in the Netherlands: a man with encephalitis following a trip to Canada*. Nederlands Tijdschrift Geneeskd. May 17;147(20):978-80.

Reisen W.K., Fang Y., Lothrop H.D., Martinez V.M., Wilson J., O'Connor P., Carney R., Cahoon-Young B., Shafii M., Brault A.C. (2006). *Overwintering of West Nile virus in Southern California.* Journal of Medical Entomology, 43 (2), pp. 344-355.

Reusken C.B.E.M., Vries A. de, Buijs J., Braks M.A.H., Hartog W. den, Scholte E.J. (2010). *First Evidence for Presence of Culex pipiens Biotype Molestus in the Netherlands, and of Hybrid Biotype Pipiens and Molestus in Northern Europe*. Journal of Vector Ecology, 35(1):210-212. 2010.

Reusken C.B.E.M., Vries A. de, Hartog W. den, Braks M.A.H., Scholte E.J. (2011). *A study of the circulation of West Nile virus in mosquitoes in a potential high-risk area for arbovirus circulation in The Netherlands, "De Oostvaardersplassen"*. Eur. Mosq. Bull. 28: 69–83.

R Core Team (2013). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. URL http://www.R-project.org/.

Reeves W. C., Brookman B., Hammon W. McD. (1948). *Studies on the flight range of certain Culex mosquitoes, using a fluorescent-dye marker, with notes on Culiseta and Anopheles*. Mosquito News (8):61-69

Schaffner F., Thiéry I., Kaufmann C., Zettor A., Lengeler C., Mathis A., Bourgouin C. (2012). *Anopheles plumbeus (Diptera: Culicidae) in Europe: A mere nuisance mosquito or potential malaria vector?* Malaria Journal, 11, art. no. 393

Scholte E., Den Hartog W., Dik M., Schoelitsz B., Brooks M., Schaffner F., Foussadier R., Braks M., Beeuwkes J. (2010). *Introduction and control of three invasive mosquito species in the Netherlands, July-October 2010.* Euro surveillance : bulletin européen sur les maladies transmissibles = European communicable disease bulletin, 15 (45)

Sequeira A.M.M., Roetman P.E.J., Daniels C.B., Baker A.K., Bradshaw C.J.A. (2014). *Distribution models for koalas in South Australia using citizen science-collected data.* Ecology and Evolution, 4 (11), pp. 2103-2114.

Service M.W. (1997). *Mosquito (Diptera: Culicidae) Dispersal - The Long and Short of It.* Journal of Medical Entomology, 34 (6), pp. 579-588.

Silvertown J. (2009). *A new dawn for citizen science.* Trends in Ecology and Evolution, 24 (9), pp. 467-471.

Takumi K., Scholte E.-J., Braks M., Reusken C., Avenell D., Medlock J.M. (2009). *Introduction, scenarios for establishment and seasonal activity of aedes albopictus in the Netherlands.* Vector-Borne and Zoonotic Diseases, 9 (2), pp. 191-196.

Tran A., Sudre B., Paz S., Rossi M., Desbrosse A., Chevalier V., Semenza J.C. (2014). *Environmental predictors of West Nile fever risk in Europe*. International Journal of Health Geographics, 13, art. no. 26

Tsuda Y., Komagata O., Kasai S., Hayashi T., Nihei N., Saito K., Mizutani M., Kunida M., Yoshida M., Kobayashi M. (2008). *A mark-release-recapture study on dispersal and flight distance of Culex pipiens pallens in an urban area of Japan*. J Am Mosq Control Assoc. 2008 Sep;24(3):339-43.

Ulrich W., Gotelli N.J. (2007). *Disentangling community patterns of nestedness and species co-occurrence.* Oikos, 116 (12), pp. 2053-2061.

Valiakos G., Papaspyropoulos K., Giannakopoulos A., Birtsas P., Tsiodras S., Hutchings M.R., Spyrou V., Pervanidou D., Athanasiou L.V., Papadopoulos N., Tsokana C., Baka A., Manolakou K., Chatzopoulos D., Artois M., Yon L., Hannant D., Petrovska L., Hadjichristodoulou C., Billinis C. (2014). *Use of wild bird surveillance, human case data and GIS spatial analysis for predicting spatial distributions of West Nile virus in Greece*. PLoS ONE, 9 (5), art. no. e96935

Veech J.A. (2013). *A probabilistic model for analysing species co-occurrence.* Global Ecology and Biogeography, 22 (2), pp. 252-260

Appendix 1 Table of Contents of the accompanied DVD

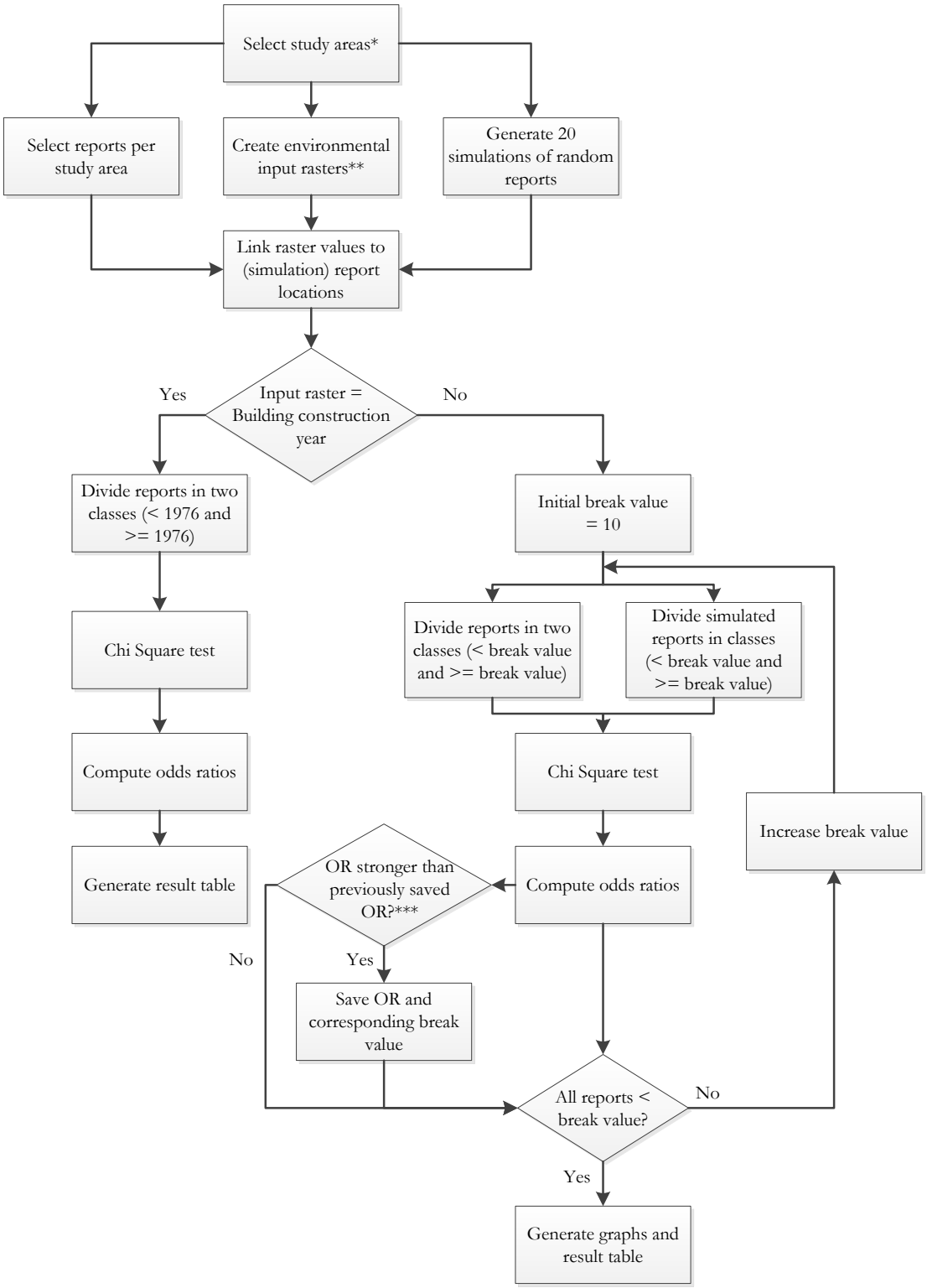
1. Report (as Microsoft Word and PDF document)
2. Presentations (Microsoft PowerPoint)
    a. Proposal presentation at Vector meeting
    b. Midterm presentation
    c. Thesis presentation at Vector meeting
    d. Colloquium
3. Datasets
    a. *Muggenradar* dataset January/February 2014
    b. *Muggenradar* dataset August/September 2014
    c. Hotspot maps
        i. Amsterdam (Jan/Feb 2014)
        ii. Rotterdam (Jan/Feb 2014)
        iii. Utrecht with Amsterdam parameters (Jan/Feb 2014)
        iv. Utrecht with Rotterdam parameters (Jan/Feb 2014)
        v. Amsterdam (Aug/Sep 2014)
        vi. Rotterdam (Aug/Sep 2014)
        vii. Utrecht with Amsterdam parameters (Aug/Sep 2014)
        viii. Utrecht with Rotterdam parameters (Aug/Sep 2014)
4. Figures, maps & tables
5. Python scripts
6. Literature (PDF files)
7. Notes of the supervisor meetings

Appendix 2 Raw *Muggenradar* Data Attributes

Attributes that are used for further analysis are shown in black, whereas attributes that were not necessary for the spatial analysis are shown in grey.

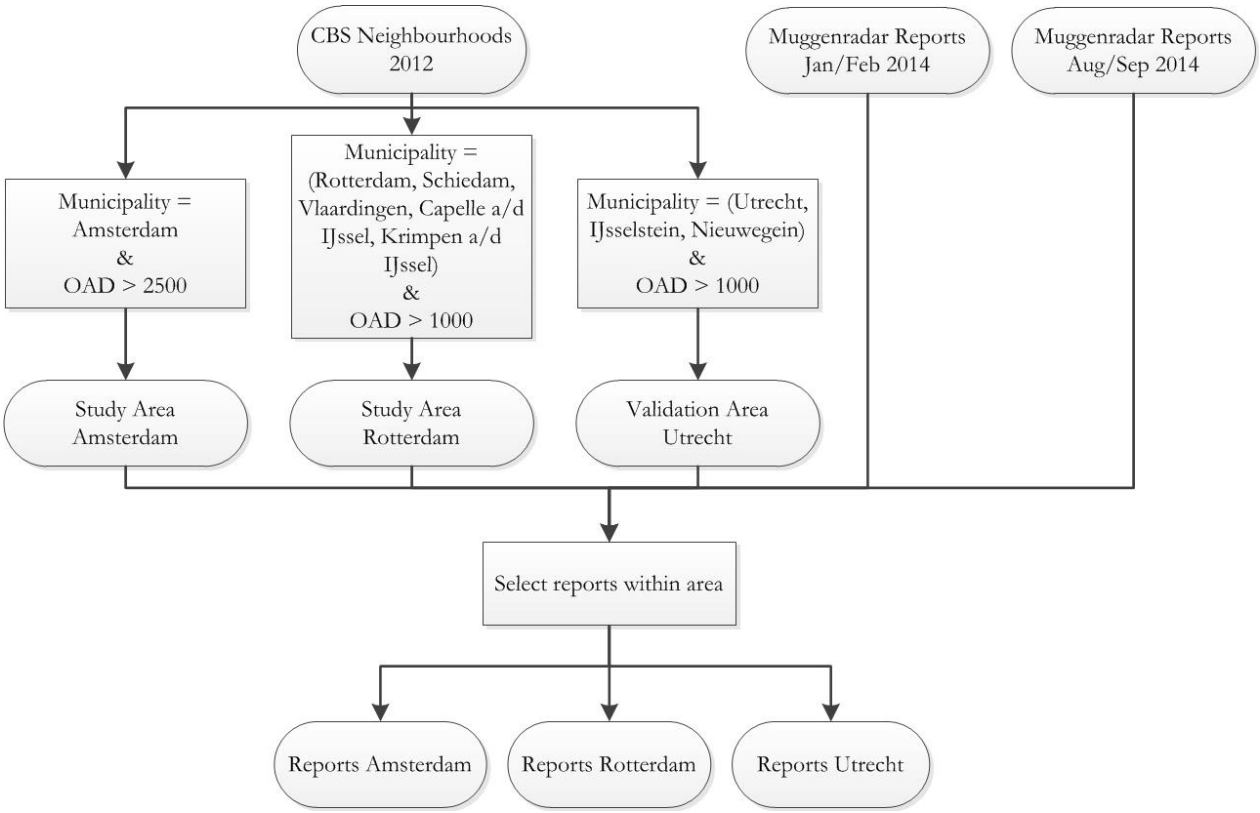
Attribute	Description	Values
Lab sample number	Reference number used in lab	Number between 1 and 3305
Unique code	6-digit code generated after filling in the online questionnaire	Random number of 6 digits
Postal code (PC6)	Postal code (PC6) where the mosquito is caught	
Postal code (PC5)	First five characters of the postal code	
Place of residence	Place of residence where the mosquito is caught	
Municipality	Municipality where the mosquito is caught	
Province	Province where the mosquito is caught	
E-mail address	E-mail address of the respondent	
Mosquito	Number that indicates if the insect is a mosquito	0 = No *Culicidae* 1 = Culicidae 2 = Not identifiable 3 = Empty envelope
Number of mosquitoes	Number of mosquitoes that were sent	
Genus	Genus of the mosquito	*Aedes* *Anopheles* *Culex* *Culiseta*
Species	Species as determined by a molecular test	*Culex pipiens pipiens* *Culex pipiens molestus* Hybrid
Quality	Quality of the insect(s)	0 = poor quality (mosquito is highly damaged) 1 = good quality
Blood-fed	Number indicating if the mosquito is blood-fed	0 = Not blood-fed 1 = Blood-fed
Reported nuisance	Nuisance reported by the respondent	Yes = Nuisance No = No nuisance
Extra unique codes	Extra unique codes obtained when the questionnaire is submitted more than once	Random number(s) of 6 digits
Remarks (Post)	Remarks sent with the sample	
Remarks (Questionnaire)	Remarks made in the online questionnaire	

Appendix 3 Flowchart of Research Question 3

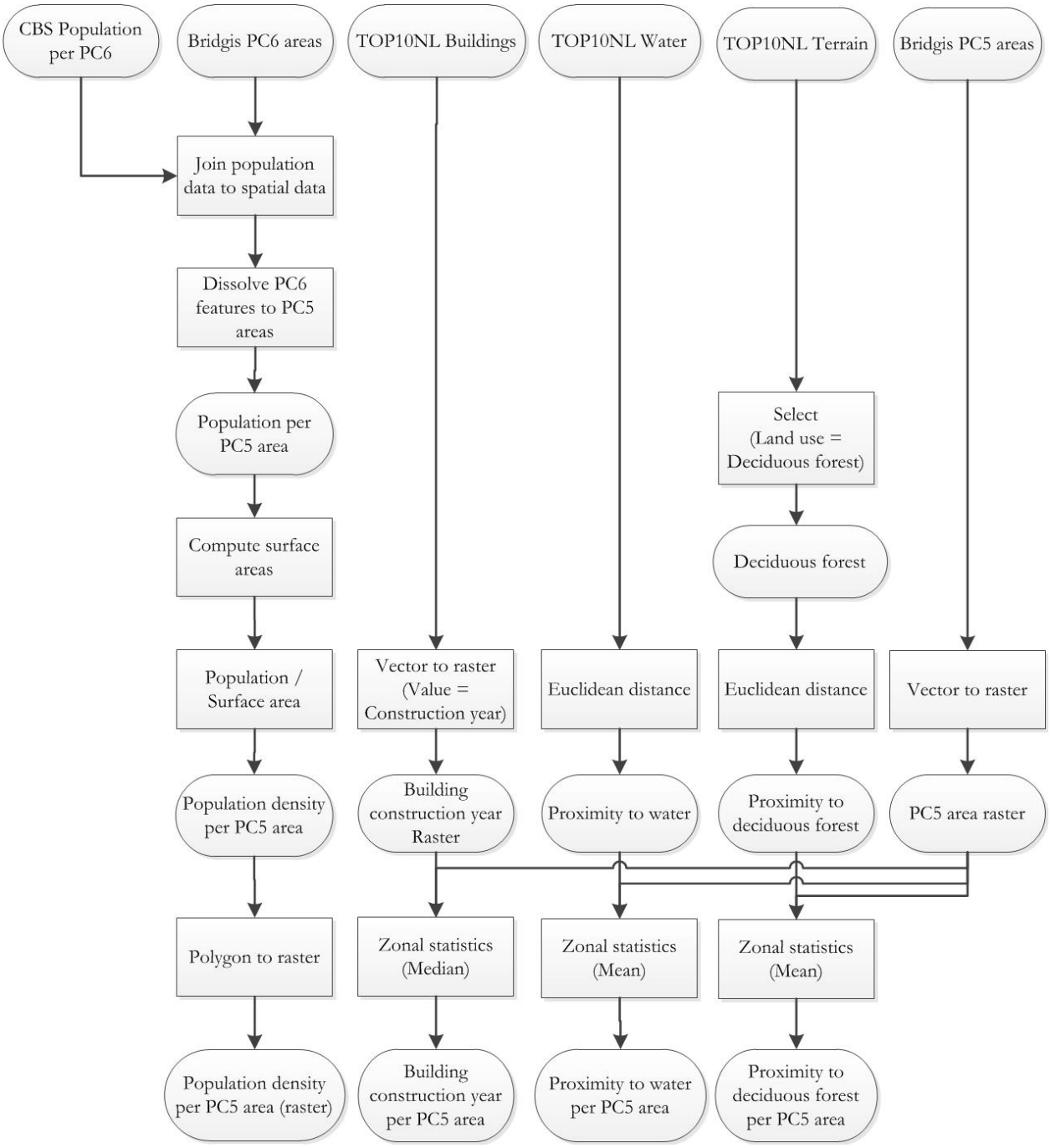


* A detailed flowchart of the study area selection can be found in Appendix 4.
** A detailed flowchart of the creation of input rasters can be found in Appendix 5.
*** The OR should be lower (in case of OR between 0-1) or higher (in case of OR of > 1) than the previously saved OR. The corresponding p-value should be below 0.05 (significant), and the report count in the first class should be outside the boundaries on the minimum and maximum simulation count. A break value in the middle 75% of the total number of reports is preferred, to avoid a small class count in one of the two classes.

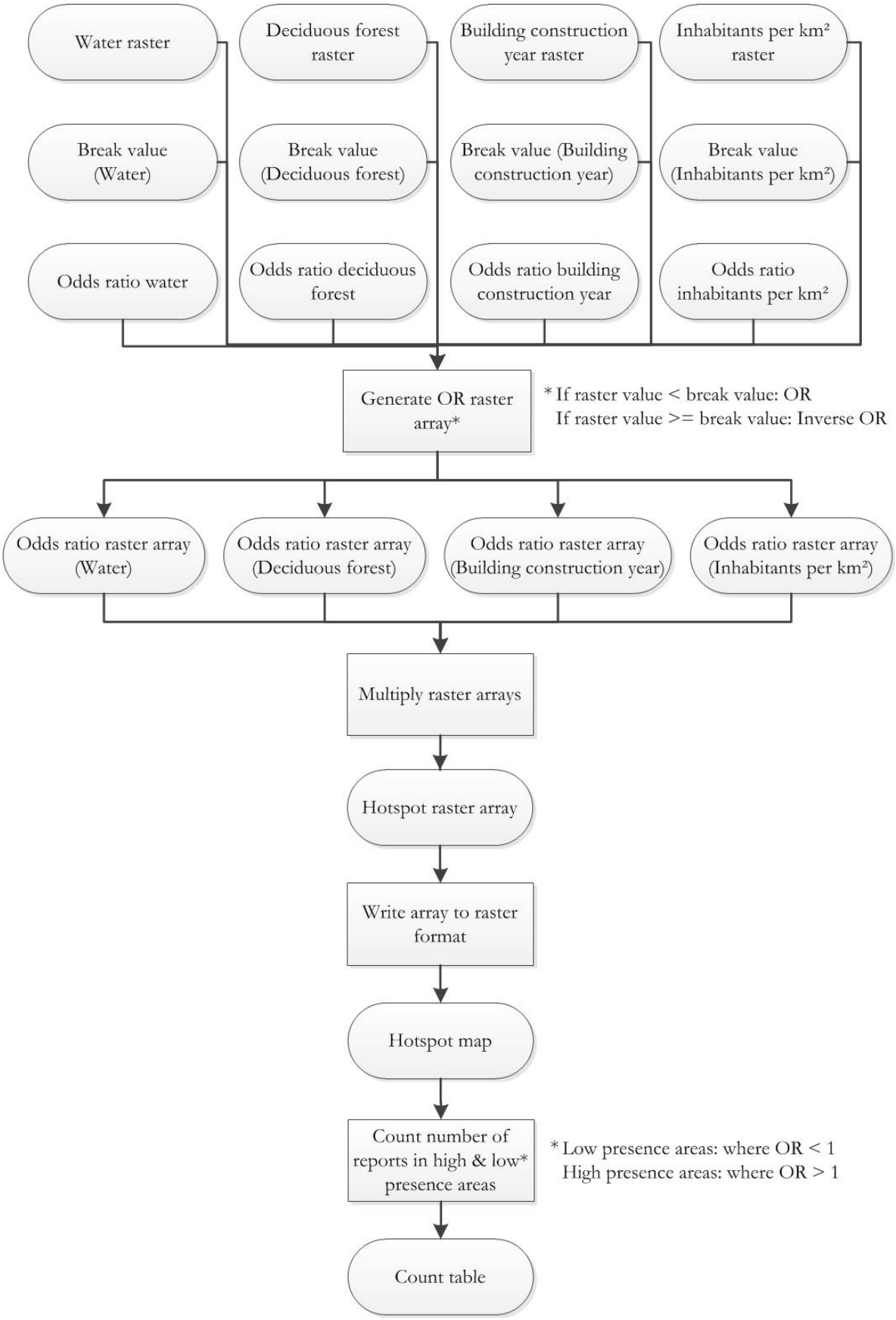
Appendix 4 Flowchart of the Study Area Selection



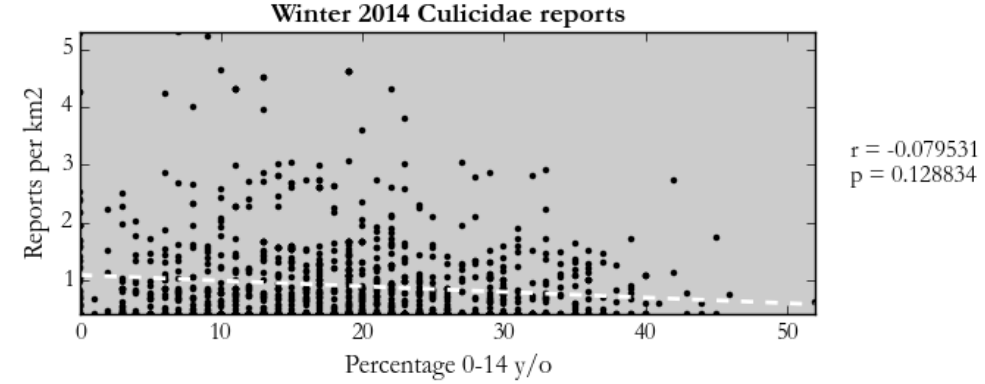
Appendix 5 Flowchart of the Pre-processing of Environmental Input Rasters

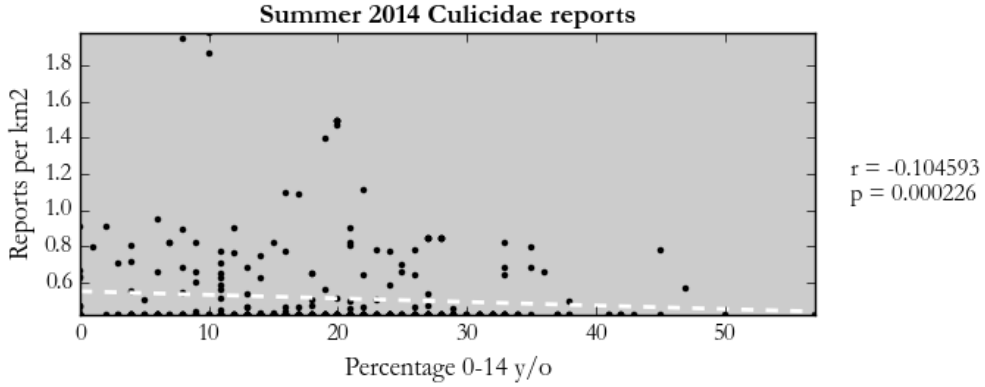


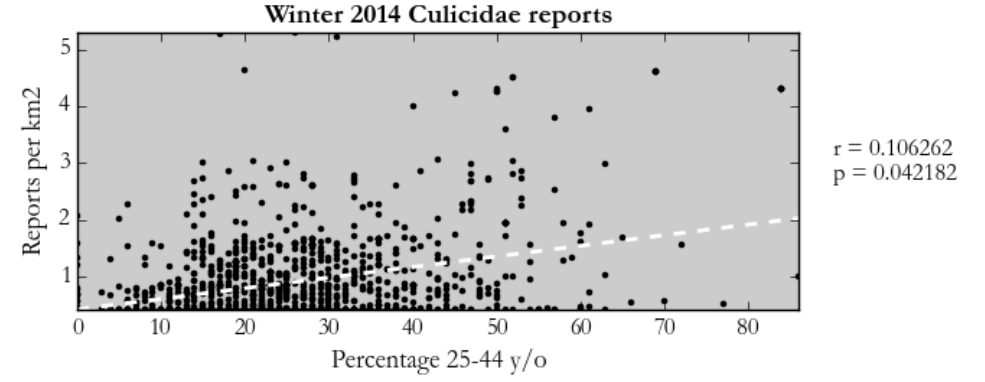
Appendix 6 Flowchart of Research Question 4

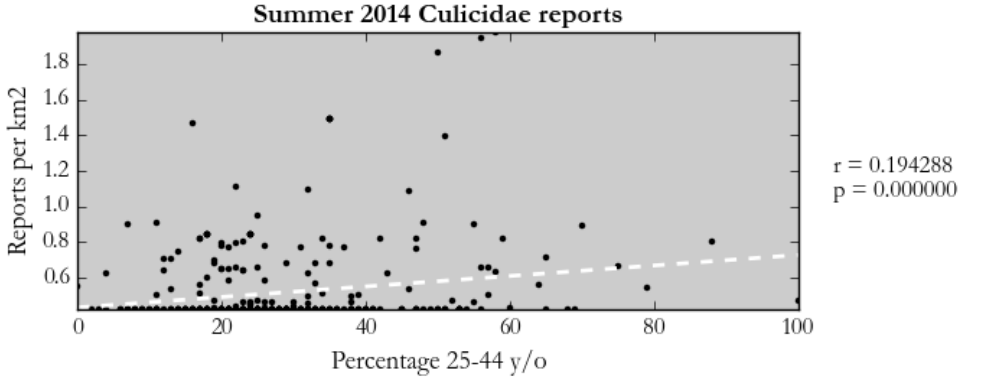


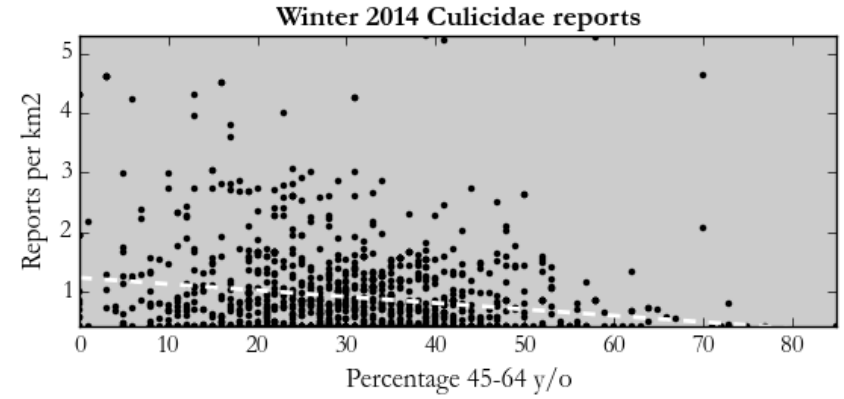
Appendix 7 Scatter Plots of (Significant) Demographic Characteristics

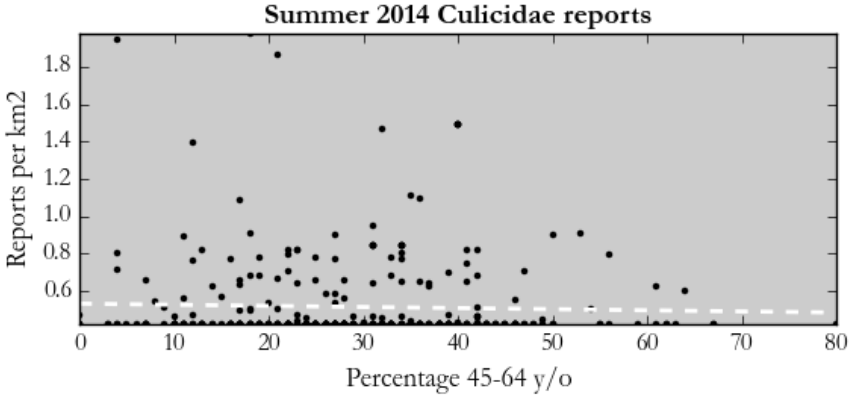


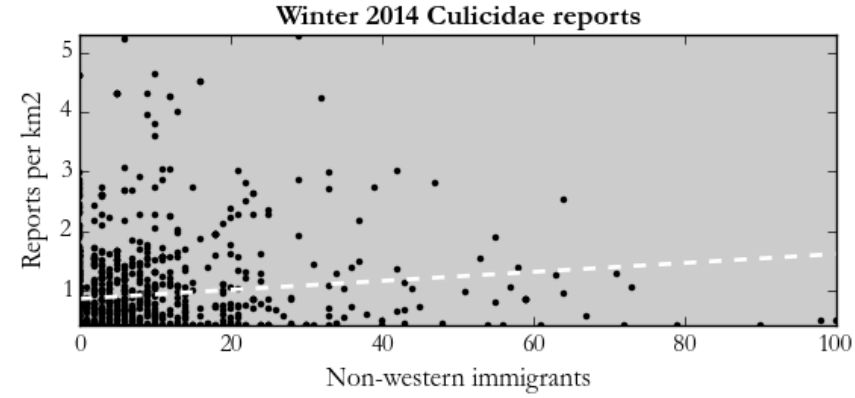


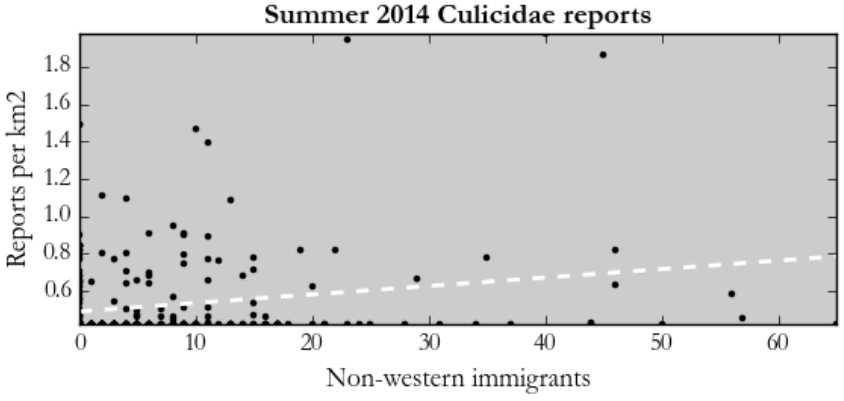


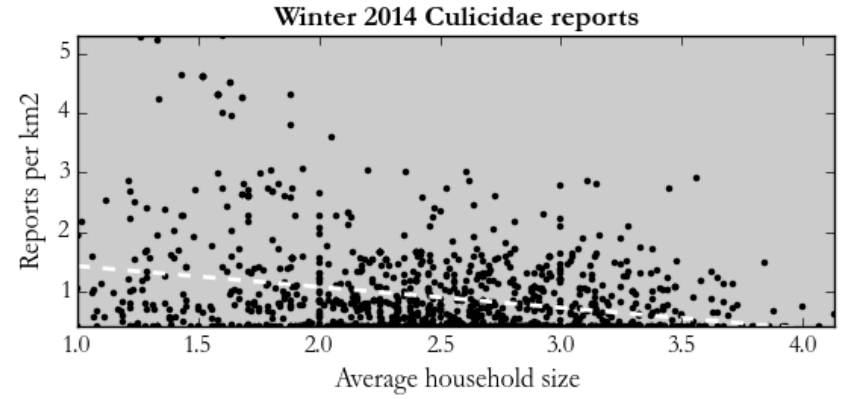


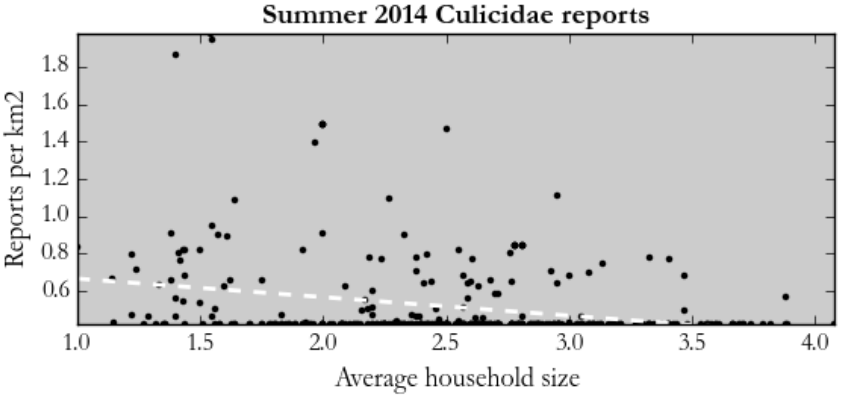

Winter 2014 Culicidae reports
Reports per km2
Percentage 45-64 y/o
r = 0.009094
p = 0.862349

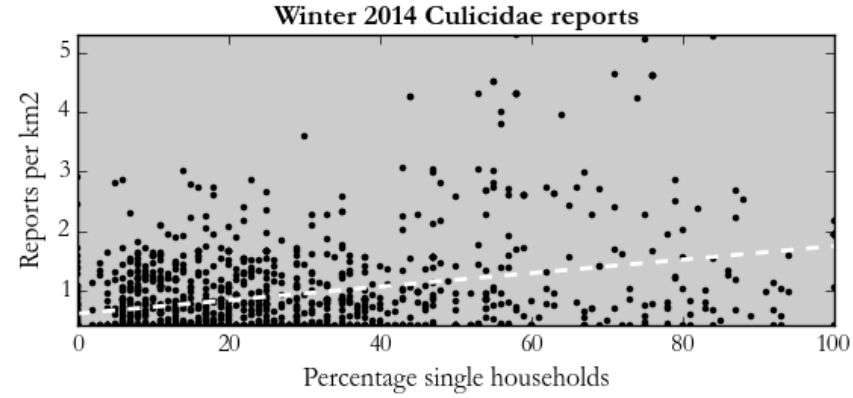

Summer 2014 Culicidae reports
Reports per km2
Percentage 45-64 y/o
r = -0.137072
p = 0.000001

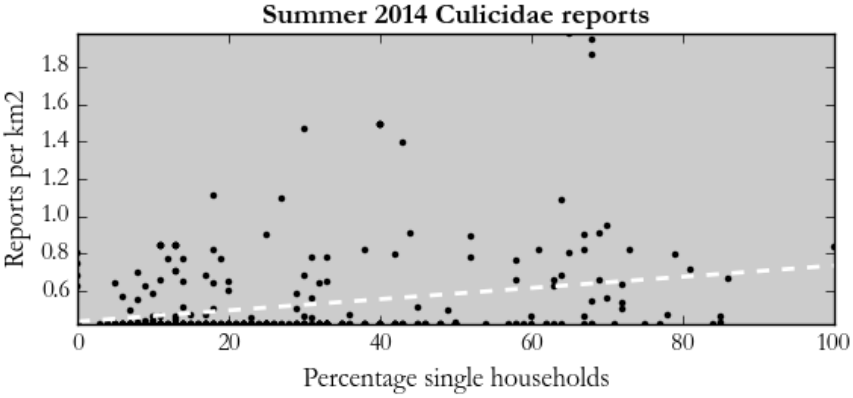

Winter 2014 Culicidae reports
Reports per km2
Non-western immigrants
r = 0.176585
p = 0.000690

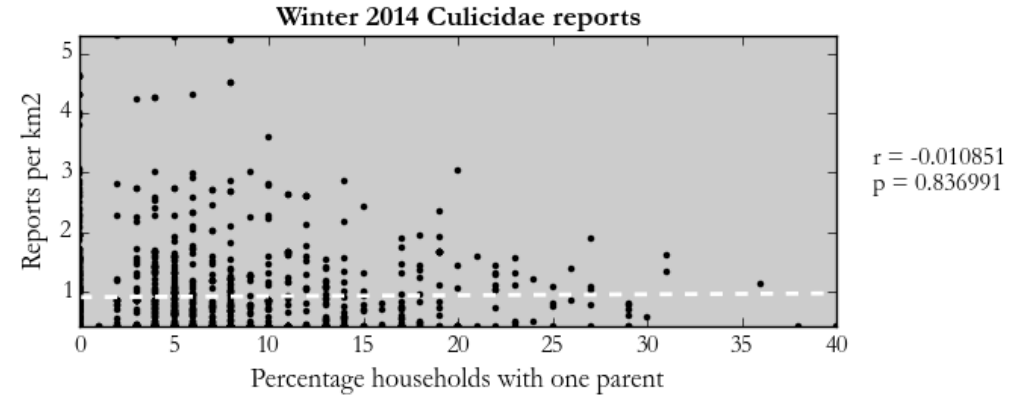

Summer 2014 Culicidae reports
Reports per km2
Non-western immigrants
r = 0.223799
p = 0.000000

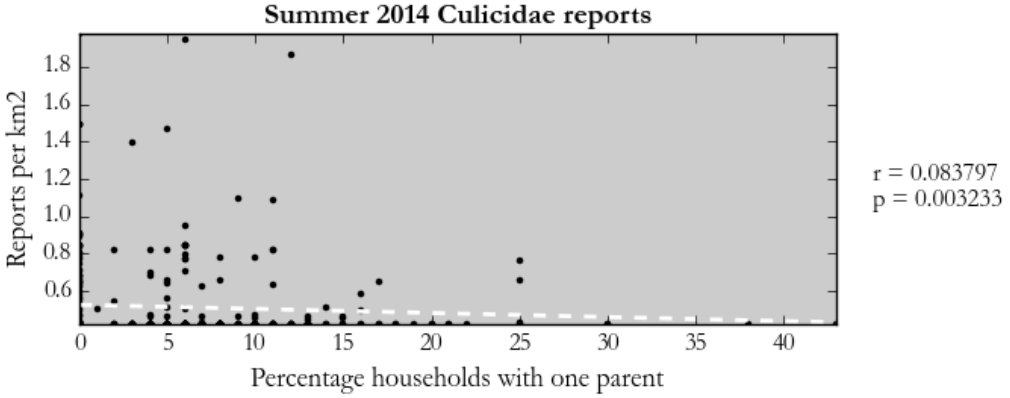

Winter 2014 Culicidae reports
Reports per km2
Average household size
r = -0.213511
p = 0.000042

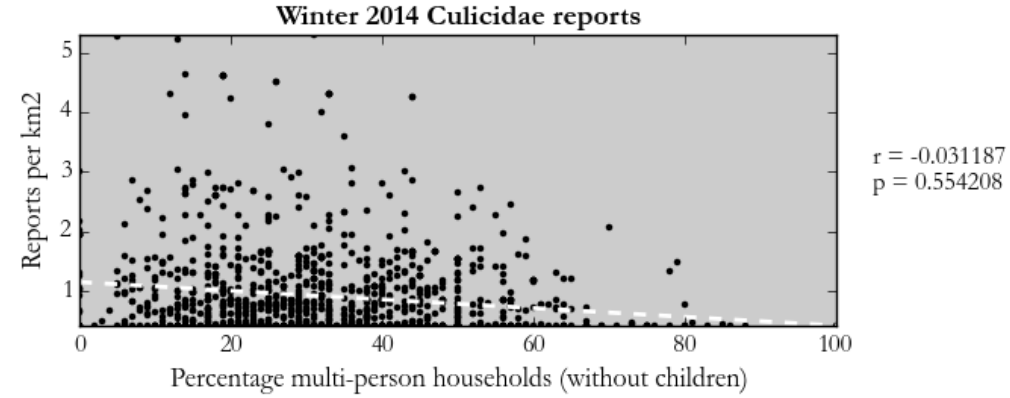

Summer 2014 Culicidae reports
Reports per km2
Average household size
r = -0.204550
p = 0.000000

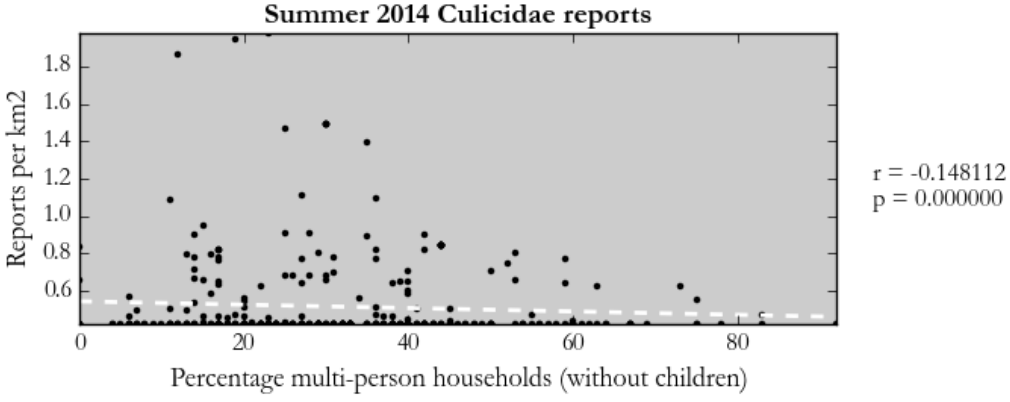

Winter 2014 Culicidae reports
Reports per km2
Percentage single households
r = 0.186553
p = 0.000359

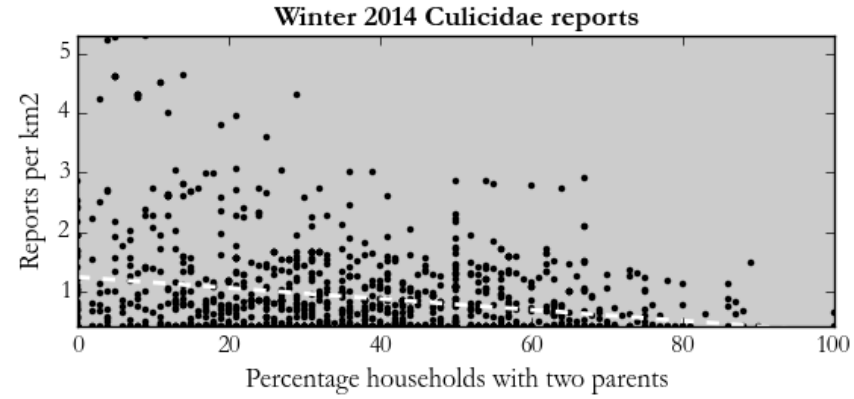

Summer 2014 Culicidae reports
Reports per km2
Percentage single households
r = 0.217146
p = 0.000000

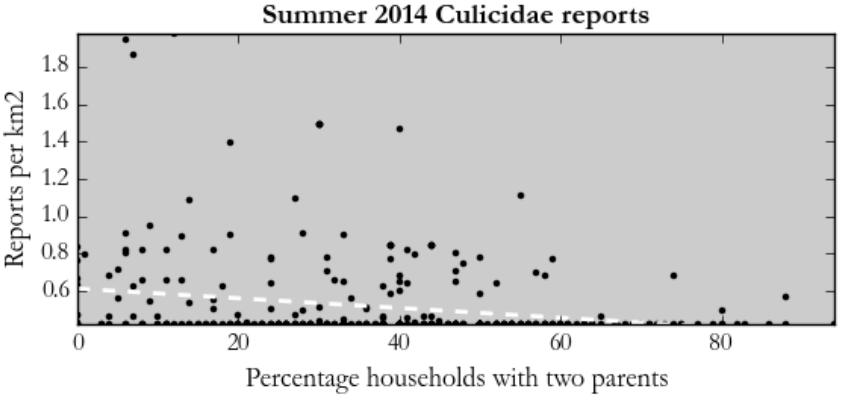

Winter 2014 Culicidae reports
Reports per km2
Percentage households with one parent
r = -0.010851
p = 0.836991

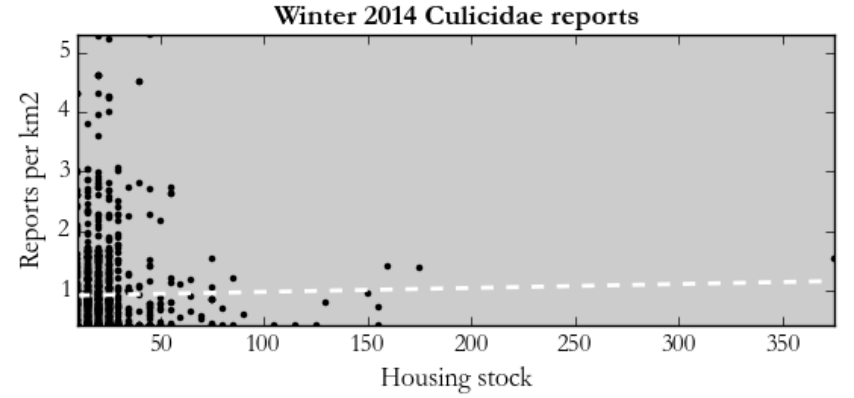

Summer 2014 Culicidae reports
Reports per km2
Percentage households with one parent
r = 0.083797
p = 0.003233

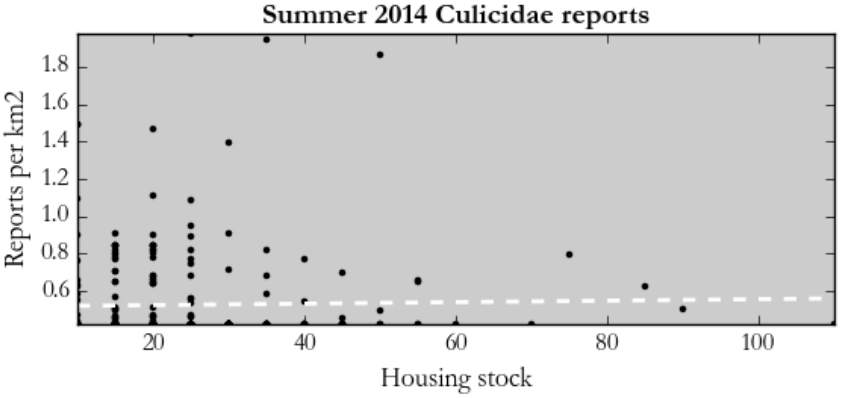

Winter 2014 Culicidae reports
Reports per km2
Percentage multi-person households (without children)
r = -0.031187
p = 0.554208

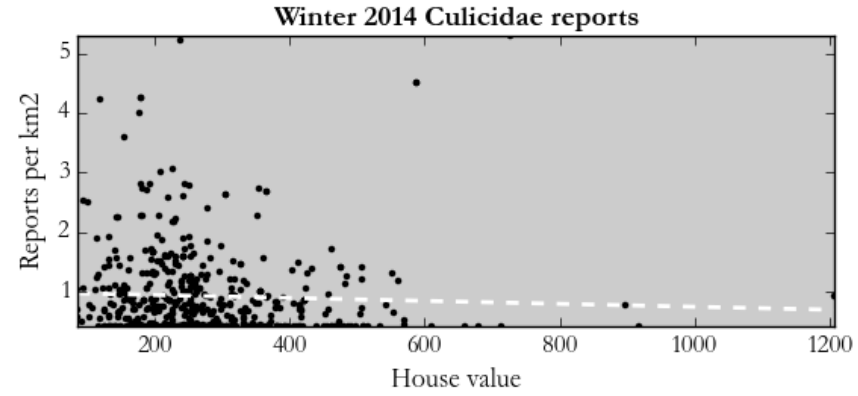

Summer 2014 Culicidae reports
Reports per km2
Percentage multi-person households (without children)
r = -0.148112
p = 0.000000

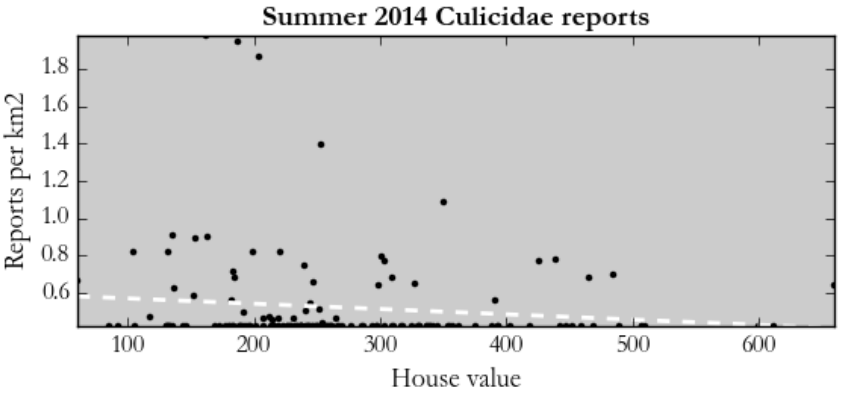

Winter 2014 Culicidae reports
Reports per km2
Percentage households with two parents
r = -0.206729
p = 0.000074


Summer 2014 Culicidae reports
Reports per km2
Percentage households with two parents
r = -0.182534
p = 0.000000


Winter 2014 Culicidae reports
Reports per km2
Housing stock
r = -0.029864
p = 0.598062


Summer 2014 Culicidae reports
Reports per km2
Housing stock
r = 0.063121
p = 0.035573


Winter 2014 Culicidae reports
Reports per km2
House value
r = -0.193441
p = 0.019315


Summer 2014 Culicidae reports
Reports per km2
House value
r = -0.115326
p = 0.008481