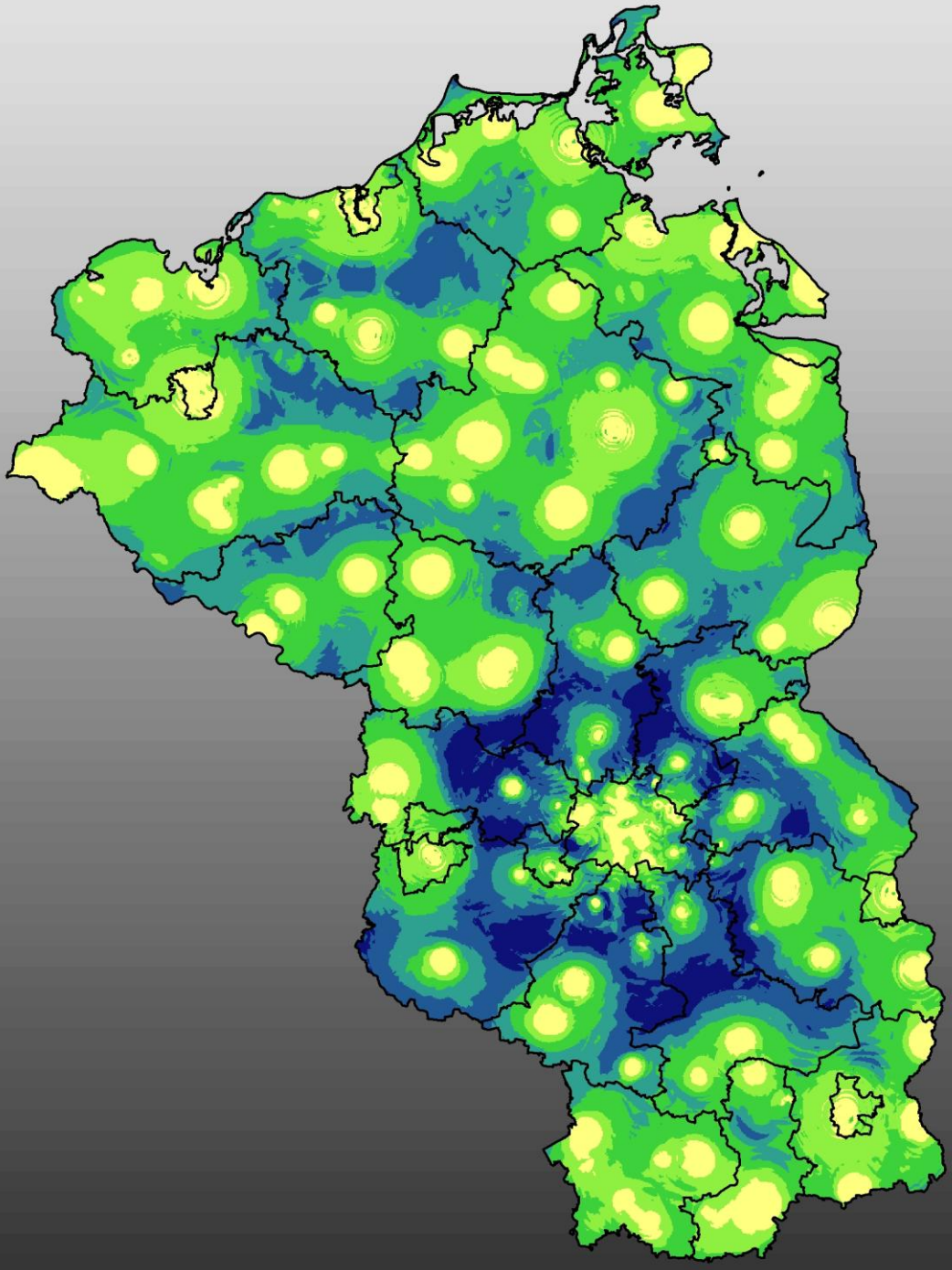
# Geographic Information Systems (GIS) in Public Health

# GEOGRAPHIC INFORMATION SYSTEMS (GIS) IN PUBLIC HEALTH

## How can GIS facilitate demand-based planning of healthcare and targeted prevention strategies?



Boris Kauhl

Cover design and layout: Boris Kauhl
Printed by: A8 Druck- und Medienservice Berlin
ISBN: 978-3-00-058708-5

The research for this dissertation was performed at the department of Health, Ethics and Society, School for Public Health and Primary Care, Faculty of Health, Medicine and Life Sciences, Maastricht University, Maastricht, the Netherlands.

# GEOGRAPHIC INFORMATION SYSTEMS (GIS) IN PUBLIC HEALTH

## How can GIS facilitate demand-based planning of healthcare and targeted prevention strategies?

**Dissertation**

To obtain the degree of Doctor at
Maastricht University on the authority of Rector Magnificus
Prof. dr. Rianne M. Letschert
in accordance with the decision of the Board of Deans,
to be defended in public
On Wednesday, 17th January 2018 at 10.00 hours.

By
Boris Kauhl

**Supervisors:**

Prof. dr. Christian Hoebe

Prof. dr. Thomas Krafft

Prof. dr. Jürgen Schweikart

**Co-supervisor:**

Dr. Nicole Dukers-Muijrers

**Assessment committee:**

Prof. dr. Maurice Zeegers (chair)

Dr. Carijn Beumer

Dr. Wim van der Hoek, RIVM Bijlthoven

Prof. dr. Klasien Horstmann

Prof. dr. med. Thomas Kistemann, University of Bonn

# Contents


| | | |
|---|---|---|
| **CHAPTER 1** | General introduction | 8 |
| **CHAPTER 2** | Estimating the spatial distribution of acute undifferentiated fever (AUF) and associated risk factors using emergency call data in India. A symptom-based approach for public health surveillance. *Health & place, 31, 111-119.* | 25 |
| **CHAPTER 3** | The spatial distribution of hepatitis C virus infections and associated determinants - An application of a geographically weighted poisson regression for evidence-based screening interventions in hotspots. *PloS one, 10(9), e0135656.* | 48 |
| **CHAPTER 4** | Do the risk factors for type 2 diabetes mellitus vary by location? A spatial analysis of health insurance claims in Northeastern Germany using kernel density estimation and geographically weighted regression. *International Journal of Health Geographics, 15(1), 38.* | 76 |
| **CHAPTER 5** | Is the current pertussis incidence only the results of testing? A spatial and space-time analysis of pertussis surveillance data using cluster detection methods and geographically weighted regression modelling. *PLoS ONE 12(3):e0172383.* | 103 |
| **CHAPTER 6** | General discussion | 135 |
| **CHAPTER 7** | **Summary** | 151 |
| | **Nederlandse samenvatting** | 155 |
| | **Acknowledgements** | 159 |
| | **Curriculum Vitae** | 162 |
| | **Publications** | 164 |
| | **Valorisation of this thesis** | 167 |

# CHAPTER 1

# General introduction

## Geographic aspects in public health

The place where people live influences many factors that shape health outcomes of the population. Social, demographic, economic and environmental processes are important determinants of health beyond the influence of the individual and are dependent upon their geographical context (1). To allocate healthcare where it is needed most and to formulate effective prevention strategies, it is not only important to examine how health and disease are distributed across space but also to examine how the environment where people live in affects health outcomes (2, 3). In recent decades, geographic aspects of health and disease have therefore become an important and fast developing field in public health (1, 4).

A qualitative interest in geographic aspects of health and disease can be traced back to the Greek physician Hippocrates (460 – 370 BC). Hippocrates noted an association between climate, seasons, water quality, individual aspects and health of the population. Quantitative approaches in the form of thematic disease maps emerged relatively late in the 18$^{th}$ century (4). Some of the more famous thematic maps are Finke`s world map of human disease and Petermann`s Cholera map of the British isles (4, 5).

Although John Snow`s famous map of Cholera deaths in London dates back to 1854 and therefore later than previous disease maps, he is often cited as the father of spatial analyses in epidemiology, medical geography and disease surveillance (6-8). What set his work apart from previously published disease maps and classified his work as spatial analysis, was the fact that he related different spatial objects to each other and generated new knowledge through this approach: By collecting data on cholera deaths and relating them with the water supply, he was able to successfully identify a contaminated water pump as the origin of the Cholera outbreak. His map was therefore not only a visualization of data, but rather an exploratory analytical tool to solve a specific task at a fine geographic scale. His work also demonstrated that epidemiological data has a geographic component and therefore shows the importance of analysing epidemiological data in a spatial context (9).

## The introduction of Geographic Information Systems to public health

The electronic storage, analysis and manipulation of spatially referenced datasets were made possible with the introduction of Geographic Information Systems (GIS) in the 1970s (9, 10). While these early GIS were mostly used for administrative

procedures, they became frequent tools for many applications, including transportation planning, forestry, emergency medical service delivery, marketing and many more (10). The use of GIS in public health in its early stages focused more on the cartographic visualization of health-related information rather than complex analyses (9). With the increasing availability of electronic, health-related, environmental, demographic and socio-economic spatial data and with technological advances in processing speed, spatial analytical methods became available to an increasing number of researchers (11). Although spatial statistical methods were already available in the 1960s, their availability was boosted in the 1990s and 2000s through further development of open source spatial statistical software such as SaTScan (12), GeoDa (13), CrimeStat (14), R (15) and GWR (16). It should be noted however, that the majority of spatial statistical methods used in public health today have their origin in quantitative geography and spatial econometrics and were rather slowly incorporated in spatial epidemiological studies (11). The relatively steep technological and methodological advance in the past two decades has helped to shift the scope of spatial analyses in public health from a purely cartographic visualization of disease occurrence towards complex analyses of the association between demographic and socio-economic population characteristics, environmental exposure factors and the spatial distribution of diseases (17). GIS are today an established tool in public health and have many purposes, such as planning of healthcare, analysis of healthcare usage and costs, disease surveillance and structural analyses of target populations for specialized treatment (1, 9, 11).

Spatial epidemiological studies using GIS are mostly based on an aggregated study design – often termed “ecological study” - in contradiction to other study designs often used in public health such as surveys (18), questionnaires (19) and cohort studies (20). Spatial epidemiological studies rely often on secondary data. Secondary data are originally collected for administrative procedures and include for example data from health insurance providers, disease registries or pharmacy sales. These data are often only available on an aggregated level and are therefore restricted to the analysis of possible demographic and socio-economic risk factor on aggregated population level. The main advantage of secondary data is that they are already collected, available and can therefore be analysed in a time- and cost-effective manner. Primary data - which are collected specifically for scientific purposes - are often used to analyse socio-economic and behavioural risk factors on an individual level and are time- and cost-intensive to collect (21). While both types of data have their advantages and disadvantages, the

analysis of risk factors based on both data sources has a similar goal: To design interventions for the main population, which is most at risk for specific health problems. This approach is inherent for both, planning and allocation of healthcare and prevention strategies, which are the scope of this dissertation.

For demand-based planning and allocation of healthcare, understanding broader, population-based processes, such as the association between ageing of the population, neighbourhood deprivation and the prevalence of chronic diseases helps to model the expected demand for healthcare. This in turn helps to allocate financial resources where they will be needed most (2, 3).

For prevention strategies however, the analysis of individual and contextual risk factors are both important to target specific behavioural as well as demographic and socio-economic risk factors. Traditionally, risk factors of diseases on an individual level are often analysed through surveys (18), questionnaires (19) and cohort studies (20). Although these study designs provide important insights how individual characteristics affect well-being and disease occurrence (22-24), there are important shortcomings associated with these study designs where GIS can provide an added value:

Analyses of individual risk factors through surveys, questionnaires and cohort studies seldom account for geographic aspects of the environment that are beyond the influence of the individual and have thus only limited use for the design of long-term public health policies. The majority of these study designs is non-spatial in nature, thus the results of these studies imply that incidence or prevalence estimates as well as the association to possible risk factors are equal across the study area (25). In reality however, diseases and associated risk factors are often heterogeneously distributed across space and are dependent upon their geographical context (1, 26). Finding areas with statistically significant higher rates of diseases is important to facilitate cost-effective prevention strategies and to target those population groups who are most in need in specific locations (1). Behavioural risk factors - which are often the main focus of questionnaires, surveys and cohort studies - are often challenging to include in practical prevention strategies as the identification of persons belonging to behavioural risk groups is challenging in the first place. Targeting specific demographic and socio-economic population characteristics has shown to be more effective to include in practical prevention strategies (27). However, individual socio-economic characteristics are seldom available in secondary data due to privacy protection. Ecological analyses based on aggregated disease counts are therefore increasingly used to make inferences

about demographic and socio-economic risk factors as this information is widely available in small-scale population data (1, 4).

The ecological analysis of epidemiological data can be divided into three basic steps: a) the identification of spatial patterns of diseases; b) the detection of areas with significantly elevated rates above average and c) the analysis of risk factors based on aggregated population, environmental or healthcare related area characteristics (1, 4).

Regression models at the ecological level are typically used to evaluate the strength of association between the occurrence of a disease and population-based area characteristics. However, the vast majority of ecological regression methods are global in nature and estimate only one single coefficient per explanatory variable, averaged over the entire study region (16, 28). In reality however, the association between the occurrence of a particular disease and population-based area characteristics varies considerably over space due to cultural, social and environmental processes on an individual and ecological level (1).

It is well documented that the prevalence of chronic diseases within the population increases with age (3, 29). However, not all elderly persons exhibit the same risk of developing a chronic condition. Persons ageing in socially disadvantaged neighbourhoods are at higher risk of developing chronic conditions, often irrespective of individual characteristics (30). Health policies targeting all elderly would be very cost-ineffective. The insight, that the association between the proportion of elderly and the prevalence of chronic conditions is stronger in more socially disadvantaged neighbourhoods, helps to facilitate demand-based planning and allocation of healthcare and to design more targeted prevention strategies. A similar problem applies to the identification of high-risk groups for infectious diseases as not all persons with specific socio-demographic characteristics automatically exhibit the same risk of infection (31, 32).

Spatial ecological studies should logically not only be capable of analysing who is at risk, but also who is where at risk. This critical information is important to facilitate cost-effective, demand-based planning and allocation of healthcare and targeted prevention strategies. This approach can be considered a core capacity of GIS in public health. This thesis therefore aims to evaluate the use of GIS and spatial epidemiological methods for planning and allocation of healthcare and targeted prevention strategies.

## Basic concepts of Geographic Information Systems

A Geographic Information System allows representing complex spatial processes and structures in the form of models. A wide range of definitions exist, with differences reflecting the personal expertise and focus of the respective authors (4). In its simplest definition, a "Geographic Information System is a technology for encoding, storing, manipulating, analysing, retrieving, transforming and displaying spatial and non-spatial data in an efficient and systematic manner" (33). It is clear from this definition that the underlying data source is the key component of a GIS.

The input data of a GIS can be distinguished by two types of data: geometric and attribute data. Geometric data represent the spatial dimension of the object and the attribute data represent the characteristics of the object (4).

## Data sources

### Geometric data

Geometric data can be represented as two different types of spatial data: Raster data and vector data.

*Raster data* describe spatial objects as rectangular cells where each cell represents information (4, 34). In public health, raster data are often used for population data independent of administrative boundaries (35) or environmental data such as land cover or temperature to evaluate the risk of disease occurrence associated with exposure to environmental factors (36).

*Vector data* can be further categorized as *points, lines and polygons*.

*Points* are included in lines and polygons. Two or more points define a line and several lines, creating an enclosed area, define a polygon (4). As standalone feature, points are often used in public health to represent an exact street address. This can be for example the exact address of an infected individual (37), the location of a general practitioner (38), a hospital (39) or other features where the exact location is of interest (40).

*Lines* are commonly used to represent street networks or the Euclidian distance to specific features (41). Of particular interest in public health is the accessibility of healthcare facilities or distance from infected individuals to possible exposure factors. Vector-based street networks are extensively used to study accessibility of populations to healthcare – often represented as driving distance or driving duration from an individual`s address to the address of a specific healthcare facility (42).

*Polygons* are by far the most commonly used type of spatial data in public health. Typically, polygons represent administrative areas such as postal codes (3), municipalities (43), counties (44) or states (45). The majority of population-based, demographic and socio-economic data is only available on an aggregated level, making administrative areas an attractive source of population data for spatial epidemiological studies (45).

**Attribute data**

As the majority of epidemiological data, such as counts of cases as well as demographic and socio-economic data, are often available as table with an administrative code per case or area characteristic, non-spatial attribute data become spatial objects when they are combined with geometric data.

**Combining different data sources for spatial epidemiological research**

The basic concept of a GIS is to model the real world in the form of different layers. These different layers may represent point data of surveillance systems (disease data), street networks, environmental data represented as raster data and socio-demographic population characteristics represented as polygon data (4). To measure the influence of different area characteristics on the occurrence of a particular disease, these different data sources are aggregated to a common administrative unit, for which all relevant data are available or can be aggregated while keeping the loss of information as little as possible.

**Spatial epidemiological methods of this thesis**

**Disease mapping**

The majority of spatial epidemiological research starts with a map of the incidence or prevalence of the disease of interest (1). The type of data available largely determines the respective method. These data are typically divided into point or polygon data.

*Point data* where each point refers to the exact street address of individuals is the smallest spatial scale, which can be used to calculate the incidence or prevalence of a disease. If both, cases and population data, are available as point data, a kernel density estimation can be used to calculate the incidence or prevalence of a disease without the limitations of relatively arbitrary administrative units. The kernel density estimation –

in its simplest form – calculates the density of cases per $km^2$ based on a grid of rectangular cells. The additional density of population (or controls) can then be used to divide the density of cases by the density of the background population to obtain the resulting incidence or prevalence of the disease of interest, irrespective of administrative boundaries (14).

In most cases, disease data are aggregated to *polygons* of administrative units to protect the privacy of the individual. A common goal of disease mapping is to display the incidence of a particular disease at the smallest administrative unit while reducing the problem of unstable disease rates due to varying population densities at small spatial scales (40, 46).

A major concern when analysing diseases based on administrative units is that these units were not designed for spatial epidemiological analyses in the first place. The population within these units usually varies considerably. This constitutes an important challenge, as the corresponding rates may be highly unstable, the smaller the administrative unit is. Consider two municipalities, where one municipality has only 1000 inhabitants and the neighbouring municipality has 10,000 inhabitants. An increase of only one case in both municipalities would increase the risk of disease occurrence tenfold in the municipality with 1000 inhabitants as compared to the municipality with 10,000 inhabitants. This large difference in the risk of disease occurrence is therefore only attributable to the underlying population density of relatively arbitrary administrative units and may not necessarily reflect the "true" risk of disease occurrence (46).

Various methods exist for reducing the instability of disease rates in small administrative units through borrowing strength from other administrative units. The most widely used approach in spatial epidemiology is Bayesian smoothing. Bayesian smoothing can be divided into three main sub-methods: Empirical Bayesian smoothing where the rates are weighted towards a global mean (13); spatial empirical Bayesian smoothing where the rates are weighted towards the mean of neighbouring areas (13) and hierarchical Bayesian smoothing using the Besag-York-Mollié model with a conditional autoregressive prior where the rates are weighted towards the local mean of neighbouring areas while allowing adjustment for socio-economic covariates (43). The Besag-York-Mollié model is widely applied in the spatial analysis of cancer where the risk of cancer after adjusting for known risk factors is of interest (43, 47). In the analysis of other diseases than cancer, the smoothed but unadjusted disease occurrence is of

main interest as risk factors are typically analysed separately in the subsequent regression analysis (3, 44, 48).

**Local cluster tests**

Although the cartographic visualization of disease risk is an important first step to visualize heterogeneity of disease risk within a region, a prioritization of areas for public health interventions based on the cartographic visualization of disease risk alone is error-prone for following reasons: In areas with few inhabitants, the incidence of a disease may be high although the underlying number of cases is fairly small. Pure visual inspection of the disease incidence may lead to small administrative units in urban areas being overlooked, while large rural administrative units with few cases may be more dominant on the map. As a consequence, more sophisticated methods are necessary to overcome these limitations and to justify the selection of areas for interventions. In this context, local cluster tests have become an important tool in public health (11). A local cluster test is a statistical test to evaluate the location and the significance of areas with higher than expected disease rates. Several local cluster tests exist with the local indicator of spatial association, the Besag-Newell test and the spatial scan statistic being the most widely used (11).

**Spatial regression modelling**

A major issue for spatial regression modelling is the presence of spatial dependence, often termed spatial autocorrelation (49). Spatial autocorrelation refers to values in a specific area displaying similar values in nearby areas and is inherent in most spatial datasets. Thus, the assumption of independence of observations, which the basic ordinary least squares (OLS) regression model assumes, is violated. This may lead to exaggerated estimates of the regression coefficients and global clustering of the residuals, deteriorating the reliability of the regression model (50). A spatial regression modelling approach differs from a non-spatial regression approach as spatial regression models assume strong variations and spatial dependence of areas as inherent in the data and an important feature, which should be accounted for (51). Non-spatial regression models in contradiction, treat strong variations and particularly outliers as nuisance, which should be removed (52). Spatial dependency of disease rates is rather the norm than the exception as the majority of diseases are clustered in space (1, 3, 47, 53). The acknowledgement of spatial dependency among observations can be considered as

inherent to geography and regional studies and is expressed in Tobler`s first law of geography: “everything is related to everything else, but near things are more related than distant things” (49, 54). Based on this assumption, several spatial regression approaches have emerged in the field of spatial econometrics, which incorporate spatial dependency and neighbourhood effects in spatial regression modelling. These approaches have been also widely applied in spatial epidemiological studies (11). The major spatial econometric regression approaches applied in public health can be further divided into simultaneous autoregressive (SAR) and conditionally autoregressive (CAR) regression models (55).

The SAR model family can be further distinguished into three different regression models, depending on where the spatial autocorrelation is thought to occur. Spatial autocorrelation may either be present in the dependent variable only (spatial-lag model), simultaneously in the dependent variable and the independent variables (lagged-mixed model) or only in the error terms and thus neither in the dependent or independent variables (spatial-error model) (56).

The CAR model is – to an extent – similar to the SAR model as it also incorporates the effect of spatial dependency in the regression parameters. The main difference between a SAR and a CAR model is that the CAR model is usually modelled in a Bayesian framework using Markov chain Monte Carlo (MCMC) simulations (57), whereas the SAR models are frequentist approaches to spatial regression modelling (11).

**Geographically weighted regression modelling**

The OLS model and both spatial econometric regression families – the SAR and CAR models – are global in nature and estimate only one single coefficient per explanatory variable. The effect of a possible explanatory variable is expected to have the same impact on the modelled outcome of interest in the entire study area (56).

However, given the influence of cultural, social and environmental processes on an individual and ecological level on disease occurrence (1), the assumption that an explanatory variable has the same effect in all locations of the study area is fairly unrealistic – especially for larger geographic areas (16). Logically, associations should be allowed to vary over space in a spatial regression approach to realistically capture the association between an explanatory variable and disease occurrence.

Geographically weighted regression (GWR) modelling - a spatial regression modelling approach capable of analysing spatially varying associations - was originally

developed to model the associations between house prices and house characteristics (16) and has since its introduction gained increasing attractiveness in spatial epidemiological studies (58, 59). GWR has been widely used to identify location-specific demographic and socio-economic population characteristics for both, chronic diseases (3, 60) as well as infectious diseases (31, 32).

The statistical concept behind GWR is to use a circular kernel, which moves over the coordinates of the study region. The centre of the kernel is the regression point. The observations within the kernel are weighted with decreasing intensity towards the edge of the kernel with observations outside the kernel receiving a weight of zero in the regression equation. GWR provides a vast amount of flexibility: The regression family can be either Gaussian, or can be specified as a generalized linear model such as Poisson or logistic regressions. The kernel can be specified with a fixed kernel in km or an adaptive kernel specified as the number of observations included inside the kernel. The optimization of the kernel-bandwidth can be specified manually or can be automated using Akaike's corrected Information Criterion (AICc), Akaikes`s Information Criterion (AIC), Cross Validation (CV) and Bayesian Information Criterion (61). The kernel may take the form of a Gaussian, bisquare, exponential, tricube and boxcar kernel (62). The goodness-of-fit statistics of a GWR can also be compared to those of a global spatial regression approach to test the hypothesis that a local regression approach provides a better fit to the data than a global approach (62, 63).

**Aims and outlines of this thesis**

The overall aim of this thesis is to evaluate how GIS and spatial epidemiological methods are useful to inform evidence-based strategies in public health. There are three main applications where GIS and spatial epidemiological methods could be potentially useful:

1. Demand-based planning and allocation of healthcare
2. Evidence-based prevention strategies
3. Detection of outbreaks for public health surveillance

For the first aim, the spatial distribution of type 2 Diabetes Mellitus based on data of a large health insurance provider in Germany was analysed. For the second aim, four case studies analysing demographic, socio-economic and environmental risk factors for acute undifferentiated fever in India, Hepatitis C in the Netherlands, type 2 Diabetes

Mellitus in Germany and pertussis in the Netherlands were conducted. For the third aim, the use of spatial epidemiological modelling and space-time cluster detection to identify possible pertussis outbreaks in the Netherlands was evaluated.

**Research questions**

This thesis aims to answer the following specific research questions:

1. Is there an added value of the spatial scan statistic to identify areas for prevention strategies (chapter 2, 3, 4, 6)?
2. How can GIS and spatial regression modelling facilitate demand-based allocation of healthcare (chapter 4)?
3. Is geographically weighted regression modelling a suitable method to identify location-specific risk groups for targeted prevention strategies (chapter 3, 4)?
   - 3.1. Which statistical properties of geographically weighted regression modelling have to be considered to obtain useful results (chapter 6)?
   - 3.2. What are current limitations of geographically weighted regression modelling (chapter 6)?
4. Can geographically weighted regression and space-time cluster detection facilitate the detection of possible pertussis outbreaks in the Netherlands (chapter 5)?

**Research design**

This thesis was based on four quantitative, spatial epidemiological studies. Question 1 was answered through the case studies on acute undifferentiated fever in India, Hepatitis C in the Netherlands and type 2 Diabetes Mellitus in Germany. The general relevance of local cluster tests is discussed in chapter 6. Question 2 was answered through the analysis of type 2 Diabetes Mellitus based on data of a large German health insurance provider. Question 3 was answered through the case studies on Hepatitis C in the Netherlands and type 2 Diabetes Mellitus in Germany. Questions 3.1. and 3.2. are discussed in chapter 6 based on the gained knowledge of the case studies applying GWR. Question 5 was answered through the case study on pertussis in the Netherlands.

**Thesis outline**

**Chapter 1:** This chapter introduces the aim of this thesis with specific focus on the use of GIS and spatial epidemiological methods in public health

**Chapter 2** presents the first case study on acute undifferentiated fever in India

**Chapter 3** presents the second case study on Hepatitis C in the Netherlands

**Chapter 4** presents the third case study on type 2 Diabetes Mellitus in Germany

**Chapter 5** presents the fourth and last case study on pertussis in the Netherlands

**Chapter 6** discusses the main findings of this thesis

**References:**


1. Cromley EK, McLafferty SL. GIS and public health: Guilford Press; 2011.
2. Grundmann N, Mielck A, Siegel M, Maier W. Area deprivation and the prevalence of type 2 diabetes and obesity: analysis at the municipality level in Germany. BMC public health. 2014;14(1):1.
3. Dijkstra A, Janssen F, De Bakker M, Bos J, Lub R, Van Wissen LJ, et al. Using spatial analysis to predict health care use at the local level: a case study of type 2 diabetes medication use and its association with demograpHic change and socioeconomic status. PloS one. 2013;8(8):e72730.
4. Schweikart J, Kistemann T. Geoinformationssysteme im Gesundheitswesen. Einführung und praktische Anwendung Heidelberg. 2004.
5. Barrett FA. Finke's 1792 map of human diseases: the first world disease map? Social science & medicine. 2000;50(7):915-21.
6. Meade MS. Medical geography: Wiley Online Library; 1988.
7. Beaglehole R, Bonita R, Kjellström T. Einführung in die Epidemiologie. Bern. Göttingen, Toronto, Seattle: Hans Huber Verlag; 1997.
8. Lawson AB, Kleinman K. Spatial and syndromic surveillance for public health: John Wiley & Sons; 2005.
9. Kistemann T, Schweikart J. Spatial turn. Bundesgesundheitsblatt-Gesundheitsforschung-Gesundheitsschutz. 2017:1-9.
10. Clarke KC, McLafferty SL, Tempalski BJ. On epidemiology and geographic information systems: a review and discussion of future directions. Emerging infectious diseases. 1996;2(2):85.
11. Auchincloss AH, Gebreab SY, Mair C, Roux AVD. A review of spatial methods in epidemiology, 2000–2010. Annual review of public health. 2012;33:107.
12. Kulldorff M. SaTScan user guide for version 9.0. 2010.
13. Anselin L. Exploring spatial data with GeoDaTM: a workbook. Urbana. 2004;51:61801.
14. Levine N. CrimeStat III: a spatial statistics program for the analysis of crime incident locations (version 3.0). Houston (TX): Ned Levine & Associates/Washington, DC: National Institute of Justice. 2004.
15. Team RDC. R: A language and environment for statistical computing. : R Foundation for Statistical Computing, Vienna, Austria.; 2008 [cited 2017 Sep. 25th]. Available from: http://www.r-project.org.
16. Fotheringham AS, Brunsdon C, Charlton M. Geographically weighted regression: John Wiley & Sons, Limited; 2003.
17. Elliott P, Wartenberg D. Spatial epidemiology: current approaches and future challenges. Environmental health perspectives. 2004:998-1006.
18. Bloom B, Cohen RA, Freeman G. Summary health statistics for US children: National Health Interview Survey, 2008. Vital and health statistics Series 10, Data from the National Health Survey. 2009 (244):1-81.
19. Burström K, Johannesson M, Diderichsen F. Health-related quality of life by disease and socio-economic group in the general population in Sweden. Health policy. 2001;55(1):51-69.
20. Reilly JJ, Armstrong J, Dorosty AR, Emmett PM, Ness A, Rogers I, et al. Early life risk factors for obesity in childhood: cohort study. Bmj. 2005;330(7504):1357.
21. Sorensen H, Sabroe S, Olsen J. A framework for evaluation of secondary data sources for epidemiological research. International journal of epidemiology. 1996;25(2):435-42.

22. Lee DS, Chiu M, Manuel DG, Tu K, Wang X, Austin PC, et al. Trends in risk factors for cardiovascular disease in Canada: temporal, socio-demographic and geographic factors. Canadian Medical Association Journal. 2009;181(3-4):E55-E66.
23. Heidemann C, Du Y, Scheidt-Nave C. Diabetes mellitus in Deutschland. 2011.
24. Wolf-Maier K, Cooper RS, Banegas JR, Giampaoli S, Hense H-W, Joffres M, et al. Hypertension prevalence and blood pressure levels in 6 European countries, Canada, and the United States. Jama. 2003;289(18):2363-9.
25. Cornberg M, Razavi HA, Alberti A, Bernasconi E, Buti M, Cooper C, et al. A systematic review of hepatitis C virus epidemiology in Europe, Canada and Israel. Liver International. 2011;31(s2):30-60.
26. Lawson AB. Statistical methods in spatial epidemiology: John Wiley & Sons; 2013.
27. Vermeiren AP, Dukers-Muijrers NH, van Loo IH, Stals F, van Dam DW, Ambergen T, et al. Identification of hidden key hepatitis C populations: an evaluation of screening practices using mixed epidemiological methods. PloS one. 2012;7(12):e51194.
28. Anselin L, Florax R. New directions in spatial econometrics: Springer Science & Business Media; 2012.
29. Ford MM, Highfield LD. Exploring the Spatial Association between Social Deprivation and Cardiovascular Disease Mortality at the Neighborhood Level. PloS one. 2016;11(1):e0146085.
30. Gilleard C, Hyde M, Higgs P. The impact of age, place, aging in place, and attachment to place on the well-being of the over 50s in England. Research on Aging. 2007;29(6):590-605.
31. Weisent J, Rohrbach B, Dunn JR. Socioeconomic determinants of geographic disparities in campylobacteriosis risk: a comparison of global and local modeling approaches. International journal of health geographics. 2012;11(1):45.
32. Feldacker C, Emch M, Ennett S. The who and where of HIV in rural Malawi: Exploring the effects of person and place on individual HIV status. Health & place. 2010;16(5):996-1006.
33. Reddy DV. Engineering Geology. New Delhi: Vikas Publishing; 2010.
34. ESRI. What is raster data? : ESRI; 2008 [cited 2016 October, 10th]. Available from: http://webhelp.esri.com/arcgisdesktop/9.2/index.cfm?TopicName=What_is_raster_data%3F.
35. Lemke D, Mattauch V, Heidinger O, Pebesma E, Hense HW. Comparing adaptive and fixed bandwidth-based kernel density estimates in spatial cancer epidemiology. International journal of health geographics. 2015;14:15. PubMed PMID: 25889018. Pubmed Central PMCID: 4389444.
36. Ali M, Emch M, Donnay JP, Yunus M, Sack RB. Identifying environmental risk factors for endemic cholera: a raster GIS approach. Health & place. 2002 Sep;8(3):201-10. PubMed PMID: 12135643.
37. Yamamura M, Freitas IMd, Santo Neto M, Chiaravalloti Neto F, Popolin MAP, Arroyo LH, et al. Spatial analysis of avoidable hospitalizations due to tuberculosis in Ribeirao Preto, SP, Brazil (2006-2012). Revista de saude publica. 2016;50.
38. Guagliardo MF. Spatial accessibility of primary care: concepts, methods and challenges. International journal of health geographics. 2004;3(1):1.
39. Yamashita T, Kunkel SR. The association between heart disease mortality and geographic access to hospitals: county level comparisons in Ohio, USA. Social science & medicine. 2010;70(8):1211-8.

40. Waller LA, Gotway CA. Applied spatial statistics for public health data: John Wiley & Sons; 2004.
41. Higgs G. A literature review of the use of GIS-based measures of access to health care services. Health Services and Outcomes Research Methodology. 2004;5(2):119-39.
42. Boscoe FP, Henry KA, Zdeb MS. A nationwide comparison of driving distance versus straight-line distance to hospitals. The Professional Geographer. 2012;64(2):188-96.
43. Pollán M, Ramis R, Aragonés N, Pérez-Gómez B, Gómez D, Lope V, et al. Municipal distribution of breast cancer mortality among women in Spain. BMC cancer. 2007;7(1):1.
44. Hipp JA. Spatial analysis and correlates of county-level diabetes prevalence, 2009-2010. Preventing chronic disease. 2015;12.
45. Pandey A, Sahu D, Bakkali T, Reddy D, Venkatesh S, Kant S, et al. Estimate of HIV prevalence and number of people living with HIV in India 2008–2009. BMJ open. 2012;2(5):e000926.
46. Lawson AB. Bayesian disease mapping: hierarchical modeling in spatial epidemiology: CRC press; 2013.
47. Johnson GD. Small area mapping of prostate cancer incidence in New York State (USA) using fully Bayesian hierarchical modelling. International journal of health geographics. 2004;3(1):1.
48. Marotta P. Assessing Spatial Relationships Between Rates of Crime and Rates of Gonorrhea and Chlamydia in Chicago, 2012. Journal of Urban Health. 2016:1-18.
49. Anselin L. Spatial econometrics: methods and models: Springer Science & Business Media; 2013.
50. Lichstein JW, Simons TR, Shriner SA, Franzreb KE. Spatial autocorrelation and autoregressive models in ecology. Ecological monographs. 2002;72(3):445-63.
51. LeSage JP. An introduction to spatial econometrics. Revue d'économie industrielle. 2008 (3):19-44.
52. ArcGIS E. Regression analysis basics [cited 2016 11th Nov.]. Available from: http://pro.arcgis.com/en/pro-app/tool-reference/spatial-statistics/regression-analysis-basics.htm - GUID-6D27B3A1-FFC6-4BF5-893F-F6D60AB2E783.
53. Barker LE, Kirtland KA, Gregg EW, Geiss LS, Thompson TJ. Geographic distribution of diagnosed diabetes in the US: a diabetes belt. American journal of preventive medicine. 2011;40(4):434-9.
54. Tobler WR. A computer movie simulating urban growth in the Detroit region. Economic geography. 1970;46(sup1):234-40.
55. Bivand RP, E., Gomez-Rubio, V. Applied Spatial Data Analysis with R. 2 ed. New York: Springer; 2013.
56. F Dormann C, M McPherson J, B Araújo M, Bivand R, Bolliger J, Carl G, et al. Methods to account for spatial autocorrelation in the analysis of species distributional data: a review. Ecography. 2007;30(5):609-28.
57. Lee D, Lee MD. Package 'CARBayes'. 2016.
58. Hu M, Li Z, Wang J, Jia L, Liao Y, Lai S, et al. Determinants of the incidence of hand, foot and mouth disease in China using geographically weighted regression models. PloS one. 2012;7(6):e38978. PubMed PMID: 22723913. Pubmed Central PMCID: 3377651.
59. Gebreab SY, Diez Roux AV. Exploring racial disparities in CHD mortality between blacks and whites across the United States: a geographically weighted regression

approach. Health & place. 2012 Sep;18(5):1006-14. PubMed PMID: 22835483. Pubmed Central PMCID: 3693935.
60. Gebreab SY, Roux AVD. Exploring racial disparities in CHD mortality between blacks and whites across the United States: a geographically weighted regression approach. Health & place. 2012;18(5):1006-14.
61. Nakaya T. GWR4 user manual. WWW Document Available online: http://www st-andrews ac uk/geoinformatics/wp-content/uploads/GWR4manual_201311 pdf (accessed on 4 November 2013). 2009.
62. Lu B, Harris P, Charlton M, Brunsdon C. The GWmodel R package: further topics for exploring spatial heterogeneity using geographically weighted models. Geo-spatial Information Science. 2014;17(2):85-101.
63. ESRI. Regression analysis basics: ESRI; 2016 [cited 2017 Jan. 30th]. Available from: http://desktop.arcgis.com/en/arcmap/10.3/tools/spatial-statistics-toolbox/regression-analysis-basics.htm.

# CHAPTER 2

# Case study on Acute Undifferentiated Fever in India

published as:

Kauhl, B., Pilot, E., Rao, R., Gruebner, O., Schweikart, J., & Krafft, T. (2015). Estimating the spatial distribution of acute undifferentiated fever (AUF) and associated risk factors using emergency call data in India. A symptom-based approach for public health surveillance. Health & place, 31, 111-119.

**Abstract:**

The **S**ystem for **E**arly-warning based on **E**mergency **D**ata (SEED) is a pilot project to evaluate the use of emergency call data with the main complaint acute undifferentiated fever (AUF) for syndromic surveillance in India. While spatio-temporal methods provide signals to detect potential disease outbreaks, additional information about socio-ecological exposure factors and the main population at risk is necessary for evidence-based public health interventions and future preparedness strategies. The goal of this study is to investigate whether a spatial epidemiological analysis at the ecological level provides information on urban-rural inequalities, socio-ecological exposure factors and the main population at risk for AUF. Our results displayed higher risks in rural areas with strong local variation. Household industries and proximity to forests were the main socio-ecological exposure factors and scheduled tribes were the main population at risk for AUF. These results provide additional information for syndromic surveillance and could be used for evidence-based public health interventions and future preparedness strategies.

## Introduction

The burden of disease in India is currently changing from being dominated by communicable diseases to chronic life-style related diseases. The overall burden of disease accounts for approx. 269 million disability adjusted life years (DALY) in India. Despite the epidemiological transition, communicable diseases still account for 50% of DALYs followed by 33% for non-communicable diseases and 17% for injuries (1). Infectious and parasitic diseases are the major contributor to communicable diseases followed by respiratory infections, diarrhoeal diseases and childhood diseases (1). Acute undifferentiated fever (2) is a first indicator for infectious diseases and is a major public health problem in India. The aetiology of AUF is fairly diverse and includes a wide range of infectious diseases such as dengue (3), malaria (4), typhoid (5), tuberculosis (6), hantavirus (7) and Japanese encephalitis (8).

Socio-economic disparities are a key driver not only of high rates of infectious diseases (9, 10), especially in rural areas (11), but also of a wide range of other health problems including neonatal mortality (12), inequalities in immunization coverage (13), mental disorders (14) and low birth-weight (15). The vulnerability to infectious diseases among various disadvantaged population sub-groups such as scheduled castes and scheduled tribes varies widely among and within the states of India (16), depending on the local interplay between agent, host and environmental factors (1). A spatial epidemiological approach using Geographic Information Systems (GIS) is therefore essential to estimate the impact of socio-economic and environmental (socio-ecological) characteristics on the incidence of infectious diseases. Such an approach has shown to deliver substantial background information for evidence-based public health interventions (17-19). However, reliable and complete surveillance data is scarce in India (20, 21), making the application of spatial epidemiological methods more challenging.

The federal structure of the Indian public health system with its variety of stakeholders and institutions, the increase of the private medical sector, the missing collaboration between the institutions and the multiplicity of vertically organized surveillance programs with their different systems of data collection complicate a uniform surveillance system (21). The Integrated Disease Surveillance Project (IDSP) was initiated in 2004 by the Ministry of Health and Family Welfare (MOHFW) with financial help of the World Bank and technical assistance of the World Health Organization (WHO) and the US Centers for Disease Control and Prevention (CDC). The

goal of this project was to connect all district hospitals and medical colleges to establish a decentralized, state-based disease and syndromic surveillance system (22). However, this approach is not spatially inclusive as the IDSP still faces problems to include data from the private medical sector and therefore underestimates the burden of disease. The current approach to estimate the burden of disease relies on fragmentary databases derived usually from public medical facilities that serve only a small fraction of the population (21). The importance of including the private medical sector into disease surveillance can best be described by the following numbers: After the turn of the millennium, 67% of all hospitals, 63% of all pharmacies and 78% of all doctors were employed within the private medical sector (11). Additionally, the IDSP still remains suboptimal for the control of infectious diseases. The surveillance data is often delayed, unreliable, inconsistent and the reporting rates display strong regional differences (23).
The **S**ystem for **E**arly-warning based on **E**mergency **D**ata (SEED) is a pilot project set up by GVK Emergency Management Research Institute (GVK EMRI), India`s largest private emergency medical service provider, and GEOMED Research to evaluate the use of emergency call data with the main complaint fever for syndromic surveillance of infectious diseases in India (24, 25). The project is closely linked to the European Emergency data-based **S**ystem for **I**nformation on, **D**etection and **A**nalysis of **R**isks and **T**hreats to **H**ealth (SIDARTHa), (26).

GVK EMRI currently operates in 14 states and 2 union territories of India, providing a chance to set up a large-scale syndromic surveillance system covering a large part of the population. The emergency call data are automatically captured using Computer Telephone Integration technology. These data are standardized, available in near real-time, spatially inclusive at fine geographic scales for the covered areas and allow the use of symptom-based data on AUF to estimate the burden of infectious diseases in areas where reliable surveillance data are not available (4, 8, 24).

While the general use of syndromic surveillance lies in the observation of spatial variations of common illnesses over time (27, 28) and the detection of potential disease outbreaks (24, 29), a purely spatial, cross-sectional epidemiological analysis at the ecological level may provide additional information about socio-economic and environmental risk factors (8, 17, 30).

Infectious diseases presenting with symptoms of fever such as malaria, dengue and typhoid are driven by socio-economic, demographic and environmental characteristics (19, 31, 32) and typically display higher rates in rural areas of India (11).

Location-based knowledge on socio-ecological exposures and the population at risk is critical to allocate scarce financial resources (11). Such knowledge informs future preparedness strategies, for example through targeted distribution of insecticide treated bed nets.

The goal of this study is therefore to examine whether a spatial epidemiological analysis at the ecological level provides background information on the main socio-ecological exposure factors and the population at risk for evidence-based public health interventions and future preparedness strategies. Specifically, we hypothesize (i) that AUF displays higher rates in rural areas as compared to urban areas (ii) that AUF is distributed unequally across space and (iii) that AUF is associated with lower socio-economic status.

## Methods

### Study area

SEED was set up as a pilot project in three districts of Andhra Pradesh (AP), India. These three districts were selected by GVK EMRI based on their proportion of infant mortality rates, female literacy, urbanization, proportion of reported fever and infection cases and proportion of scheduled caste and scheduled tribe population to ensure a representative sample within Andhra Pradesh (25). Srikakulam district was chosen for this study because it has the largest proportion of fever among the three selected districts. A community level household survey estimated the prevalence of fever to be 16.7%. A more detailed analysis revealed that 18% of these fever cases were attributable to malaria, 8% to typhoid and 4% to dengue and the remaining 72% to AUF (25). The district is characterized by a long coastline in the east and forested areas in the northern and north-western parts. Srikakulam has a population of 2.54 mio inhabitants according to the Census of India 2001 (33). The smallest administrative units in rural areas of India are villages, which can be defined as areas with (i) a maximum population of 5,000 inhabitants, (ii) a maximum of 75% of the male population employed in the non-agricultural sector and (iii) a maximum population density of 400 inhabitants per $km^2$ (33). Mandals are the smallest administrative unit in AP for which a wide variety of population statistics are available and comprise between 27,141 and 187,132 inhabitants in Srikakulam district (33). The district is predominantly rural and contains 11% of urban population, which is far lower than the average of 27.3% in Andhra Pradesh (33). The literacy rate may be considered as low with only 54% as compared to

60% for the AP average. Srikakulam has a lower proportion of scheduled caste population with 9.5% as compared to the AP average of 16.0%. The proportion of scheduled tribes is slightly higher with 7.1% than the AP average of 7%.

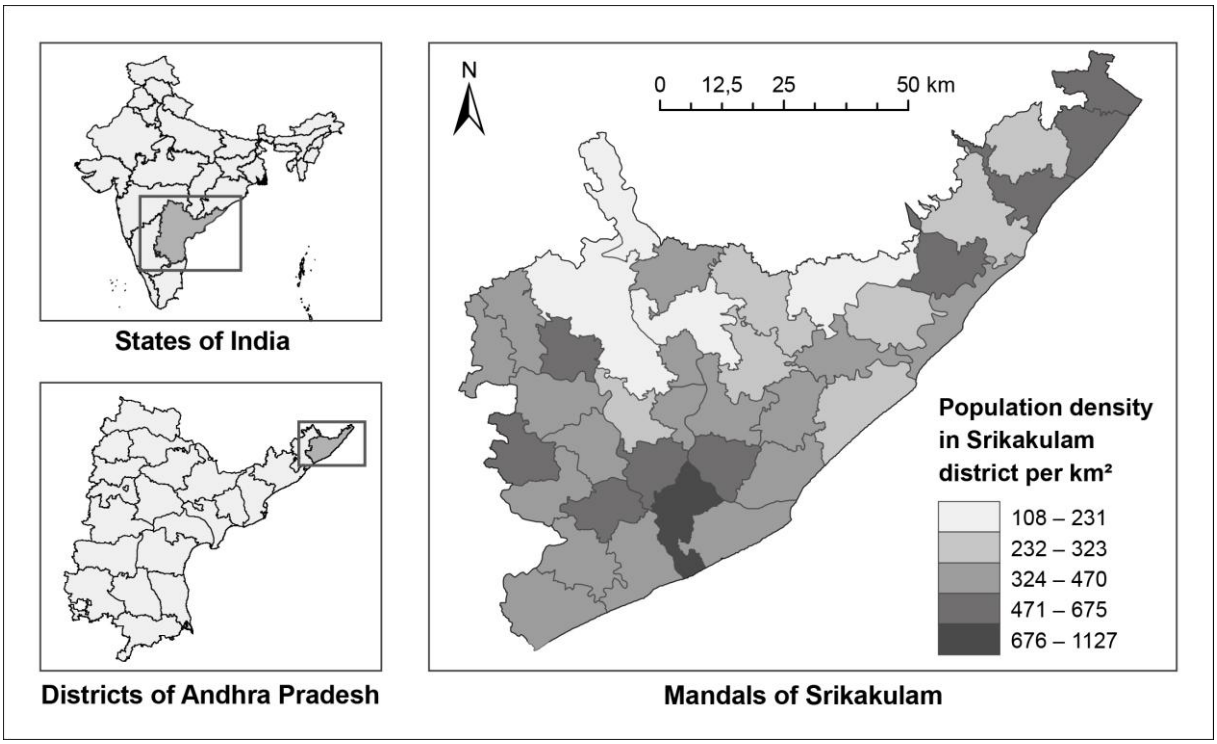



**Fig.1:** Study Area, 2008

## Data

### Outcome variable

Emergency call data with the main complaint AUF were used as indicator for infectious diseases. The emergency call data were provided by GVK EMRI and were available for the time period January 1st to December 31st, 2008. 8,062 AUF calls were recorded for the year 2008 in Srikakulam district. The emergency call data were available on village level and were aggregated to mandal level to be able to use population data, which were obtained from the Census of India 2001 (33). The calculated risk expressed as the number of AUF cases for 2008 per 100,000 inhabitants for each mandal was used as the dependent variable in the regression model.

### Explanatory variables

We included environmental factors associated with vector-borne diseases resulting in AUF such as rainfall (34, 35) and proximity to forests (36, 37). Annual Rainfall data were obtained for the year 2008 based on mandal level from the

Directorate of Economics and Statistics, Hyderabad, Andhra Pradesh, India. Data on forest cover were downloaded from Open Street Map (38). We visually checked the accuracy of the Open Street Map layer. Although not 100% accurate, we found this dataset superior than available raster datasets and sufficient for our analysis. The distance to forests was calculated as distance of the AUF emergency calls on village level to forest, averaged per mandal. To determine whether AUF follows a distinctive socio-economic gradient, we included several socio-economic variables from the Census of India 2001. An overview over all candidate explanatory variables is given in table 1. These variables include the sex ratio for the total population as well as for the child population (aged 0-6) measured as number of female persons per 1,000 male persons. The proportion of scheduled caste and scheduled tribes represents the lowest socio-economic status since these two population groups are historically disadvantaged and have the lowest socio-economic status within the Indian society (39). Literacy rate contains all persons aged seven and older, who are able to read and write in any language. Literacy rate is an important predictor for understanding health-education messages and awareness of health-programs (40). Employment status was split in several categories: General work participation and proportion of main workers were included as indicator for the ability to pay for health-related costs. The variable non-workers includes persons with no personal income and is therefore an indicator for the proportion of persons unable to pay out of pocket for medical expenses. Cultivation and agricultural labour were included as potential predictors for exposure to zoonotic diseases resulting in AUF (7, 41). Household industries are traditionally home-based and are characterized by their high level of exploitation and are another indicator of low socio-economic status (16). Other workers was included as indicator for a higher socio-economic status since this category encompasses work, which requires higher levels of education and therefore generates higher wages such as teachers, municipal servants and government employees. The variable population density was calculated as number of inhabitants per $km^2$. All socio-economic variables and their definitions were obtained from the Census of India 2001 (33). Although the census data used in this study are from 2001, new data from the Census of India were not yet available on mandal level during the time of the analysis.

**Table 1:** Candidate explanatory variables. N = 38

| Variable | Source | Mean | SD |
|---|---|---|---|
| Children aged 0-6 | Census of India | 13.4% | 0.9% |
| Sex ratio total population | Census of India | 1014 | 38 |
| Sex ratio child population | Census of India | 969 | 22 |
| Scheduled caste | Census of India | 9.4% | 4.1% |
| Scheduled tribe | Census of India | 7.1% | 14.6% |
| Literacy rate | Census of India | 54% | 6.7% |
| Work participation | Census of India | 48.5% | 4.3% |
| Main workers | Census of India | 35.1% | 4.9% |
| Cultivation labour | Census of India | 23% | 6.4% |
| Agricultural labour | Census of India | 47.7% | 8.8% |
| Household industries | Census of India | 4% | 1.5% |
| Other workers | Census of India | 25.3% | 12.3% |
| Population density | Census of India | 420.6 | 172.9 |
| Annual Rainfall / mandal | Dir. of Econ. and Stat. | 1027mm | 331mm |
| Distance to forests | Open Street Map | 9.5km | 8.1km |

**Analytical methods**

The methodology applied in this study follows closely the recommendations of the CDC to investigate suspected clusters of cancers (42) and has also been widely applied in a comparable manner to investigate clusters of infectious diseases (8, 17): We created a thematic map displaying the relative risk for each administrative unit; determined spatial clusters where the number of observed cases is higher than the expected cases; and applied a kernel density estimation to visualize the number of cases on village level. We then determined significant explanatory variables through OLS regression.

**Exploratory disease mapping**

To facilitate visual interpretation of the underlying disease process, the relative risk (RR) was calculated for each administrative unit. A map of the RR displays the ratio of observed to expected cases for each administrative unit and represents how much more common an event in this location is as compared to the global average (43). Spatial

Empirical Bayes smoothing of the relative risk was considered useful in this study since the population at risk displayed a strong variation between the administrative areas. This leads to a large variance of the relative risk in areas where the underlying population is small and a small variance in areas where the underlying population is large (44). The RR estimates were smoothed towards a local mean by using a nearest neighbour approach. The neighbours were defined as areas sharing a common edge or boundary (45). We preferred a locally weighted Empirical Bayes smoothing approach over a global approach due to the occurrence of local clusters inherent in our data. The computation was carried out using the EBlocal function of the spdep package available in R (46, 47). For visualisation, the results were then imported in ESRI ArcGIS 10.1.

**Local cluster detection**

To determine administrative areas where the number of observed cases is significantly higher than the expected cases, the spatial scan statistic was applied to search for local clusters of elevated RR. We used the Poisson model where, under the null hypothesis, the cases of AUF follow an inhomogeneous Poisson process (48). We selected the number of AUF cases in 2008, population from the census of India 2001 and the centroid coordinates for each mandal as necessary input data. The spatial scan statistic imposes a circular scanning window over the study area, flexibly in size and position. In this study, we evaluated clusters with 10% of the population at risk. This was done to detect spatial clusters as precisely as possible since the default setting of 50% is more likely to produce results of no practical use (49). The spatial scan statistic compares the observed and expected number of cases inside the scanning window to the area outside the scanning window. The calculation of the maximum likelihood is based on the number of observed and expected cases inside and outside the scanning window. The scanning window with the maximum likelihood and more cases than expected is the most likely cluster. The statistical significance is based on 999 Monte-Carlo replications where the null hypothesis of complete spatial randomness is rejected in this study if the p-value is less than 0.05 (50). The application of the spatial scan statistic was performed in Kulldorf`s SaTScan software version 9.2 (51).

**Kernel density estimation**

The kernel density estimation was used as a complementary tool to visualize the spatial distribution of AUF emergency calls within the spatial clusters. The mandals to

calculate the RR are fairly large spatial units and therefore mask important variations on village level. The kernel density estimation is an interpolation technique that creates a continuous surface derived from a point pattern that allows an easier identification of densely distributed features. This is done by placing a symmetrical mathematical function over each point, the so-called kernel, which has its peak directly over the point with decreasing intensity towards the edge of the function. The distance from the point towards the edges is the bandwidth and determines the amount of smoothing inherent in the kernel density estimation (52). Of the 8,062 fever emergency calls in 2008, 7,366 (91.4%) could be successfully matched with an already existing geodatabase, which contained the coordinates of the village centroids. These village coordinates served as input point pattern for the analysis. In this study, we chose a quartic distribution as mathematical function for the kernel and evaluated bandwidths of 1, 3 and 5 kilometres. We found that a bandwidth of 3 km yielded the best results for our analysis. The calculation of the kernel density estimation was performed using the CrimeStat III software (53). The results were imported in ESRI ArcGIS 10.1 and were displayed together with the layers for forest cover.

**Regression analysis**

The next step of our analysis was to model the influence of potential explanatory variables on the incidence of AUF. We specified our explanatory variables using following criteria: The coefficients are statistically significant and have the expected sign; the explanatory variables do not display multicollinearity and the residuals are normally distributed and are not spatially autocorrelated (54). In order to achieve normality of the dependent variable, the dependent variable was transformed using a natural log-transformation (55, 56). To find a meaningful model, we used a data-mining tool called Exploratory Regression, which is available in ESRI ArcGIS 10.1. This tool is comparable to a step-wise regression. However, this tool identifies variable combinations in an OLS regression model that meet all requirements outlined above (18, 57). The most parsimonious model with the lowest AIC value was used for further analysis. We then applied OLS regression in OpenGeoDa 1.2.0 (58). The calculation of Moran`s I to detect spatial autocorrelation of the residuals was based on first order queen contiguity where neighbours share a common edge or corner (54).

## Results

### Difference between urban and rural risks for acute undifferentiated fever

The overall incidence of AUF was 317 per 100,000 inhabitants. Higher risks could generally be observed in purely rural areas (RR = 1.20, 95% CI: 0.64 - 1.75) as compared to mandals containing urban areas (RR = 0.66, 95% CI: 0.36 - 0.96).

### Spatial inequalities of acute undifferentiated fever

Higher risks were concentrated in the northern parts of the district in close proximity to forests (Fig. 2). The spatial scan statistic detected two clusters. The most significant cluster was located in Seethampeta mandal ($p<0.001$, RR = 9.7, 1621 cases), which is characterized by a high proportion of forest cover. The second cluster consisted of the three mandals Meliaputti, Nandigam and Tekkali ($p<0.001$, RR = 1.74, 943 cases), which are also characterized by their high proportion of forest cover. The kernel density estimation revealed that AUF cases were concentrated in close proximity to, and within a forest in Seethampeta mandal. Especially Seethampeta village stands out with 824 AUF cases. This village contains the largest number of AUF cases per village within the study area. The second largest number of AUF cases within Seethampeta mandal was observed in Pedarama village with 131 cases in close proximity to Seethampeta village. In the second spatial cluster, three concentrations stand out: The town Tekkali with 160 cases, the village Nandigam with 74 cases and the village Meliaputti with 108 cases. A spatial pattern from Meliaputti heading into the forest is visible, leading through the village Padda with 41 cases and Nelabonthu with 29 cases.

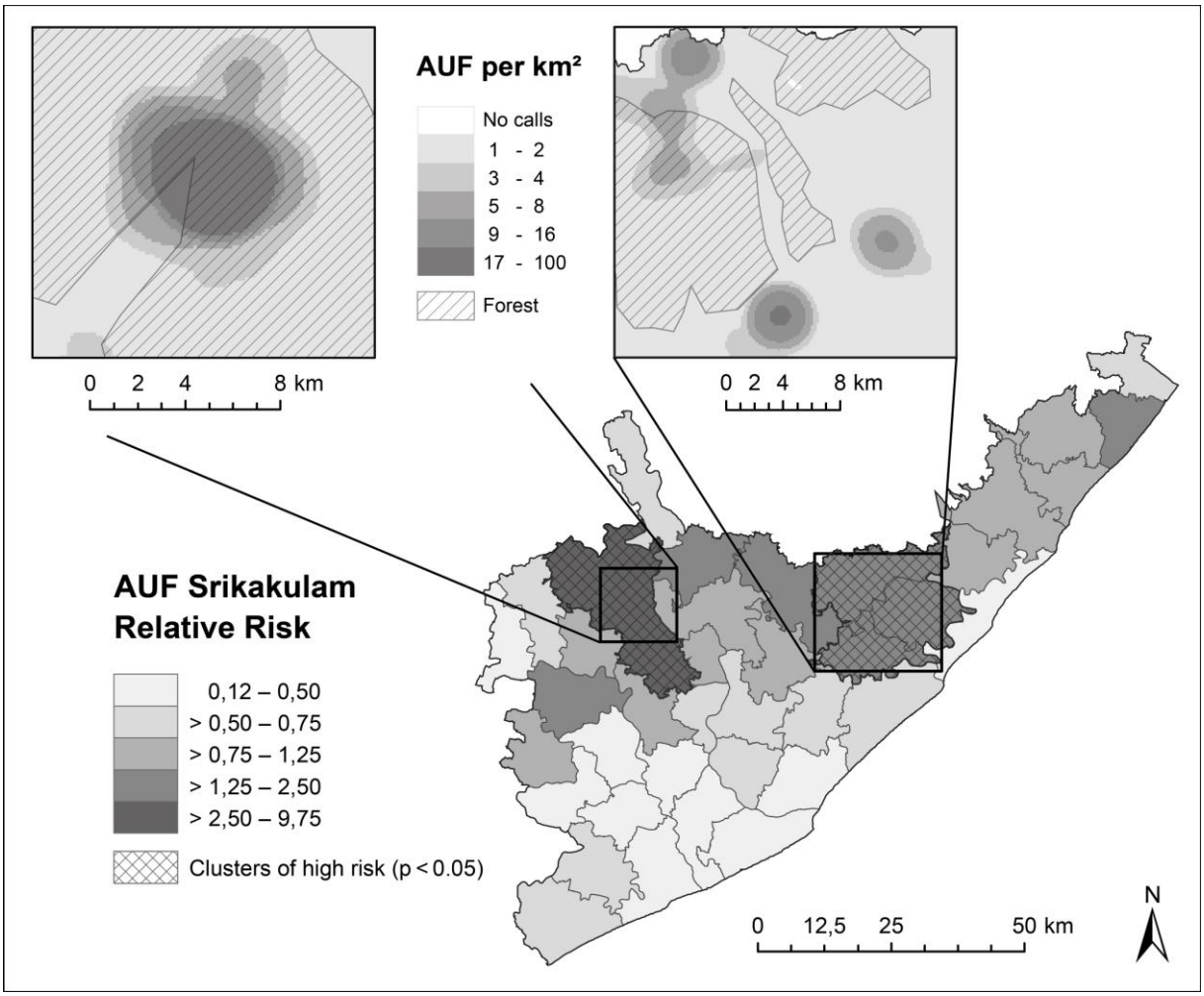



**Fig. 2:** Spatial clustering of AUF in Srikakulam District, 2008. Crosshatched areas indicate local clusters detected by the spatial scan-statistic.

**Socio-ecological exposure factors for acute undifferentiated fever**

The model with the lowest AIC value and the most plausible explanation was used as the final OLS regression model, which included three explanatory variables: Percentage of scheduled tribe population, distance to forests and proportion of household industries. This model explained 66.2% of the variation in AUF emergency calls (Adj. R-squared: 0.6619). The model met all requirements for a properly specified OLS model: The model performance was overall statistically significant (F-statistic: 25.151, $p<0.001$). The coefficients had the expected signs (table 2) and did not display multicollinearity (multicollinearity condition number 7.408). The residuals were normally distributed (Jarque-Bera test: 1.186, $p>0.05$) and were not spatially autocorrelated (Moran`s I: 0.611, $p>0.05$). The Lagrange multiplier tests (LM-lag and LM-error) did not show any spatial dependence (LM-lag: 0.511, $p>0.05$; LM-error: 0.199, $p>0.05$) implying that a spatial error model or a spatial lag model would not enhance the

analysis. The coefficients revealed that the incidence of AUF is positively associated with the proportion of scheduled tribes. An increase of 1% of scheduled tribes population will increase the incidence of AUF by 3%. Proportion of household industries was also positively associated with the incidence of AUF. An increase of 1% of household industries will increase the incidence of AUF by 11.6%. The distance to forest was negatively associated with the occurrence of AUF. 1km more distance to forests will decrease the incidence of AUF by 0.047%.

**Table 2:** Ordinary least squares (OLS) regression coefficients

| Variable | Coefficient | Standard error | Probability |
|---|---|---|---|
| Scheduled tribes | 3.04758 | 0.58251 | <0.001 |
| Househ. industries | 11.60569 | 5.50590 | <0.05 |
| Dist. to forest (m) | -0.04664 | 0.01055 | <0.001 |

By examining the spatial distribution of scheduled tribes (fig. 3), it becomes evident that the proportion of scheduled tribes has a strong link to the incidence of AUF. Especially Seethampeta mandal stands out. In this area, the relative risk is almost 10 times as high as the district average while the proportion of scheduled tribes with 87% is almost 12 times as high as the district average. Comparable findings can also be observed for Meliaputti and Pathapatnam; the incidence of AUF and the proportion of scheduled tribes in these mandals are twice as high as compared to the district average. In contradiction, in the mandals around Srikakulam city, very low relative risks and very low proportions of scheduled tribes can be observed. The association between household industries (fig. 4) however, are not as clear as for scheduled tribes. While it is obvious that household industries have no influence on the occurrence of AUF in Seethampeta mandal, the influence in the mandals Meliaputti, Pathapatnam and the northern mandals Kaviti and Kanchili is probably higher. The incidence of AUF shows a strong link to forested areas, especially in the most significant cluster in Seethampeta mandal but also in the second significant cluster in the mandals Meliaputti, Nandigam and Tekkali.

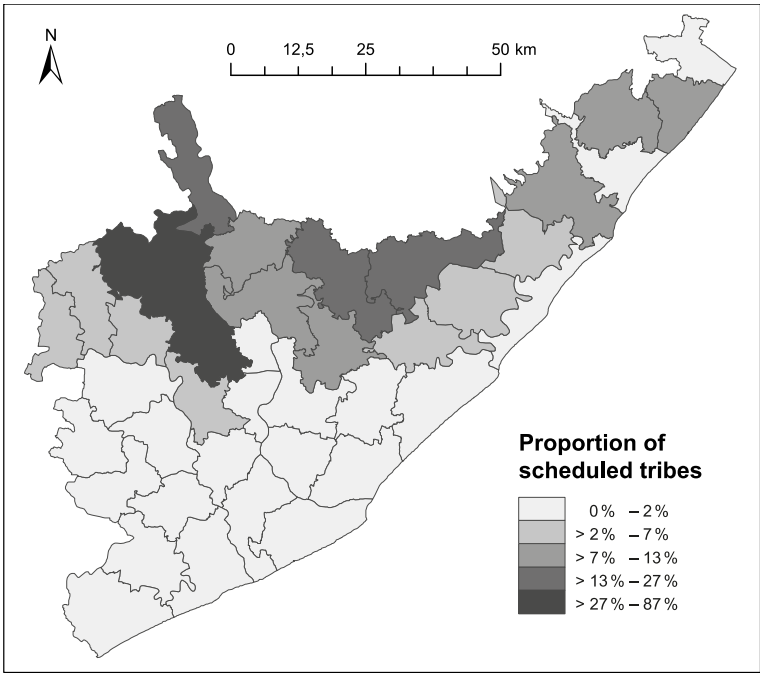



**Fig. 3:** Proportion of scheduled tribes

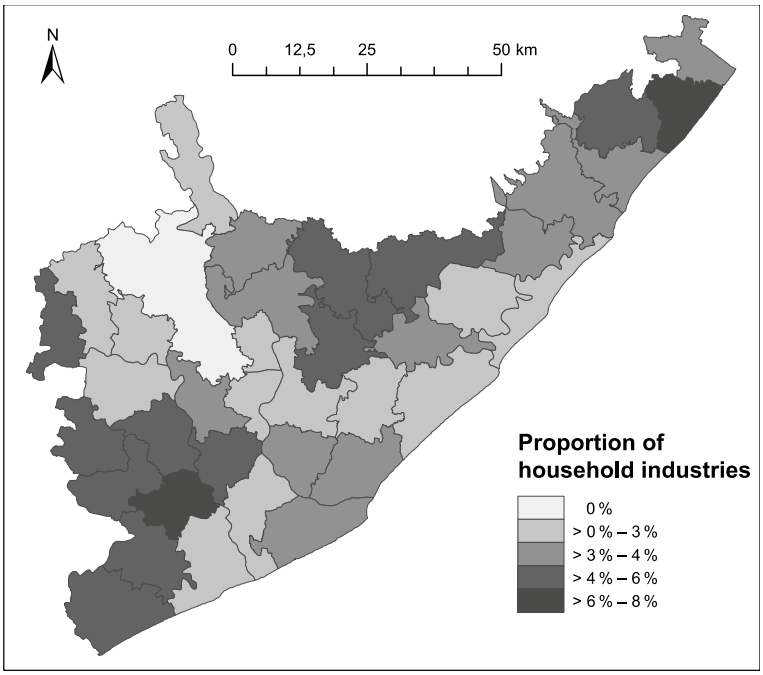



**Fig. 4:** Proportion of household industries

## Discussion

The main findings of this study were that (i) rural areas display higher risk towards AUF as compared to urban areas (ii) that AUF is unequally distributed across mandals in Srikakulam and (iii) that scheduled tribes are the main population at risk

and household industries and proximity to forests are important socio-ecological risk factors for AUF.

**Higher risk of acute undifferentiated fever in rural areas**

Our results suggest that the risk of AUF is higher in rural areas as compared to urban areas. These results correspond to previous findings for infectious diseases resulting in fever and corresponds well to the current health situation in rural India (11). A nationally representative survey estimated the spatial distribution of the incidence of deaths attributable to malaria in India. 90% of estimated deaths attributable to malaria occurred in rural areas and displayed strong local variation (59). Dengue fever in contrast, changed over time from being predominantly urban in India to gaining a strong impact in rural areas, especially in areas with dense forest (60). A wide range of other diseases such as diarrhoeal diseases and diseases carried through the air are more common in rural areas than in urban areas such as typhoid and tuberculosis and are attributable to unclean water, exposure to unhealthy living conditions and poor nutrition (11). The high correlation between the proportion of AUF emergency calls to the total emergency demand and the proportion of fever within the population as indicated through the community level household survey (25) indicates that AUF emergency calls might be a realistic indicator to estimate the burden of infectious diseases within the population. The spatial inclusiveness of these data is not only likely to show a higher incidence of infectious diseases than surveillance data would suggest but provides additionally a more reliable foundation to analyse risk factors associated with AUF than the current data of the IDSP, which is suffering from unreliable and strong regional differences in reporting rates (23).

**Spatial inequalities of acute undifferentiated fever**

The disease mapping approach and the spatial scan statistic revealed that AUF displays strong local variation, both on mandal level as well as on village level. This variation at local level as well as the occurrence of local clusters corresponds well to previous findings analysing infectious diseases using the spatial scan statistic (61-63). However, the full potential of the spatial scan statistic was not employed here as we followed a purely spatial approach and did not search for spatio-temporal clusters. A prospective spatio-temporal cluster detection might provide an additional value for an

early-warning system based on EMS data to detect potential disease outbreaks as early as possible (29).

**Socio-ecological exposure factors for acute undifferentiated fever**

Based on the OLS regression, we found that proportion of scheduled tribes, proportion of household industries and proximity to forests were predictors of AUF and explained 66.2% of the spatial variation of risk towards AUF. AUF risk was strongly associated with the proportion of scheduled tribes and is therefore highly correlated to the most disadvantaged population group. These results correspond to other findings from the literature showing that low socio-economic status is an important predictor of elevated rates of infectious diseases in India (9, 64, 65), especially in rural areas (10, 11). Indigenous population groups belong to the poorest and most disadvantaged population groups in India and research on the health of this population group is often restricted to a sample of a specific indigenous population group (66). Indigenous people are living often close to forest areas and are disease prone as access to health services often is limited (67). Our results deliver statistical evidence for this relationship. Resulting interventions could be aimed directly at remote tribal populations to identify the underlying reasons that lead to a high vulnerability to infectious diseases. These reasons might consist of adverse distribution and poor treatment capacities of public primary healthcare facilities (68), lower willingness to attend public or private health care facilities due to high out of pocket costs and loss of productivity due to absence from work (69). The detection of spatial clusters might indicate areas for collecting blood samples to identify the underlying pathogens causing AUF (70). The significant association of AUF to scheduled tribes and household industries in turn might lead to initiatives such as the provision of insecticide treated bed nets (71) or indoor residual spraying (72).

**Limitations**

Our study has several limitations: The emergency calls with the main complaint AUF comprise a very broad category of potential underlying infectious diseases. This allows only an estimation of the impact of socio-economic and environmental determinants on the general incidence of certain infectious diseases but does not necessarily allow a first clue about the underlying disease itself. As shown in other studies, the complaint fever could be divided into more specific syndromes such as acute

encephalitis syndrome (AES) (4). Such an approach has shown to allow a detailed spatial analysis of landscape risk-factors associated with Japanese encephalitis (8) and could result in more detailed knowledge about contributing ecological factors. The use of emergency calls for this study limits the explanatory power for urban areas. Due to the higher availability of transportation as well as higher availability of medical infrastructure in urban areas, the use of emergency medical services may not be the first option to use. (73). This highlights the need to incorporate other data-sources as well. Urban Malaria is a major public health problem in India (64) and a strong contributor to the overall number of AUF cases in Srikakulam (25). Another limitation could be the knowledge of GVK EMRI`s 108 toll free emergency hotline. We were unable to verify if the service is equally popular within the district or if there are any notable spatial gaps of advertisement. Potentially, this could lead in areas with high advertisement to more frequent use and in areas with low advertisement to an under-utilization of this service. It would be interesting to compare the results of our analysis with results based on laboratory confirmed cases of infectious diseases to see whether our results differ widely from results conducted using laboratory confirmed cases. However, given the current scenario of disease surveillance in India, such a comparison is not possible (21). The number of explanatory variables available from the Census of India on mandal level was very limited. We would have favoured to include different age groups as additional explanatory variables to analyse which age group is most at risk. Additionally, other important variables such as bed-net use, housing materials, accessibility to health care providers and distance to water bodies as indicator for a potential vulnerability to vector-borne diseases (18) were not available for this study. The administrative units we used in this study are fairly large areas. Although we displayed the number of cases on village level using a kernel density estimation, we could not display the incidence rate or create a spatial regression model on this scale since the necessary census data were not available on village level during the time of the analysis. This limitation decreases the explanatory power of the kernel density estimation. Additionally, we would have favoured a Geographically Weighted Regression (GWR) to account for spatial heterogeneity. However, since our study area consisted only of 38 administrative units, the results would have been unreliable. Páez et al. point out that the use of GWR for small datasets with less than 160 administrative is not advisable (74). Current studies benefitting from the application of GWR usually focus on fairly large datasets (17, 18, 30, 75). This limitation underlines, that future research on risk factors should focus on

analysing AUF emergency calls on larger areas such as whole states to be able to capture spatial heterogeneity of socio-economic and environmental determinants within regression models. Such an approach could enhance the use of symptom-based data to explain the range of contributing factors to AUF.

## Conclusions

We used EMS data with the main complaint acute undifferentiated fever as indicator for infectious diseases and linked AUF to socio-ecological exposure factors.

Our results display that the spatial distribution of AUF follows closely the current scenario of infectious diseases in India as it reflects a higher vulnerability to fever in rural areas, spatial heterogeneity at local levels and a strong association with lower socio-economic status. This in turn highlights the value of AUF emergency calls to monitor the spatial distribution of infectious diseases in areas where reliable surveillance data are not available. Additionally, our approach shows that an epidemiological analysis at the ecological level using emergency call data could be used to identify main socio-ecological exposure factors and the main population at risk. These results might be relevant for future preparedness strategies and targeted, evidence-based public health interventions and provide additional information for syndromic surveillance. Our approach also stresses the importance and possibilities of including private medical institutions in surveillance activities. We hypothesize that our approach is useful not only for Srikakulam district, but also could be an effective way of guiding evidence-based public health interventions and future preparedness strategies in India where spatial EMS data are available.

## Acknowledgement

The authors would like to thank the German Federal Ministry of Education and Research (BMBF) and the Indian Council for Medical Research (ICMR) in the field of Public Health 2009-2011 (Grant Agreement No.: BMBF:IND08/005;ICMR:INDO/TRC/612/09-IHD) as well as the German Research Foundation (DFG) (GR 4302/1-1).

The authors would also like to thank Dr. Christoph Staubach for his valuable GIS modelling input and helpful discussions.

## References


1. Gupte MD, Ramachandran V, Mutatkar RK. Epidemiological profile of India: historical and contemporary perspectives. Journal of biosciences. 2001 Nov;26(4 Suppl):437-64. PubMed PMID: 11779959.
2. Greco A, Fester N, Engel AT, Kaufer BB. Role of the short telomeric repeat region in Marek's disease virus replication, genomic integration, and lymphomagenesis. Journal of virology. 2014 Dec;88(24):14138-47. PubMed PMID: 25275118. Pubmed Central PMCID: 4249155.
3. Reller ME, Bodinayake C, Nagahawatte A, Devasiri V, Kodikara-Arachichi W, Strouse JJ, et al. Unsuspected dengue and acute febrile illness in rural and semi-urban southern Sri Lanka. Emerging infectious diseases. 2012 Feb;18(2):256-63. PubMed PMID: 22304972. Pubmed Central PMCID: 3310451.
4. Joshi R, Colford JM, Jr., Reingold AL, Kalantri S. Nonmalarial acute undifferentiated fever in a rural hospital in central India: diagnostic uncertainty and overtreatment with antimalarial agents. The American journal of tropical medicine and hygiene. 2008 Mar;78(3):393-9. PubMed PMID: 18337332.
5. Gasem MH, Wagenaar JF, Goris MG, Adi MS, Isbandrio BB, Hartskeerl RA, et al. Murine typhus and leptospirosis as causes of acute undifferentiated fever, Indonesia. Emerging infectious diseases. 2009 Jun;15(6):975-7. PubMed PMID: 19523308. Pubmed Central PMCID: 2727336.
6. Abrahamsen SK, Haugen CN, Rupali P, Mathai D, Langeland N, Eide GE, et al. Fever in the tropics: aetiology and case-fatality - a prospective observational study in a tertiary care hospital in South India. BMC infectious diseases. 2013;13:355. PubMed PMID: 23899336. Pubmed Central PMCID: 3750507.
7. Chandy S, Yoshimatsu K, Boorugu HK, Chrispal A, Thomas K, Peedicayil A, et al. Acute febrile illness caused by hantavirus: serological and molecular evidence from India. Transactions of the Royal Society of Tropical Medicine and Hygiene. 2009 Apr;103(4):407-12. PubMed PMID: 19237179.
8. Robertson C, Pant DK, Joshi DD, Sharma M, Dahal M, Stephen C. Comparative spatial dynamics of Japanese encephalitis and acute encephalitis syndrome in Nepal. PloS one. 2013;8(7):e66168. PubMed PMID: 23894277. Pubmed Central PMCID: 3718805.
9. Gupta S, Shenoy VP, Mukhopadhyay C, Bairy I, Muralidharan S. Role of risk factors and socio-economic status in pulmonary tuberculosis: a search for the root cause in patients in a tertiary care hospital, South India. Tropical medicine & international health : TM & IH. 2011 Jan;16(1):74-8. PubMed PMID: 21091857.
10. Pascual Martinez F, Picado A, Roddy P, Palma P. Low castes have poor access to visceral leishmaniasis treatment in Bihar, India. Tropical medicine & international health : TM & IH. 2012 May;17(5):666-73. PubMed PMID: 22385129.
11. Patil AV, Somasundaram KV, Goyal RC. Current health scenario in rural India. The Australian journal of rural health. 2002 Apr;10(2):129-35. PubMed PMID: 12047509.
12. Kumar C, Singh PK, Rai RK, Singh L. Early neonatal mortality in India, 1990-2006. Journal of community health. 2013 Feb;38(1):120-30. PubMed PMID: 22797909.
13. Lauridsen J, Pradhan J. Socio-economic inequality of immunization coverage in India. Health economics review. 2011;1(1):11. PubMed PMID: 22828353. Pubmed Central PMCID: 3497714.
14. Shidhaye R, Patel V. Association of socio-economic, gender and health factors with common mental disorders in women: a population-based study of 5703

married rural women in India. International journal of epidemiology. 2010 Dec;39(6):1510-21. PubMed PMID: 21037247. Pubmed Central PMCID: 2992631.
15. Bharati P, Pal M, Bandyopadhyay M, Bhakta A, Chakraborty S, Bharati P. Prevalence and causes of low birth weight in India. Malaysian journal of nutrition. 2011 Dec;17(3):301-13. PubMed PMID: 22655452.
16. Raju S, Atkins, P.J., Kumar, N., Townsend, J.G., Corbridge, S., Duvvury, N., Evans, I., Harriss, B., Kumar, S., Oughton, E.A. Atlas of Women and Men in India. New Delhi, India: Zubaan Books; 1999.
17. Weisent J, Rohrbach B, Dunn JR, Odoi A. Socioeconomic determinants of geographic disparities in campylobacteriosis risk: a comparison of global and local modeling approaches. International journal of health geographics. 2012 Oct 13;11(1):45. PubMed PMID: 23061540. Pubmed Central PMCID: 3528622.
18. Haque U, Scott LM, Hashizume M, Fisher E, Haque R, Yamamoto T, et al. Modelling malaria treatment practices in Bangladesh using spatial statistics. Malaria journal. 2012 Mar 05;11:63. PubMed PMID: 22390636. Pubmed Central PMCID: 3350424.
19. Khormi HM, Kumar L. Modeling dengue fever risk based on socioeconomic parameters, nationality and age groups: GIS and remote sensing based case study. The Science of the total environment. 2011 Oct 15;409(22):4713-9. PubMed PMID: 21906782.
20. Aparajita C, Ramanakumar A. Burden of disease in rural India: an analysis through cause of death. The Internet Journal of Third World Medicine. 2005;2(2).
21. John TJ, Dandona L, Sharma VP, Kakkar M. Continuing challenge of infectious diseases in India. Lancet. 2011 Jan 15;377(9761):252-69. PubMed PMID: 21227500.
22. Kant L, Krishnan SK. Information and communication technology in disease surveillance, India: a case study. BMC public health. 2010;10(Suppl 1):S11.
23. Gaikwad A, Oruganti, R, Singh, V, Anchala, R, Rao, B.M., Ramswaropp. Reporting pattern in Integrated Disease Surveillance Project (IDSP) in Andhra Pradesh. Indian Emergency Journal. 2010;5(1):13-6.
24. Author. Details removed for peer review. 2011.
25. Jena B, Prasad M, Murthy S, Ramanarao G. Demand pattern for Medical Emergency Services for Infectious Diseases in Andhra Pradesh - A Geo-spatial Temporal Analysis of Fever Cases. Indian Emergency Journal. 2010;5(1):5 - 8.
26. Office E-NC. European Emergency Data-based Syndromic Surveillance System: European Emergency Data Group - Research Network Coordination Office; 2014 [cited 2014 11 March]. Available from: http://sidartha.eu/.
27. Cooper DL, Smith GE, Regan M, Large S, Groenewegen PP. Tracking the spatial diffusion of influenza and norovirus using telehealth data: a spatiotemporal analysis of syndromic data. BMC Med. 2008;6:16. PubMed PMID: 18582364. Pubmed Central PMCID: 2464582.
28. Horst MA, Coco AS. Observing the spread of common illnesses through a community: using Geographic Information Systems (GIS) for surveillance. Journal of the American Board of Family Medicine : JABFM. 2010 Jan-Feb;23(1):32-41. PubMed PMID: 20051540.
29. van den Wijngaard CC, van Asten L, van Pelt W, Doornbos G, Nagelkerke NJ, Donker GA, et al. Syndromic surveillance for local outbreaks of lower-respiratory infections: would it work? PloS one. 2010;5(4):e10406. PubMed PMID: 20454449. Pubmed Central PMCID: 2861591.
30. Hu M, Li Z, Wang J, Jia L, Liao Y, Lai S, et al. Determinants of the incidence of hand, foot and mouth disease in China using geographically weighted regression

models. PloS one. 2012;7(6):e38978. PubMed PMID: 22723913. Pubmed Central PMCID: 3377651.
31. Winskill P, Rowland M, Mtove G, Malima RC, Kirby MJ. Malaria risk factors in north-east Tanzania. Malaria journal. 2011;10:98. PubMed PMID: 21507217. Pubmed Central PMCID: 3094229.
32. Corner RJ, Dewan AM, Hashizume M. Modelling typhoid risk in Dhaka metropolitan area of Bangladesh: the role of socio-economic and environmental factors. International journal of health geographics. 2013 Mar 16;12:13. PubMed PMID: 23497202. Pubmed Central PMCID: 3610306.
33. Commissioner RGC. Census Data 2001 / Metadata New Delhi, India: Ministry of Home Affairs, Government of India; 2001 [cited 2014 13 March]. Available from: http://censusindia.gov.in/.
34. Reid HL, Haque U, Roy S, Islam N, Clements AC. Characterizing the spatial and temporal variation of malaria incidence in Bangladesh, 2007. Malaria journal. 2012;11:170. PubMed PMID: 22607348. Pubmed Central PMCID: 3465176.
35. Alzahrani AG, Al Mazroa MA, Alrabeah AM, Ibrahim AM, Mokdad AH, Memish ZA. Geographical distribution and spatio-temporal patterns of dengue cases in Jeddah Governorate from 2006-2008. Transactions of the Royal Society of Tropical Medicine and Hygiene. 2013 Jan;107(1):23-9. PubMed PMID: 23222946.
36. Haque U, Sunahara T, Hashizume M, Shields T, Yamamoto T, Haque R, et al. Malaria prevalence, risk factors and spatial distribution in a hilly forest area of Bangladesh. PloS one. 2011;6(4):e18908. PubMed PMID: 21533048. Pubmed Central PMCID: 3080915.
37. de Castro MC, Monte-Mor RL, Sawyer DO, Singer BH. Malaria risk on the Amazon frontier. Proceedings of the National Academy of Sciences of the United States of America. 2006 Feb 14;103(7):2452-7. PubMed PMID: 16461902. Pubmed Central PMCID: 1413719.
38. Geofabrik. India: Open Street Map; 2014 [cited 2014 22 February]. Available f rom: http://download.geofabrik.de/asia/india.html.
39. Mohindra KS, Haddad S, Narayana D. Women's health in a rural community in Kerala, India: do caste and socioeconomic position matter? Journal of epidemiology and community health. 2006 Dec;60(12):1020-6. PubMed PMID: 17108296. Pubmed Central PMCID: 2465509.
40. Kumar S, Quinn SC. Existing health inequalities in India: informing preparedness planning for an influenza pandemic. Health policy and planning. 2012 Sep;27(6):516-26. PubMed PMID: 22131367. Pubmed Central PMCID: 3529628.
41. Khan SM, Debnath C, Pramanik AK, Xiao L, Nozaki T, Ganguly S. Molecular evidence for zoonotic transmission of Giardia duodenalis among dairy farm workers in West Bengal, India. Veterinary parasitology. 2011 Jun 10;178(3-4):342-5. PubMed PMID: 21324592.
42. National Center for Environmental Health CDCAG. Investigating suspected cancer clusters and responding to community concerns: guidelines from CDC and the Council of State and Territorial Epidemiologists. MMWR Recommendations and reports : Morbidity and mortality weekly report Recommendations and reports / Centers for Disease Control. 2013 Sep 27;62(RR-08):1-24. PubMed PMID: 24067663.
43. Berke O. Exploratory spatial relative risk mapping. Preventive veterinary medicine. 2005;71(3):173-82.
44. Lawson AB, Biggeri AB, Boehning D, Lesaffre E, Viel JF, Clark A, et al. Disease mapping models: an empirical evaluation. Disease Mapping Collaborative Group.

Statistics in medicine. 2000 Sep 15-30;19(17-18):2217-41. PubMed PMID: 10960849.
45. Waller L, Gotway C. Applied spatial statistics for public health data. Hoboken, NJ: John Wiley and Sons, Inc.; 2004.
46. Bivand R. spdep: Spatial dependence: weighting schemes, statistics and models. R package version 0.5-71. 2014 [cited 2014 12 February].
47. Team RC. A language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing; 2013 [cited 2014 14 January]. Available from: http://www.r-project.org.
48. Kulldorff M. A Spatial Scan Statistic. Communications in statistics: theory and methods. 1997;26(6):1481 - 96.
49. Chen J, Roth RE, Naito AT, Lengerich EJ, Maceachren AM. Geovisual analytics to enhance spatial scan statistic interpretation: an analysis of U.S. cervical cancer mortality. International journal of health geographics. 2008;7:57. PubMed PMID: 18992163. Pubmed Central PMCID: 2596098.
50. Kulldorf M, Athas WF, Feuer EJ, Miller BA, Key CR. Evaluating Cluster Alarms: A Space-Time Scan Statistic and Brain Cancer in Los Alamos, New Mexico. American journal of public health. 1998;88(9):1377-80.
51. Kulldorff M. SaTScanTM User Guide for version 9.2 2013 [cited 2013 8 September].
52. Smith SC, Bruce CW. CrimeStat III User Workbook Washington, DC: The National Institute of Justice; 2008 [cited 2014 12 March]. Available from: https://http://www.icpsr.umich.edu/CrimeStat/workbook/CrimeStat_Workbook.pdf.
53. Levine N. Crime mapping and the Crimestat program. Geographical analysis. 2006;38(1):41-56.
54. Anselin L. Exploring Spatial Data with GeoDaTM : A Workbook Urbana, Illinois, USA: Spatial Analysis Laboratory, Department of Geography, University of Illinois at Urbana-Champaign; 2005 [cited 2014 2 March]. Available from: https://geodacenter.asu.edu/system/files/geodaworkbook.pdf.
55. ESRI. Regression analysis basics 2009 [cited 2014 11 March]. Available from: http://webhelp.esri.com/arcgisdesktop/9.3/index.cfm?TopicName=Regression_analysis_basics.
56. Zhou Y, Hallisey EJ, Freymann GR. Identifying perinatal risk factors for infant maltreatment: an ecological approach. International journal of health geographics. 2006;5:53. PubMed PMID: 17144919. Pubmed Central PMCID: 1698478.
57. ESRI. How Exploratory Regression works 2013 [cited 2014 February 6th]. Available from: http://resources.arcgis.com/en/help/main/10.1/index.html - //005p00000054000000.
58. Anselin L, Syabri I, Kho Y. GeoDa: an introduction to spatial data analysis. Geographical analysis. 2006;38(1):5-22.
59. Dhingra N, Jha P, Sharma VP, Cohen AA, Jotkar RM, Rodriguez PS, et al. Adult and child malaria mortality in India: a nationally representative mortality survey. Lancet. 2010 Nov 20;376(9754):1768-74. PubMed PMID: 20970179. Pubmed Central PMCID: 3021416.
60. Gupta B, Reddy BP. Fight against dengue in India: progresses and challenges. Parasitology research. 2013 Apr;112(4):1367-78. PubMed PMID: 23455936.

61. Haque U, Huda M, Hossain A, Ahmed SM, Moniruzzaman M, Haque R. Spatial malaria epidemiology in Bangladeshi highlands. Malaria journal. 2009;8:185. PubMed PMID: 19653914. Pubmed Central PMCID: 2732922.
62. Toan do TT, Hu W, Quang Thai P, Hoat LN, Wright P, Martens P. Hot spot detection and spatio-temporal dispersion of dengue fever in Hanoi, Vietnam. Global health action. 2013;6:18632. PubMed PMID: 23364076. Pubmed Central PMCID: 3556563.
63. Liebman KA, Stoddard ST, Morrison AC, Rocha C, Minnick S, Sihuincha M, et al. Spatial dimensions of dengue virus transmission across interepidemic and epidemic periods in Iquitos, Peru (1999-2003). PLoS neglected tropical diseases. 2012;6(2):e1472. PubMed PMID: 22363822. Pubmed Central PMCID: 3283551.
64. Kumar DS, Andimuthu R, Rajan R, Venkatesan MS. Spatial trend, environmental and socioeconomic factors associated with malaria prevalence in Chennai. Malaria journal. 2014;13:14. PubMed PMID: 24400592. Pubmed Central PMCID: 3893554.
65. Sur D, von Seidlein L, Manna B, Dutta S, Deb AK, Sarkar BL, et al. The malaria and typhoid fever burden in the slums of Kolkata, India: data from a prospective community-based study. Transactions of the Royal Society of Tropical Medicine and Hygiene. 2006 Aug;100(8):725-33. PubMed PMID: 16455118.
66. Subramanian SV, Davey Smith G, Subramanyam M. Indigenous health and socioeconomic status in India. PLoS medicine. 2006 Oct;3(10):e421. PubMed PMID: 17076556. Pubmed Central PMCID: 1621109.
67. Balgir R, editor Tribal health problems, disease burden and ameliorative challenges in tribal communities with special emphasis on tribes of Orissa. Proceedings of National Symposium on "Tribal Health" 19th-20th October; 2006.
68. Duggal R. Public health expenditures, investment and financing under the shadow of a growing private sector. Review of Healthcare in India, CEHAT, Mumbai. 2005.
69. Nayar K. Social exclusion, caste & health: a review based on the social determinants framework. Indian Journal of Medical Research. 2007;126(4):355.
70. Phuong HL, de Vries PJ, Nga TT, Giao PT, Hung LQ, Binh TQ, et al. Dengue as a c ause of acute undifferentiated fever in Vietnam. BMC infectious diseases. 2006;6(1):123.
71. Lengeler C. Insecticide-treated bed nets and curtains for preventing malaria. The Cochrane database of systematic reviews. 2004;2(2).
72. Bousema T, Stevenson J, Baidjoe A, Stresman G, Griffin JT, Kleinschmidt I, et al. The impact of hotspot-targeted interventions on malaria transmission: study protocol for a cluster-randomized controlled trial. Trials. 2013;14:36. PubMed PMID: 23374910. Pubmed Central PMCID: 3576332.
73. Saddichha S, Saxena MK, Vibha P, Methuku M. Neurological emergencies in India--lessons learnt and strategies to improve outcomes. Neuroepidemiology. 2009;33(3):280-5. PubMed PMID: 19696519.
74. Páez A, Farber S, Wheeler D. A simulation-based study of geographically weighted regression as a method for investigating spatially varying relationships. Environment and Planning-Part A. 2011;43(12):2992.
75. Tsai PJ. Scrub typhus and comparisons of four main ethnic communities in taiwan in 2004 versus 2008 using geographically weighted regression. Global journal of health science. 2013 May;5(3):101-14. PubMed PMID: 23618480.

# CHAPTER 3

# Case study on Hepatitis C in the Netherlands

published as:

Kauhl, B., Heil, J., Hoebe, C. J., Schweikart, J., Krafft, T., & Dukers-Muijrers, N. H. (2015). The spatial distribution of hepatitis C virus infections and associated determinants—An application of a geographically weighted poisson regression for evidence-based screening interventions in hotspots. *PloS one*, *10*(9), e0135656.

**Abstract**

**Background:** Hepatitis C Virus (HCV) infections are a major cause for liver diseases. A large proportion of these infections remain hidden to care due to its mostly asymptomatic nature. Population-based screening and screening targeted on behavioural risk groups had not proven to be effective in revealing these hidden infections. Therefore, more practically applicable approaches to target screenings are necessary. Geographic Information Systems (GIS) and spatial epidemiological methods may provide a more feasible basis for screening interventions through the identification of hotspots as well as demographic and socio-economic determinants.

**Methods:** Analysed data included all HCV tests (n=23,800) performed in the southern area of the Netherlands between 2002-2008. HCV positivity was defined as a positive immunoblot or polymerase chain reaction test. Population data were matched to the geocoded HCV test data. The spatial scan statistic was applied to detect areas with elevated HCV risk. We applied global regression models to determine associations between population-based determinants and HCV risk. Geographically weighted Poisson regression models were then constructed to determine local differences of the association between HCV risk and population-based determinants.

**Results:** HCV prevalence varied geographically and clustered in urban areas. The main population at risk were middle-aged males, non-western immigrants and divorced persons. Socio-economic determinants consisted of one-person households, persons with low income and mean property value. However, the association between HCV risk and demographic as well as socio-economic determinants displayed strong regional and intra-urban differences.

**Discussion:** The detection of local hotspots in our study may serve as a basis for prioritization of areas for future targeted interventions. Demographic and socio-economic determinants associated with HCV risk show regional differences underlining that a one-size-fits-all approach even within small geographic areas may not be appropriate. Future screening interventions need to consider the spatially varying association between HCV risk and associated demographic and socio-economic determinants.

## Introduction

Hepatitis C virus (HCV) infections are a major cause of liver diseases and are the leading cause for liver cirrhosis worldwide (1). The World Health Organization estimates that 123 million people globally are infected with HCV (2). A major challenge for public health response to HCV is its mostly asymptomatic nature and therefore the limited number of HCV positive individuals aware of their HCV status. Infected, but undiagnosed persons are an important source for further transmission (3). Several studies estimated the proportion of asymptomatic infections to account for 70% (4, 5) to 90% (6) of acute infections, leading to only a small proportion of infected individuals seeking medical attention for symptoms related to HCV infection (7). It is estimated that less than one-third of HCV infected individuals are aware of their HCV status (8-10). Many infections are either undetected or are detected at a late stage. Highly effective therapeutic options for HCV are becoming available, (11, 12) but logically only to persons, who's HCV infection is diagnosed.

To provide an opportunity for treatment of HCV positive persons, which are yet undiagnosed and therefore currently hidden to care, preventive screening is necessary.
The HCV prevalence and its associated risk factors varies considerable between countries (13, 14). Past interventions focused on the population in general were not very cost-effective, especially in countries where the overall HCV prevalence is low. In the Netherlands, the HCV prevalence in the Dutch adult population is estimated to be relatively low with 0.2%, although estimates vary between 0.1 and 0.4%, depending largely on the study design and population studied (15, 16). A meta analysis on the effectiveness of screening interventions suggests that for low HCV prevalence populations, pre-screening selection criteria should be used to increase efficiency (17). The World Health Organisation (WHO) recommends in its new guidelines to offer HCV tests to people with high risk behaviour and to people from high risk populations (18). These target populations include transmission risk groups such as injecting drug users (IDU) (5, 10, 11), blood transfusion recipients (3), surgery and dialysis patients (13), professionals in patient care (5), immigrants from endemic countries (13), persons with low socio-economic status (5) and HIV positive men who have sex with men (MSM) (3). However, screening approaches to target these risk groups have not been shown to be effective in revealing the totality of hidden cases as the identification of people who belong to such risk groups in the first place appeared to be quite challenging. Furthermore, in the Netherlands it had been shown that a substantial part (25%) of all

HCV infections is not attributable to any of the aforementioned risk groups and is therefore not included in screening interventions targeted at risk groups (16). Although the prevalence of HCV in the US is higher with an estimated 2% (19), the Center for Disease Control (CDC) similar to the WHO advises screening of persons in risk groups (IDU, blood transfusion or organ transplant recipients before July 1992, health care personnel with history of exposure and born to an HCV-positive mother) (20). However, these criteria appeared also in the US difficult to include in practical screening interventions (10). As a result, future screening interventions need to find characteristics of HCV that are more practically applicable than the risk groups and behavioural factors outlined above.

Other relevant factors than behavioural and demographic risk factors associated with HCV are socio-economic characteristics. As for many infectious diseases, including HCV, lower socio-economic status tends to be associated with higher prevalence (1, 13, 21, 22). The identification of socio-economic determinants provides a more practically applicable basis for screening interventions (10), as population characteristics are typically available within population data (23). The application of Geographic Information Systems (GIS) is essential to display the spatial heterogeneity of disease risk and to quantify the impact of socio-economic determinants on the incidence of infectious diseases (22, 24).

Exploratory disease mapping and local cluster tests have shown to help identifying areas with statistically significant high risks (often referred to as hotspots or clusters) for prioritizing future interventions for Hepatitis C in the mainland of China (25) as well as many other infectious diseases including HIV (26), *Chlamydia trachomatis* and *Neisseria gonorrhea* (27).

The increasing availability of a wide range of population-based variables allows a detailed analysis of demographic and socio-economic determinants of disease risk using spatial regression models at the ecological level (24, 28, 29).

With respect to HCV, it has been shown that not only prevalence varies between and within countries, but also the association between risk factors and HCV prevalence (13), highlighting the necessity to account for local variation in spatial regression models for HCV.

In settings where strong local variation of the association between disease risk and possible determinants can be expected, geographically weighted Poisson regression models (GWPR) have proven to be very effective to measure the spatially varying

association between possible determinants and disease risk. This in turn, often led to the conclusion that the determinants of a specific disease depend largely where infected populations live, allowing public health preventions to be targeted directly at those population groups, that are most at risk in a specific location (30-32).

The aim of this paper is therefore to (i) determine hotspots for future screening interventions using the spatial scan statistic and (ii) to assess demographic and socio-economic determinants of HCV risk within these hotspots using GWPR to facilitate targeted, evidence-based screening interventions aimed directly at risk-groups.

## Data and Methods

### Ethics Statement

The medical ethics committee of the Maastricht University Medical Centre (Maastricht, the Netherlands) approved the study (11-4-136) and waived the need for consent to be collected from participants. Since retrospective data originated from standard care (in which one can opt-out for the use of their data for scientific research) and were analyzed anonymously, no further informed consent for data analysis was obtained.

### Dependent Variable

The dependent variable consisted of the HCV diagnoses that were made in the southern part of the province Limburg, the Netherlands between January 1st, 2002 and December 31st, 2008, comprising an adult population of 500,955 in 2008 (10, 33). The diagnoses were retrieved from HCV test data that were provided by three hospital laboratories, which perform tests on HCV upon request of nearly all care providers serving the area. In total 23,800 HCV tests were conducted of which 823 unique patients were tested positive. According to screening procedures in the Netherlands, HCV antibodies were detected with an ELISA. Confirmation was performed with an immunoblot and/or polymerase chain reaction (PCR). When an acute infection was suspected or when the patient was HIV positive or on hemodialysis, only PCR was used for screening. In the current study, we defined a positive confirmation test or PCR as a positive case. Of these 823 unique positive individuals, 781 had valid postal codes assigned and were included in the analysis. Next to postal code and HCV test result, the laboratory dataset included sex and age (10).

**Explanatory Variables**

We assessed several demographic and socio-economic variables for their association with HCV risk. The data for these variables were downloaded from the Central Bureau for Statistics Netherlands. In this study, we used data and map sources from the Statline database 2009 (33) (Table 1). The data were available on neighbourhood level and had to be matched to the four-digits postal codes of the HCV data. A neighbourhood is a part of a municipality with a homogenous socio-economic structure (33, 34). Due to privacy restrictions, socio-economic data on neighbourhood level is only available for neighbourhoods with more than 50 persons, 200 persons, 10 households and 70 households, depending on the respective variable (33). We therefore aggregated to the four-digits postal codes based on those neighbourhoods, for which socio-economic data was made available.

Demographic variables included stratified population data for 2012 on four-digits postal code level (10) (16). The population data was extracted from customised data by Statistics Netherlands (Extraction date: 20/02/2013).

Socio-economic variables included marital status (proportion of residents that were married, unmarried, divorced, or widowed)(35), proportion of non-western immigrants (16), proportion of one-person households, proportion of households without children, average income (10, 36), proportion of persons having low income (36) (defined as an income below 19,200 Euro per year (33)), households having low purchasing power (defined as households having less than 9,250 Euro available per year (33)), households having low income (36) (defined as households with an annual income below 25,100 Euro (33)), households below social minimum and mean property value as indicator for potential area deprivation (10, 33).

**Table 1:** Explanatory variables.

| Variable | Average | Min | Max |
|---|---|---|---|
| Married (%) | 44.7 | 1.0 | 62.0 |
| Unmarried (%) | 40.9 | 96.0 | 15.0 |
| Divorced (%) | 3.8 | 0.0 | 10.6 |
| Widowed (%) | 6.8 | 0.0 | 48.0 |
| Non-western immigrants (%) | 4.8 | 0.0 | 14.2 |

| | | | |
|---|---|---|---|
| One-person households (%) | 36.4 | 7.0 | 71.8 |
| Households w/o children (%) | 32.1 | 20.7 | 57.0 |
| Social welfare recipients (%) | 1.2 | 0.5 | 14.6 |
| Average income (in 1000 Euro) | 20.5 | 14.2 | 29.1 |
| Persons with low income (%) | 2.1 | 0.5 | 15.5 |
| HH with low purch. power (%) | 8.9 | 0.0 | 20.0 |
| Households with low income (%) | 45.9 | 17.0 | 67.0 |
| HH below social minimum (%) | 9.3 | 0.0 | 22.0 |
| Mean prop. value (in 1000 Euro) | 213.6 | 116.7 | 433 |
| Males aged 36 – 45 (%) | 6.4 | 0.0 | 12.2 |
| Males aged 46 – 55 (%) | 8.2 | 0.0 | 11.2 |

**Exploratory Disease Mapping**

We calculated the prevalence rate of HCV and the relative risk (RR) for the adult population aged between 16 and 65.

The RR estimates provide useful information how common HCV infection in a specific location is as compared to the global baseline (37). We additionally applied spatial empirical Bayes smoothing since the population at risk displayed strong regional variation. This leads to a large variance of the prevalence rate and the relative risk especially in areas where the underlying population is small (38). Due to strong regional variation in the HCV prevalence, we applied a local smoothing approach. The prevalence rates and the RR were therefore smoothed towards a local mean where the neighbours were defined as areas sharing a common edge and a common boundary (39). The calculation of the spatial empirical Bayes smoothing was carried out using OpenGeoDa 1.2.0 (40) and the results were then imported in ESRI ArcGIS 10.1.

**Global Cluster Detection**

To test whether there is spatial autocorrelation of the HCV prevalence, we used Moran`s I. Moran`s I is a widely used global cluster test, which determines the degree of clustering or dispersion within a data set. The resulting values may range from 1 (perfect correlation), 0 (complete spatial randomness) to -1 (perfect dispersed) (41). For the HCV data, a positive spatial autocorrelation means that postal code areas with

high HCV prevalence are close to other postal code areas with high HCV prevalence. For this study, we defined adjacency as postal code areas sharing a common edge or corner. The presence of global clustering justified the subsequent local cluster analysis. The computation of Moran`s I was carried out in OpenGeoDa 1.2.0 (40)

**Local Cluster Detection**

The spatial scan statistic has been widely applied in several spatial-epidemiological studies to detect local clusters with statistically significant elevated risk of infectious diseases (22, 26, 42, 43). The spatial scan statistic is a local cluster test, which identifies the location and the statistical significance of local clusters (26). We applied a Poisson purely spatial model where the number of HCV cases follows an inhomogeneous Poisson process (44). The input data for this model consisted of the number of positive individuals per postal code, the number of adults aged between 16 and 65 and the centroid coordinates for each area. The spatial scan statistic imposes a circular scanning window, which is flexibly in size and position and gradually moves over all coordinates, evaluating all potential cluster locations and sizes up to either a user-defined maximum radius, a user defined maximum percentage of the population at risk or the default value of up to 50% of the population at risk (45).

In our study, the purpose of the spatial scan statistic was to detect areas with significantly elevated risk of diagnosed HCV, which can serve as a basis for the prioritization of future screening interventions (46, 47). We set the maximum population at risk to not exceed 5% of the adult population. This was done to detect local clusters as precisely as possible since the default settings of 50% of the population at risk are more likely to produce clusters of no practical use (48). The computation was carried out using the SaTScan software version 9.2 (45).

**Spatial Regression**

**Ordinary Least Squares Regression**

To specify a meaningful geographically weighted Poisson regression model, we conducted several steps: First, we performed a natural log-transformation of the dependent variable. We then used a data-mining tool called Exploratory Regression in ESRI ArcGIS 10.1. to determine potential candidate explanatory variables. This tool evaluates all possible variable combinations that form a properly specified ordinary

least squares (OLS) regression model. Exploratory regression is comparable to a step-wise regression (31). However, it evaluates all possible variable combinations based on following criteria: (i) the coefficients are statistically significant, (ii): the explanatory variables are free from multicollinearity, (iii): the residuals are normally distributed and (iv): the residuals do not display spatial autocorrelation (31, 49, 50).

Based on the results of the exploratory regression, we determined overall model significance, the presence of heteroscedasticity and a wide range of diagnostics by creating an OLS regression model in OpenGeoDa 1.2.0 (40) with the same dependent and explanatory variables as suggested by the exploratory regression.

**Geographically Weighted Regression**

Since the OLS regression is a global regression model, it estimates the strength of the relationship between the dependent variable and the explanatory variables averaged over the whole study area. However, the larger the study area, the more unlikely it is that one single coefficient per explanatory variable reflects the true underlying spatial relationship between the dependent variable and the explanatory variable since spatial data tend to vary over space. Global statistics tend to lead to the conclusion that relationships between variables are equal across the entire study area whereas local statistics can show the falsity of this assumption by displaying how the relationships vary across space (51). The geographically weighted regression (GWR) method is therefore an extension to the traditional standard regression methodology and estimates a wide range of local parameters and diagnostics.

The Poisson distribution within the GWR framework is currently the most suitable for disease data, especially if observed counts of cases are low in specific areas (52-54). The dependent variable was specified within the geographically weighted Poisson regression (GWPR) as the observed number of HCV cases per postal code and the offset variable was specified as the number of adult persons per postal code. The GWPR model calculates an additional global Poisson regression model, which can be compared to the results of the global OLS model to test the hypothesis that a Poisson regression is more suitable for HCV than the traditional OLS regression. The explanatory variables for the global and local Poisson regression models were the same variables that were found to be significant as specified by the OLS model. The centroids of each postal code were used as input coordinates. The GWPR model then uses a kernel and fits for each coordinate a regression equation where the coordinate in the centre of the

kernel is the regression point. The data points inside the kernel are weighted from the centre of the kernel towards the edge of the kernel. Data points outside the kernel receive a weight of zero and are not included in the regression equation. For each coordinate, the data points are weighted differently so that each regression point has a unique regression equation. We used an adaptive kernel size so that in rural areas where data points are sparse, the kernel bandwidth will increase in size and will decrease in urban areas where data points are plentiful. The size of the bandwidth for each kernel and regression point is optimized using Akaike`s Information Criterion (AIC) (51). To facilitate interpretation of the regression coefficients of the GWPR, the coefficients were exponentiated to show an increase or decrease of the relative risk of the dependent variable per one-unit change in the respective explanatory variable (52). Statistical significance for each coefficient per postal code was calculated using pseudo t-values (51). The statistic behind the GWPR method is described in detail elsewhere (52). The computation of the GWPR was carried out using the GWR4 software (55).

## Results

### Spatial Distribution of Hepatitis C Prevalence Among Adults

The prevalence and the risk estimates between the postal code areas varied widely, ranging from 0 to 1.02% of the adult population per postal code. The overall prevalence rate among adults was 0.19% of the total adult population. There was a clear urban-rural divide within the study area. Areas with higher risks were strongly concentrated within the urban areas of Heerlen, Maastricht and to a lower extent in Sittard-Geleen (Fig. 1). Moran`s I revealed significant positive global autocorrelation of the HCV prevalence (Moran`s I = 0.43, $p<0.001$), indicating that postal codes with higher risks are close to each other.

The spatial scan statistic detected five significant local clusters (Fig. 1). These are postal codes with statistically significant elevated risk of diagnosed HCV. All clusters could be observed within the three urban areas of the study area (Table 2). In total, these clusters contain 268 (34%) of all observed HCV infections in the study area.

**Table 2:** Significant clusters with high HCV risk as determined by the spatial scan statistic.

| Cluster nr. | Location | RR | Cases | P-value |
|---|---|---|---|---|
| 1 | Southern part of Heerlen (3 postal codes) | 4.30 | 91 | <0.001 |
| 2 | Northern part of Heerlen (2 postal codes) | 2.83 | 60 | <0.001 |
| 3 | Northern part of Maastricht (1 postal code) | 4.03 | 31 | <0.001 |
| 4 | Centre of Maastricht (3 postal codes) | 1.91 | 71 | <0.001 |
| 5 | East. part of Sittard-Geleen (1 postal code) | 3.29 | 15 | <0.05 |

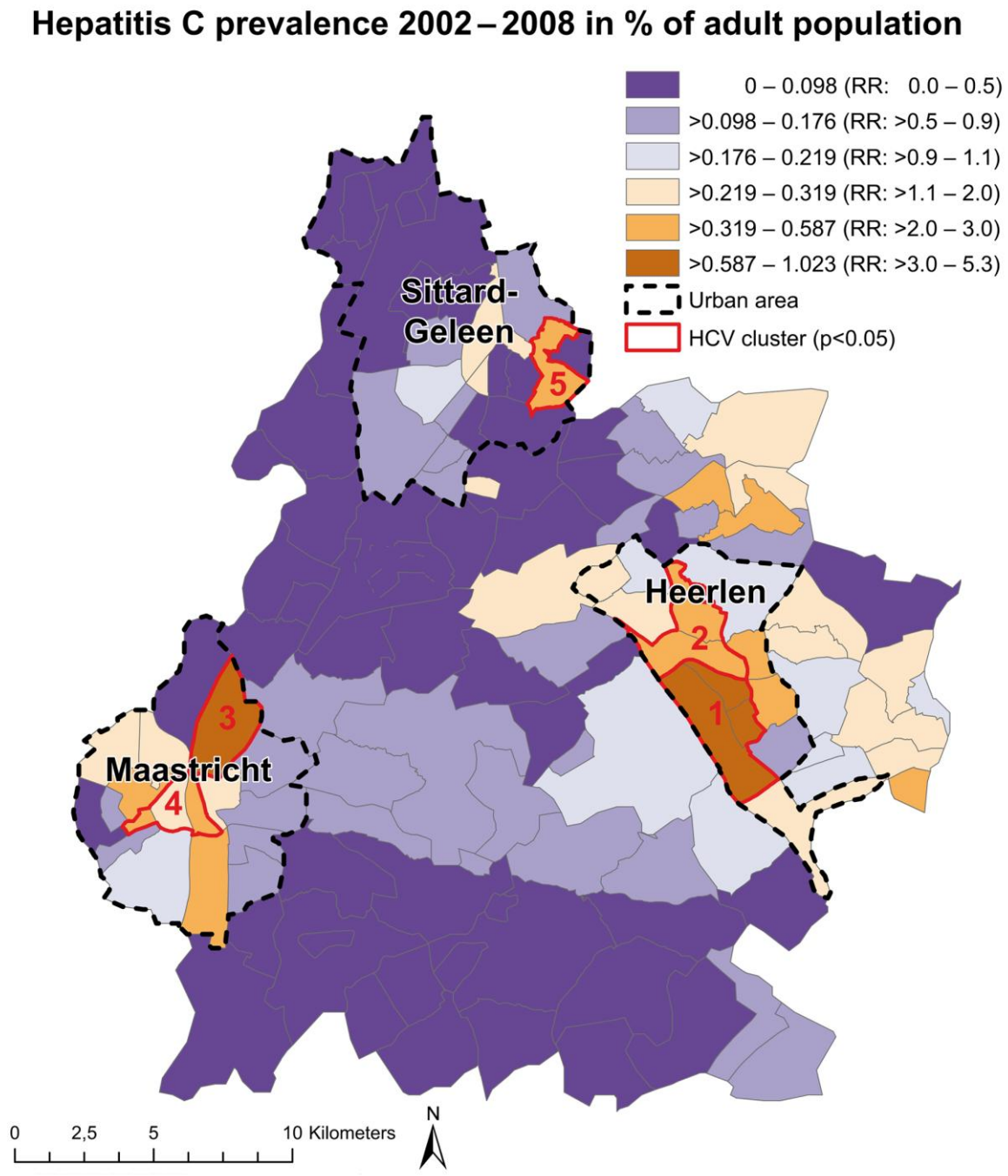



**Fig. 1:** Spatial Distribution of HCV prevalence and RR, 2002 – 2008.

**Demographic and Socio-economic Determinants of HCV**

As our analysis was exploratory in nature, we were interested in determining variable combinations based on population data that delivered a plausible explanation of HCV risk. We therefore identified two models that met all requirements for a properly specified OLS model that delivered a plausible explanation of the HCV prevalence. The AICc value of both OLS models differed only by 3, justifying a comparison of both models (56). The first model consisted of the following explanatory variables that were overall

positively associated with HCV risk: (i) proportion of divorced persons, (ii) proportion of one-person households, (iii) proportion of non-western immigrants and (iv) proportion of males aged 36 – 45. The second model consisted of the variables (i) average income per person, (ii) one-person households, (iii) mean property value and (iv) males aged 36 – 45. Variables for the second model were overall positively associated with HCV risk except for average income and mean property value, that both showed an inverse association. The same variables that were found to be significant in the OLS models were then used for further analysis in the global and local Poisson models.

By comparing model performance in terms of the goodness-of-fit AICc statistic (Table 3), the model with the lowest AICc value is the model with the best fit (24). Based on this criterion, for both Model 1 and Model 2 the AICc value suggests that the global Poisson regression had a better fit than the OLS regression. However, the local Poisson regression outperformed both global regression approaches. The local Poisson regression of model 2 was the overall best-fitting regression model in terms of the AICc value as well as the percentage of local deviance explained.

**Table 3:** Comparison of global and local models.

| Model | Local Deviance Explained | AICc |
|---|---|---|
| | **Model 1** | |
| OLS | 0.50 | 495 |
| Global Poisson | 0.47 | 372 |
| Local Poisson | 0.53 | 334 |
| | **Model 2** | |
| OLS | 0.50 | 498 |
| Global Poisson | 0.48 | 360 |
| Local Poisson | 0.55 | 323 |

**Results of the Geographically Weighted Poisson Regression**

**Model 1**

The results of the local Poisson model revealed strong local differences of the regression coefficients within the local clusters of elevated HCV risk (Table 4). The impact of the proportion of divorced persons on HCV risk was strongest in cluster 3 and 4 in Maastricht and cluster 5 in the northern part of Sittard-Geleen. In Heerlen, the impact of divorced persons was lowest (Fig. 2). The impact of one-person households displayed intra-urban differences as well as regional differences. The association between the proportion of one-person households and HCV risk was strongest in the northern part of Heerlen (cluster 2) and the southern part of Sittard-Geleen (cluster 5). In cluster 3 and 4 in Maastricht, the impact of one-person households was overall lower than in the other urban areas and clusters. However, the northern part of Maastricht displayed a stronger association of one-person households to HCV risk than the southern part (Fig. 2). The association between the proportion of non-western immigrants and HCV risk was only significant in cluster 3 and 4 in Maastricht and surrounding areas (Fig. 2). Also, the association between the proportion of males aged 36 – 45 years and HCV risk displayed large regional differences; its impact was only significant in cluster 5 in Sittard-Geleen, followed by clusters 3 and 4 in Maastricht and the rural areas in between (Fig. 2).

**Table 4:** Significant ($p<0.05$) coefficients per HCV cluster for model 1.

| HCV Cluster nr. | Determinants | Coefficient |
|---|---|---|
| 1 | Divorced persons | 1.102 |
| | One-person households | 1.034 |
| 2 | Divorced persons | 1.096 |
| | One-person households | 1.036 |
| 3 | Divorced persons | 1.161 |
| | One-person households | 1.035 |
| | Non-western immigrants | 1.036 |
| | Males aged 36 - 45 | 1.157 |
| 4 | Divorced persons | 1.164 |
| | One-person households | 1.034 |

| | | |
|---|---|---|
| | Non-western immigrants | 1.035 |
| | Males aged 36 - 45 | 1.146 |
| 5 | Divorced persons | 1.159 |
| | One-person households | 1.036 |
| | Males aged 36 - 45 | 1.161 |

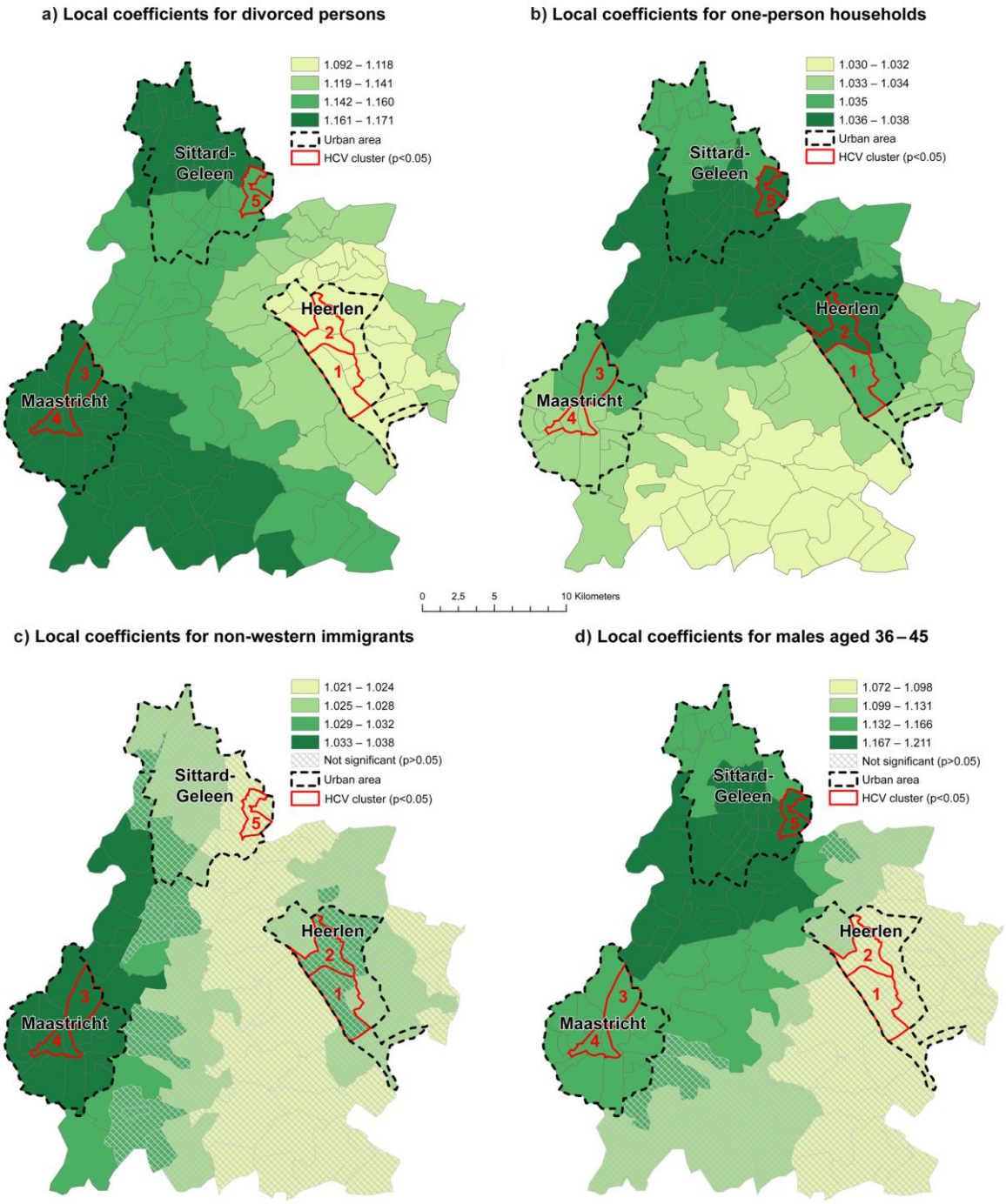



**Fig. 2:** Local coefficients for model 1.

**Model 2**

Comparable to the first model, the second model revealed strong local differences of the coefficients within the HCV clusters (Table 5). The association of HCV risk to average income was overall negative, indicating that a lower income is associated with a higher HCV risk. The local coefficients however, revealed that this association is not in the whole study area significant and negative. Average income is only significant inversely associated with HCV risk in cluster 5 in Sittard-Geleen and one postal code area in Maastricht (Fig. 3). The proportion of one-person households was positively associated with HCV risk in cluster 5 in Sittard-Geleen and the northern postal codes of Maastricht in cluster 3. This association decreased in strength towards cluster 1 and 2 in Heerlen (Fig. 3). Mean property value was negatively associated to HCV risk in all areas but the association displayed strong regional and intra-urban differences and was strongest in the southern postal codes of Heerlen in cluster 1 (Fig. 3). The association between the proportion of males aged 36-45 and HCV risk displayed a similar pattern as observed in model 1. The association was only significant in the northern parts of Maastricht in cluster 3 and 4, the southwestern parts of Sittard-Geleen in cluster 5 and areas in between (Fig. 3).

**Table 5:** Significant ($p<0.05$) coefficients per HCV cluster for model 2

| HCV Cluster nr. | Determinants | Coefficient |
|---|---|---|
| 1 | One-person households | 1.038 |
| | Mean property value | -1.009 |
| 2 | One-person households | 1.039 |
| | Mean property value | -1.008 |
| 3 | One-person households | 1.047 |
| | Mean property value | -1.006 |
| | Males aged 36 - 45 | 1.213 |
| 4 | One-person households | 1.045 |
| | Mean property value | -1.006 |
| | Males aged 36 - 45 | 1.186 |
| 5 | Average income per person | -1.070 |

| | |
|---|---|
| One-person households | 1.046 |
| Mean property value | -1.006 |
| Males aged 36 - 45 | 1.161 |

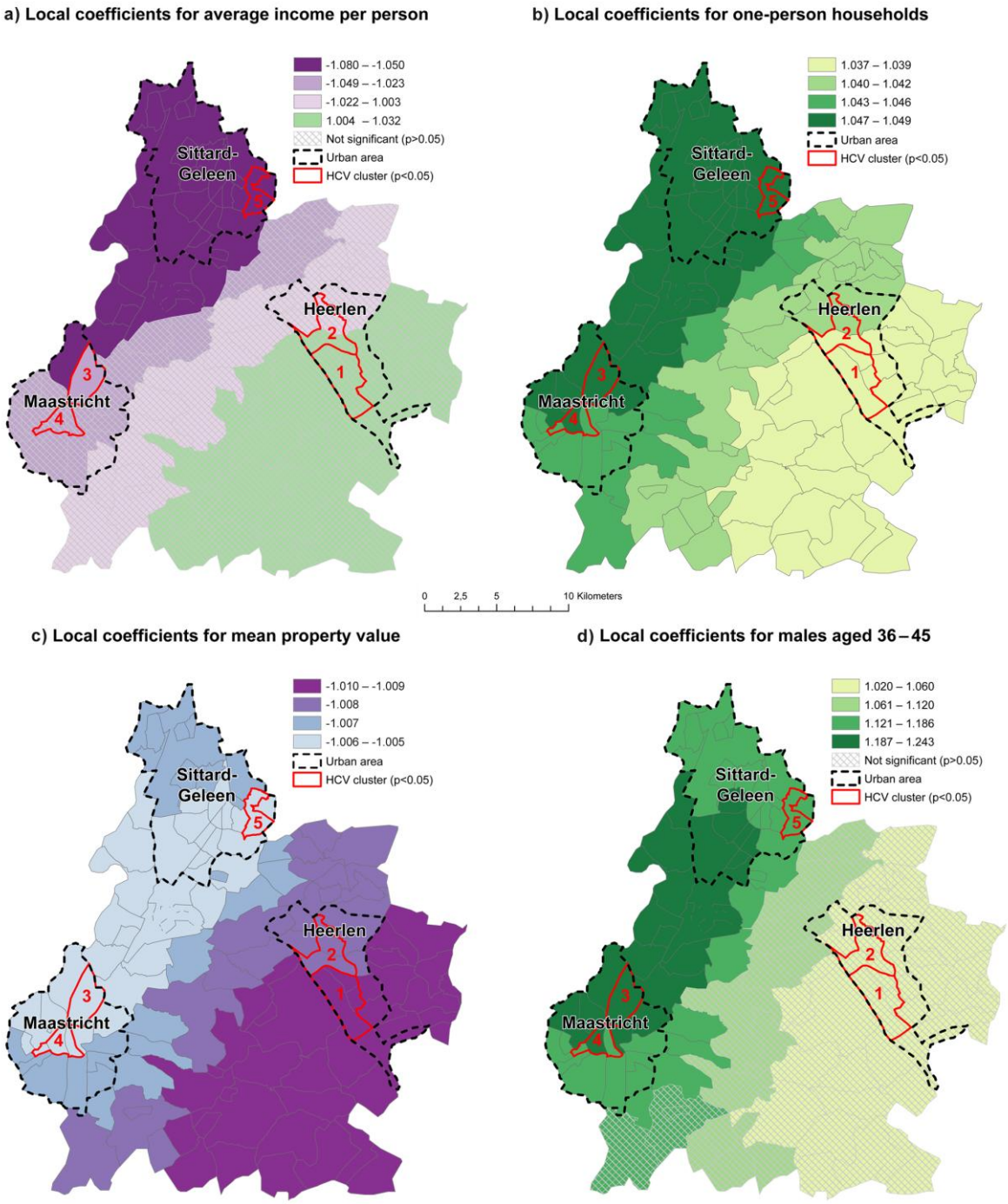



**Fig. 3:** Local coefficients for model 2.

**Discussion**

The prevalence of HCV varies geographically within the province of South Limburg and clusters were located in urban areas. The main population at risk were divorced persons, male residents aged 36 – 45 and non-western immigrants residing in the area. Socio-economic determinants associated with HCV risk included one-person households, low income at individual level and areas with low mean property value. The associations between these determinants and HCV risk displayed strong regional and intra-urban differences.

The overall prevalence of diagnosed HCV cases was 0.19%, which is in the range of previous overall estimations of the HCV prevalence within the Dutch population (7, 57). However, the prevalence showed strong local variations with prevalences ranging between 0 and 1.023%,

Five local clusters of significantly elevated HCV risk were detected. These clusters were located in the three urban areas in the region. These results suggest that HCV risk is higher in urban areas than in rural areas and clusters geographically. Thereby, HCV prevalence does not only vary between countries, as was noted before (13, 14) but also on small geographic scales such as postal code areas. The small-scale variation of HCV prevalence corresponds with findings of another spatial analysis of HCV in a higher prevalence country (58). Local clustering of HCV prevalence in urban areas is typical for a wide range of infectious diseases, including HIV (26), *Neisseria gonorrhoea* (42) and *Chlamydia trachomatis* (27). The detection of local clusters in our study may serve as a basis for prioritization of areas for future targeted and evidence-based screening interventions (26, 42). However, it should be noted that only a third of all HCV cases were detected in these clusters. The other cases showed a more random distribution over the region.

To what extent would these demographic and socio-economic determinants be of additional value to focus prevention strategies? When assuming that the population-based determinants represent the actual individual-based risk factors, then all determinants revealed here may indicate who are the key populations for HCV. Targeting these risk factors in the areas identified as clusters could serve as a practically applicable basis for prioritization of future screening interventions.

While there is a wide range of literature available about the prevalence of HCV infections and its associated risk factors (13, 14), only a local analysis as employed here may help to understand the patterns of HCV infections and its associations to socio-economic

determinants to effectively use available financial resources for targeted screening efforts.

The proportion of residents that were divorced was found to be associated with HCV risk over the complete study region. Marital status had been previously associated with HCV risk, yet findings were inconsistent (59-62). Being divorced could be a proxy for sexual and economic instability. The found association between HCV risk and divorced persons may therefore serve as basis for future research on the role of marital status and potential high-risk sexual behaviour on HCV transmission in the study area. Non-western immigrants were identified as ethnic risk group in our study. Although this association corresponds well to previous studies focusing on risk factors of HCV in the Netherlands (15, 16), the association of non-western immigrants to HCV risk was only significant in Maastricht. Potentially, in the other cities, immigrants from eastern-European countries might be more relevant as ethnic risk group (13, 15).

Males aged 36 – 45 were another main demographic risk group identified in our analysis confirming US findings (3). It is considered unlikely that this association can be for a large part explained by HIV positive MSM, as they comprise an important but only small part of the HCV cases in the Netherlands. (63). However, the association between males aged 36 – 45 and HCV risk was only significant in the western part of the study area. One-person households were identified as a risk factor relating to household size. Although this association to HCV may not be obvious at first, it is in line with our findings that divorced persons are an overall risk factor for HCV and could be a potential additional proxy for sexual and economic instability. This finding may additionally serve as a basis for future research on the role of one-persons households and HCV transmission. Mean property value and low income at personal level were important socio-economic determinants associated with HCV risk (35) and are in line with other studies showing that low socio-economic status is an important risk factor for HCV (10, 13, 36). However, our study demonstrated that low income at personal level was only significant in the urban area of Sittard-Geleen, while mean property value was found to be overall significant within the study area. Although this corresponds well to previous findings (10, 13, 36), it highlights the importance of including several markers for low socio-economic status on personal, household and area level to understand how these different measures of low socio-economic status impact the prevalence of HCV infections.

Several determinants were associated with HCV risk in the complete study region while others were only associated in certain regions; but all associations showed regional variance. The strong spatial differences observed suggest that the importance of demographic and socio-economic determinants to characterize the HCV key population may depend largely on the area where the HCV infected individual lives. Our findings are therefore in line with other studies applying GWR for infectious diseases (24, 30, 64).

In all clusters, an association was observed between HCV risk and divorced persons, one-person households and low mean property value. The proportion of middle-aged males were only associated to HCV in the clusters 3-5, and the proportion of non-western immigrants were only associated in the clusters 3 and 4. Income at personal level was only inversely associated in cluster 5. Thus, the impact of demographic and socio-economic determinants differed across the study area for the identified clusters.

**Limitations**

First, the spatial analysis of this study was based on the four-digits postal code areas of the Netherlands. Although this spatial aggregation may be considered as a fine geographic scale (34), the prevalence rate of HCV follows the potentially arbitrary administrative boundaries of these postal codes. The results of our analysis might differ if a different level of aggregation had been chosen. This problem is often referred to as the modifiable areal unit problem (MAUP) and has not only an impact on the spatial distribution of HCV risk and the location of the detected clusters, but also on the results of the ecological regression analysis (65). For our study, it would have been favourable to use street-level addresses of the HCV positive persons and underlying population at risk to analyse the spatial distribution of HCV without the limitation of arbitrary administrative boundaries (26). This would not only allow a precise localization of HCV clusters, but could offer the chance to perform a geographically weighted logistic regression to provide more detailed insights on the spatially varying association between HCV risk and associated socio-economic and demographic determinants (51). However, the HCV laboratory data as well as the population data used in this study were not available on this scale.

Second, it is unknown whether testing was motivated by the individuals due to symptoms related to HCV infection or was advised by a general practitioner due to prior

knowledge of potential exposure factors of the tested individual. It is also unknown whether geographical, demographic or socio-economic determinants may have been associated with access to testing services (e.g. by distance, lack of knowledge, illiteracy) hence may have influenced the observed associations. The tested persons might therefore differ from the general population. During the initial data analysis, we tested the association of tested persons to demographic or socio-economic population characteristics through an additional exploratory regression model with the log-transformed percentage of tested persons as dependent variable. However, the exploratory regression analysis could not find demographic or socio-economic population characteristics that delivered a properly specified OLS regression model.

Additionally, we compared the spatial pattern of the ratio of HCV positive persons to tested persons with the ratio of positive persons to the adult population. Both approaches displayed a similar spatial pattern. An additional cluster analysis using a Bernoulli model in SaTScan with the number of negative tested persons as controls (45) could be used to test whether the location of spatial clusters will change when using the negative tested persons as denominator. This might additionally indicate, whether testing is performed randomly or follows different spatial patterns that cannot yet be explained by population or demographic characteristics that were available for this study. However, we applied only a Poisson model as our goal was to compare the HCV prevalence within our study area to previous estimates of the HCV prevalence in the Netherlands, which would not be possible when applying a case-control study design.

In our study, we consider the geographical spread of diagnosed HCV as a realistic representation of the diagnosed HCV prevalence among the adult population since the proportion of tested persons could not be properly explained by demographic or socio-economic population characteristics and the two compared ratios displayed a similar spatial pattern.

Third, the demographic and socio-economic determinants examined are practically applicable but are hampered by lack of precision as they are based on population data and not on an individual level. Population data provide population characteristics per neighbourhood. Therefore, additional research is needed to study whether the population-based determinants for key populations actually capture the individuals comprising such key population.

Fourth, we previously estimated that up to 66% of all HCV-positive patients in the study region were hidden to current screening practices (10). As a result, cases that

were diagnosed may differ from the cases that were still hidden with respect to the variables studied here.

Given the limitations outlined above, it is unknown to what extent the clusters and the demographic and socio-economic determinants really reflect the hidden population. A proof-of-principle intervention targeting postal codes in a detected cluster is currently being set up to reveal whether the hidden HCV infected individuals are appropriately addressed by our detected clusters and determinants. Additionally, we may have missed associations of potential determinants not captured in our analyses, as these were unavailable in the population databases such as educational level.

Also, the population-based determinants used in this study were taken from the Statline database 2009 as this was the earliest population data to include socio-economic variables and the customized stratified demographic data on sex and age were only available for 2012 but not for the years between 2002 – 2008. Although this might influence the results, it is unlikely that this has a strong impact as the demographic composition in the Netherlands remained relatively stable within the last few years (66).

The application of a Geographically Weighted Poisson regression clearly demonstrated spatial variability of the coefficients and underlined that future screening interventions for HCV clearly have to take into account the spatially varying association between demographic and socio-economic determinants. However, Paéz et al. point out that the use of GWPR delivers more robust results when applied on large datasets containing more than 160 administrative units (67). Therefore, future research applying GWPR for HCV should focus on larger areas such as whole countries to gain more robust insights on the spatial variation of determinants for HCV (23, 29, 30). The reproducibility of our study would allow a similar analysis for the whole of the Netherlands.

**Conclusions**

In this study, we used spatial epidemiological methods to analyse the spatial distribution of HCV and its associated demographic and socio-economic determinants. Our results revealed strong regional differences not only of the HCV prevalence but also of the association between demographic and socio-economic determinants and HCV risk. These findings underline that a one-size-fits-all approach is not appropriate and that future screening interventions need to take into account the spatially varying

demographic and socio-economic determinants for HCV. Our approach may not only be useful for South-Limburg, but may be useful in other countries as well.

**Acknowledgements**

The authors acknowledge the three medical microbiology laboratories for providing the laboratory data: Dick van Dam and Monique Manders (Orbis Medical Centre, Sittard, the Netherlands), Frans Stals and Jos Bus (Atrium Medical Center Parkstad, Heerlen, The Netherlands), and Inge van Loo and Gert Grauls (Maastricht University Medical Centre, Maastricht, The Netherlands). The authors would also like to thank Jennifer Ilius for enhancing the maps using Adobe Illustrator.

## References


1. Shepard CW, Finelli L, Alter MJ. Global epidemiology of hepatitis C virus infection. The Lancet infectious diseases. 2005 Sep;5(9):558-67. PubMed PMID: 16122679.
2. Perz J, editor Estimated global prevalence of hepatitis C virus infection. 42nd Annual Meeting; 2004: Idsa.
3. Recommendations for prevention and control of hepatitis C virus (HCV) infection and HCV-related chronic disease. Centers for Disease Control and Prevention. MMWR Recommendations and reports : Morbidity and mortality weekly report Recommendations and reports / Centers for Disease Control. 1998 Oct 16;47(RR-19):1-39. PubMed PMID: 9790221.
4. Gebo KA, Bartlett JG. Management of hepatitis C: a review of the NIH Consensus Development Conference. The Hopkins HIV report : a bimonthly newsletter for healthcare providers / Johns Hopkins University AIDS Service. 2002 Sep;14(5):suppl i-iv. PubMed PMID: 12240644.
5. Alter MJ, Margolis HS, Bell BP, Bice SD, Buffington J, Chamberland M, et al. Recommendations for prevention and control of hepatitis C virus (HCV) infection and HCV-related chronic disease. MMWR Morbidity and mortality weekly report. 1998;47(1).
6. Desenclos J. The challenge of hepatitis C surveillance in Europe. Euro surveillance: bulletin Europeen sur les maladies transmissibles= European communicable disease bulletin. 2003;8(5):99-100.
7. Slavenburg S, Verduyn-Lunel F, Hermsen J, Melchers W, Te Morsche R, Drenth J. Prevalence of hepatitis C in the general population in the Netherlands. Neth J Med. 2008;66(1):13-7.
8. Culver DH, Alter MJ, Mullan RJ, Margolis HS. Evaluation of the effectiveness of targeted lookback for HCV infection in the United States—interim results. Transfusion. 2000;40(10):1176-81.
9. Singer ME, Younossi ZM. Cost effectiveness of screening for hepatitis C virus in asymptomatic, average-risk adults. The American journal of medicine. 2001;111(8):614-21.
10. Vermeiren AP, Dukers-Muijrers NH, van Loo IH, Stals F, van Dam DW, Ambergen T, et al. Identification of hidden key hepatitis C populations: an evaluation of screening practices using mixed epidemiological methods. PloS one. 2012;7(12):e51194.
11. Dore GJ, Matthews GV, Rockstroh J. Future of hepatitis C therapy: development of direct-acting antivirals. Current opinion in HIV and AIDS. 2011;6(6):508-13.
12. Ghany MG, Nelson DR, Strader DB, Thomas DL, Seeff LB. An update on treatment of genotype 1 chronic hepatitis C virus infection: 2011 practice guideline by the American Association for the Study of Liver Diseases. Hepatology. 2011;54(4):1433-44.
13. Cornberg M, Razavi HA, Alberti A, Bernasconi E, Buti M, Cooper C, et al. A systematic review of hepatitis C virus epidemiology in Europe, Canada and Israel. Liver International. 2011;31(s2):30-60.
14. Hahné SJ, Veldhuijzen IK, Wiessing L, Lim T-A, Salminen M, van de Laar M. Infection with hepatitis B and C virus in Europe: a systematic review of prevalence and cost-effectiveness of screening. BMC infectious diseases. 2013;13(1):181.
15. Vriend HJ, de Coul ELO, Van De Laar TJ, Urbanus AT, Van Der Klis FR, Boot HJ. Hepatitis C virus seroprevalence in The Netherlands. The European Journal of Public Health. 2012;22(6):819-21.

16. Vriend H, Van Veen M, Prins M, Urbanus A, Boot H, OP DE COUL E. Hepatitis C virus prevalence in The Netherlands: migrants account for most infections. Epidemiology and infection. 2013;141(06):1310-7.
17. Zuure FR, Urbanus AT, Langendam MW, Helsper CW, van den Berg CH, Davidovich U, et al. Outcomes of hepatitis C screening programs targeted at risk groups hidden in the general population: a systematic review. BMC public health. 2014;14:66. PubMed PMID: 24450797. Pubmed Central PMCID: 4016146. Epub 2014/01/24. eng.
18. Organization WH. Guidelines for the screening, care and treatment of persons with hepatitis C infection. 2014.
19. Smith BD, Yartel AK. Comparison of Hepatitis C Virus Testing Strategies: Birth Cohort Versus Elevated Alanine Aminotransferase Levels. American journal of preventive medicine. 2014;47(3):233-41.
20. Smith BD, Morgan RL, Beckett GA, Falck-Ytter Y, Holtzman D, Teo C-G, et al. Recommendations for the identification of chronic hepatitis C virus infection among persons born during 1945-1965. MMWR Recommendations and reports : Morbidity and mortality weekly report Recommendations and reports / Centers for Disease Control. 2012;61(RR-4):1-32.
21. Du P, Lemkin A, Kluhsman B, Chen J, Roth RE, MacEachren A, et al. The roles of social domains, behavioral risk, health care resources, and chlamydia in spatial clusters of US cervical cancer mortality: not all the clusters are the same. Cancer causes & control : CCC. 2010 Oct;21(10):1669-83. PubMed PMID: 20532608. Epub 2010/06/10. eng.
22. Kauhl B, Pilot E, Rao R, Gruebner O, Schweikart J, Krafft T. Estimating the spatial distribution of acute undifferentiated fever (AUF) and associated risk factors using emergency call data in India. A symptom-based approach for public health surveillance. Health & place. 2015;31:111-9.
23. Shoff C, Yang TC. Spatially varying predictors of teenage birth rates among counties in the United States. Demographic research. 2012 Sep 11;27(14):377-418. PubMed PMID: 23144587. Pubmed Central PMCID: 3493119.
24. Weisent J, Rohrbach B, Dunn JR, Odoi A. Socioeconomic determinants of geographic disparities in campylobacteriosis risk: a comparison of global and local modeling approaches. International journal of health geographics. 2012 Oct 13;11(1):45. PubMed PMID: 23061540. Pubmed Central PMCID: 3528622.
25. Wang L, Xing J, Chen F, Yan R, Ge L, Qin Q, et al. Spatial Analysis on Hepatitis C Virus Infection in Mainland China: From 2005 to 2011. PloS one. 2014;9(10):e110861.
26. Tanser F, Barnighausen T, Cooke GS, Newell ML. Localized spatial clustering of HIV infections in a widely disseminated rural South African epidemic. International journal of epidemiology. 2009 Aug;38(4):1008-16. PubMed PMID: 19261659. Pubmed Central PMCID: 2720393. Epub 2009/03/06. eng.
27. Bush KR, Henderson EA, Dunn J, Read RR, Singh A. Mapping the core: chlamydia and gonorrhea infections in Calgary, Alberta. Sexually transmitted diseases. 2008;35(3):291-7.
28. Zheng S, Cao CX, Cheng JQ, Wu YS, Xie X, Xu M. Epidemiological features of hand-foot-and-mouth disease in Shenzhen, China from 2008 to 2010. Epidemiology and infection. 2014 Aug;142(8):1751-62. PubMed PMID: 24139426.
29. Tsai PJ. Scrub typhus and comparisons of four main ethnic communities in taiwan in 2004 versus 2008 using geographically weighted regression. Global journal of health science. 2013 May;5(3):101-14. PubMed PMID: 23618480.

30. Hu M, Li Z, Wang J, Jia L, Liao Y, Lai S, et al. Determinants of the incidence of hand, foot and mouth disease in China using geographically weighted regression models. PloS one. 2012;7(6):e38978. PubMed PMID: 22723913. Pubmed Central PMCID: 3377651.
31. Haque U, Scott LM, Hashizume M, Fisher E, Haque R, Yamamoto T, et al. Modelling malaria treatment practices in Bangladesh using spatial statistics. Malaria journal. 2012 Mar 05;11:63. PubMed PMID: 22390636. Pubmed Central PMCID: 3350424.
32. Lin CH, Wen TH. Using geographically weighted regression (GWR) to explore spatial varying relationships of immature mosquitoes and human densities with the incidence of dengue. International journal of environmental research and public health. 2011 Jul;8(7):2798-815. PubMed PMID: 21845159. Pubmed Central PMCID: 3155330.
33. Netherlands S. Statline 2015 [cited 2015 March 6]. Available from: http://statline.cbs.nl/Statweb/.
34. Dijkstra A, Janssen F, De Bakker M, Bos J, Lub R, Van Wissen LJ, et al. Using spatial analysis to predict health care use at the local level: a case study of type 2 diabetes medication use and its association with demographic change and socioeconomic status. PloS one. 2013;8(8):e72730. PubMed PMID: 24023636. Pubmed Central PMCID: 3758350.
35. Garfein RS, Vlahov D, Galai N, Doherty MC, Nelson KE. Viral infections in short-term injection drug users: the prevalence of the hepatitis C, hepatitis B, human immunodeficiency, and human T-lymphotropic viruses. American journal of public health. 1996;86(5):655-61.
36. Meffre C, Le Strat Y, Delarocque-Astagneau E, Dubois F, Antona D, Lemasson JM, et al. Prevalence of hepatitis B and hepatitis C virus infections in France in 2004: social factors are important predictors after adjusting for known risk factors. Journal of medical virology. 2010 Apr;82(4):546-55. PubMed PMID: 20166185.
37. Berke O, Grosse Beilage E. Spatial relative risk mapping of pseudorabies-seropositive pig herds in an animal-dense region. Journal of veterinary medicine B, Infectious diseases and veterinary public health. 2003 Sep;50(7):173-82. PubMed PMID: 14535929.
38. Lawson AB, Biggeri AB, Boehning D, Lesaffre E, Viel JF, Clark A, et al. Disease mapping models: an empirical evaluation. Disease Mapping Collaborative Group. Statistics in medicine. 2000 Sep 15-30;19(17-18):2217-41. PubMed PMID: 10960849.
39. Waller L, Gotway C. Applied spatial statistics for public health data. Hoboken, NJ: John Wiley and Sons, Inc.; 2004.
40. Anselin L. Exploring Spatial Data with GeoDaTM : A Workbook Urbana, Illinois, USA: Spatial Analysis Laboratory, Department of Geography, University of Illinois at Urbana-Champaign; 2005 [cited 2014 2 March]. Available from: https://geodacenter.asu.edu/system/files/geodaworkbook.pdf.
41. Moran PA. Notes on continuous stochastic phenomena. Biometrika. 1950:17-23.
42. Jennings JM, Curriero FC, Celentano D, Ellen JM. Geographic identification of high gonorrhea transmission areas in Baltimore, Maryland. American journal of epidemiology. 2005;161(1):73-80.
43. Alencar CH, Ramos AN, Jr., dos Santos ES, Richter J, Heukelbach J. Clusters of leprosy transmission and of late diagnosis in a highly endemic area in Brazil: focus on different spatial analysis approaches. Tropical medicine & international health : TM & IH. 2012 Apr;17(4):518-25. PubMed PMID: 22248041. Epub 2012/01/18. eng.

44. Kulldorff M. A Spatial Scan Statistic. Communications in statistics: theory and methods. 1997;26(6):1481 - 96.
45. Kulldorff M. SaTScanTM User Guide for version 9.2 2013 [cited 2013 8 September].
46. Coleman M, Coleman M, Mabuza AM, Kok G, Coetzee M, Durrheim DN. Using the SaTScan method to detect local malaria clusters for guiding malaria control programmes. Malaria journal. 2009;8(68):10.1186.
47. Jones RC, Liberatore M, Fernandez JR, Gerber SI. Use of a prospective space-time scan statistic to prioritize shigellosis case investigations in an urban jurisdiction. Public health reports. 2006;121(2):133.
48. Chen J, Roth RE, Naito AT, Lengerich EJ, Maceachren AM. Geovisual analytics to enhance spatial scan statistic interpretation: an analysis of U.S. cervical cancer mortality. International journal of health geographics. 2008;7:57. PubMed PMID: 18992163. Pubmed Central PMCID: 2596098.
49. Poole MA, O'Farrell PN. The assumptions of the linear regression model. Transactions of the Institute of British Geographers. 1971:145-58.
50. ESRI. How Exploratory Regression works 2013 [cited 2014 February 6th]. Available from: http://resources.arcgis.com/en/help/main/10.1/index.html - //005p00000054000000.
51. Fotheringham AS, Brunsdon C, Charlton M. Geographically weighted regression: the analysis of spatially varying relationships: John Wiley & Sons; 2003.
52. Nakaya T, Fotheringham AS, Brunsdon C, Charlton M. Geographically weighted Poisson regression for disease association mapping. Statistics in medicine. 2005 Sep 15;24(17):2695-717. PubMed PMID: 16118814.
53. Lovett AA, Bentham C, Flowerdew R. Analysing geographic variations in mortality using poisson regression: the example of ischaemic heart disease in England and Wales 1969–1973. Social science & medicine. 1986;23(10):935-43.
54. Lovett A, Flowerdew R. Analysis of count data using poisson regression*. The Professional Geographer. 1989;41(2):190-8.
55. Nakaya T. GWR4 user manual 2012. Available from: http://www. st-andrews. ac. uk/geoinformatics/wp-content/uploads/GWR4manual_201311. pdf.
56. Burnham KP, Anderson DR. Model selection and multimodel inference: a practical information-theoretic approach: Springer Science & Business Media; 2002.
57. Veldhuijzen IK, van Driel HF, Vos D, de Zwart O, van Doornum GJ, de Man RA, et al. Viral hepatitis in a multi-ethnic neighborhood in the Netherlands: results of a community-based study in a low prevalence country. International Journal of Infectious Diseases. 2009;13(1):e9-e13.
58. Mujeeb SA, Shahab S, Hyder AA. Geographical display of health information: study of hepatitis C infection in Karachi, Pakistan. Public health. 2000 Sep;114(5):413-5. PubMed PMID: 11035468.
59. Alter MJ, Kruszon-Moran D, Nainan OV, McQuillan GM, Gao F, Moyer LA, et al. The prevalence of hepatitis C virus infection in the United States, 1988 through 1994. New England journal of medicine. 1999;341(8):556-62.
60. Bao YP, Liu ZM, Lian Z, Li JH, Zhang RM, Zhang CB, et al. Prevalence and correlates of HIV and HCV infection among amphetamine-type stimulant users in 6 provinces in China. Journal of acquired immune deficiency syndromes. 2012 Aug 1;60(4):438-46. PubMed PMID: 22481605.
61. Rodrigues Neto J, Cubas MR, Kusma SZ, Olandoski M. Prevalence of hepatitis C in adult users of the public health service of Sao Jose dos Pinhais--Parana. Revista

brasileira de epidemiologia = Brazilian journal of epidemiology. 2012 Sep;15(3):627-38. PubMed PMID: 23090309.
62. Cavlek TV, Margan IG, Lepej SZ, Kolaric B, Vince A. Seroprevalence, risk factors, and hepatitis C virus genotypes in groups with high-risk sexual behavior in Croatia. Journal of medical virology. 2009 Aug;81(8):1348-53. PubMed PMID: 19551819.
63. van de Laar TJ, van der Bij AK, Prins M, Bruisten SM, Brinkman K, Ruys TA, et al. Increase in HCV incidence among men who have sex with men in Amsterdam most likely caused by sexual transmission. Journal of Infectious Diseases. 2007;196(2):230-8.
64. Tsai PJ, Yeh HC. Scrub typhus islands in the Taiwan area and the association between scrub typhus disease and forest land use and farmer population density: geographically weighted regression. BMC infectious diseases. 2013;13:191. PubMed PMID: 23627966. Pubmed Central PMCID: 3648375.
65. Fotheringham AS, Wong DW. The modifiable areal unit problem in multivariate statistical analysis. Environment and planning A. 1991;23(7):1025-44.
66. OECD. Demographic Change in the Netherlands: Strategies for resilient labour markets. The Netherlands: OECD, 2013.
67. Páez A, Farber S, Wheeler D. A simulation-based study of geographically weighted regression as a method for investigating spatially varying relationships. Environment and Planning-Part A. 2011;43(12):2992.

# CHAPTER 4

# Case study on Type 2 Diabetes Mellitus in Germany

published as:

Kauhl, B., Schweikart, J., Krafft, T., Keste, A., & Moskwyn, M. (2016). Do the risk factors for type 2 diabetes mellitus vary by location? A spatial analysis of health insurance claims in Northeastern Germany using kernel density estimation and geographically weighted regression. *International Journal of Health Geographics*, *15*(1), 38.

**Abstract**

**Background:** The provision of general practitioners (GPs) in Germany still relies mainly on the ratio of inhabitants to GPs at relatively large scales and barely accounts for an increased prevalence of chronic diseases among the elderly and socially underprivileged populations. Type 2 Diabetes Mellitus (T2DM) is one of the major cost-intensive diseases with high rates of potentially preventable complications. Provision of healthcare and access to preventive measures is necessary to reduce the burden of T2DM. However, current studies on the spatial variation of T2DM in Germany are mostly based on survey data, which do not only underestimate the true prevalence of T2DM, but are also only available on large spatial scales. The aim of this study is therefore to analyse the spatial distribution of T2DM at fine geographic scales and to assess location-specific risk factors based on data of the AOK health insurance.

**Methods:** To display the spatial heterogeneity of T2DM, a bivariate, adaptive kernel density estimation (KDE) was applied. The spatial scan statistic (SaTScan) was used to detect areas of high risk. Global and local spatial regression models were then constructed to analyze socio-demographic risk factors of T2DM.

**Results:** T2DM is especially concentrated in rural areas surrounding Berlin. The risk factors for T2DM consist of proportions of 65 – 79 year olds, 80+ year olds, unemployment rate among the 55 – 65 year olds, proportion of employees covered by mandatory social security insurance, mean income tax, and proportion of non-married couples. However**,** the strength of the association between T2DM and the examined socio-demographic variables displayed strong regional variations.

**Conclusion:** The prevalence of T2DM varies at the very local level. Analyzing point data on T2DM of northeastern Germany`s largest health insurance provider thus allows very detailed, location-specific knowledge about increased medical needs. Risk factors associated with T2DM depend largely on the place of residence of the respective person. Future allocation of GPs and current prevention strategies should therefore reflect the location-specific higher healthcare demand among the elderly and socially underprivileged populations.

## Introduction

The prevalence of chronic diseases and therefore the projectable utilization of healthcare depend strongly on the demographic and socio-economic composition of the respective population (1-3). International studies suggest a strong relationship between the proportion of elderly, low socio-economic status and a higher prevalence of chronic diseases (2, 4-6). However, planning of GPs in Germany still relies mainly on the ratio of inhabitants to GPs at fairly large scales (7) and does neither sufficiently reflect the location-specific higher prevalence of chronic diseases among the elderly and population groups with a lower socio-economic status, nor the accessibility of GPs in rural areas (8).

With the ongoing demographic transition and migration processes from rural to urban areas, the gap between demand and supply of health care is already widening in Germany. While the ageing of the population and therefore the prevalence of chronic diseases increases in rural areas, the availability of GPs decreases (9). To meet the increased demand for healthcare especially in rural areas, it is important to identify locations with higher healthcare demand as spatially precise as possible. Additional knowledge about the population groups, which are most at risk in specific locations is necessary to effectively plan the future provision of GPs and immediate preventive measures where they are needed most.

Type 2 Diabetes Mellitus (T2DM) is a major public health threat with an increasing prevalence among the general population worldwide (4, 10, 11) and especially in Germany (3). Prevention and access to healthcare are necessary not only to prevent a further increase but also to prevent severe complications such as lower-extremity amputation (12) or stroke (10).

Despite behavioral risk factors such as lack of physical exercise, dietary deficits and smoking (13), a wide range of studies additionally highlights an association between age, lower socioeconomic status and T2DM (4, 14-16).

Geographic information systems (GIS) and spatial regression models at the ecological level have gained increasing attention in recent years as this approach allows an analysis of possible risk factors that are often unavailable on an individual level due to privacy restriction (15, 17). For T2DM, this approach might help to identify the population groups, which are most in need for the provision of healthcare and access to preventive measures. However, several studies point out that socio-demographic risk factors for T2DM, but also for a wide range of other diseases depend largely on the place

of residence of the respective individual (4, 14, 15, 17, 18). As a consequence, a one-size fits all solution seems therefore inappropriate for effective public health strategies and allocation of healthcare (15).

Analyzing the spatial distribution of T2DM and associated risk factors in Germany is challenging, as epidemiological data on chronic diseases is seldom publicly available (19). Only few studies have examined the spatial distribution of T2DM in Germany (16, 20-23). However, the majority of these studies are based upon data from Germany's largest telephone survey of the Robert-Koch-Institute (GEDA) (16, 20, 21). A spatial analysis of this data source is therefore restricted to fairly large areas such as the counties in Germany (16, 21), or includes only a selection of municipalities (20). Analyses based on surveys however, tend to underestimate the prevalence of T2DM as persons with a higher socioeconomic status are more likely to respond than persons with a lower socioeconomic status (20, 21). Therefore, such surveys have only limited use for a demand-driven planning and allocation of healthcare and prevention strategies.

Health insurance in Germany is generally mandatory and approximately 86% of the population are covered by one of the statutory health insurance providers, 10% are covered by private health insurance providers and the remaining 4% are covered by the state (24). However, there are large socio-demographic differences between members of the various statutory health insurances (25). As the provision and allocation of primary healthcare in Germany is planned and organized by the association of statutory health insurance physicians in accordance with the statutory health insurance providers (7), it is necessary for each health insurance provider to engage in planning of primary healthcare based on an empirical evaluation of the medical demand of their respective insurants.

At the federal level, 1671 inhabitants per 1 GP at the spatial scale of central areas (Mittelbereiche) of the Federal Agency of Building and Urban Development (BBSR) is the target-ratio for the allocation of GPs in Germany (7). The association of statutory health insurance physicians defines over- or undersupply as deviation from this ratio by 110% and 50%, respectively and has to undertake appropriate measures if over- or undersupply exists (7) . However, this ratio was established in the 1990s (7) and does not recognize an increased prevalence of T2DM and other chronic diseases in location-specific population groups. The association of statutory health insurance physicians has reacted to this criticism by incorporating a demographic factor and allowing deviations

from the established inhabitants to GP ratio for areas with increased medical demand in their revised planning guidelines (7). However, due to the lack of reliable, small-scale public health data on chronic diseases, an increased medical demand of a location-specific population group is still difficult to detect (16, 20-23). To realistically capture such an increased demand for healthcare, more reliable sources than survey data and spatial analyses at smaller scales are necessary than it is currently possible with survey data in Germany.

In this context, health insurance claims of the AOK Nordost have several advantages over survey data: (a) This data source represents a large sample of northeastern Germany`s population, (b) can be analyzed on a fine geographic scale and (c) prevalence estimates of health insurance claims are not depending on the response rate of participants and are therefore a more realistic estimate of the "true" prevalence of chronic conditions than survey data (26). Ultimately, a spatial analysis of this data source might provide new and inclusive insights on the spatial distribution of chronic diseases and population-based risk factors.

The goal of our paper is therefore to (i) analyze the spatial distribution of T2DM based on health insurance claims of northeastern Germany`s largest statutory health insurance provider; (ii) to evaluate possible risk factors using global ecological regression models and (iii) to examine the spatially varying association between socio-demographic risk factors and T2DM.

## Methods

### Dependent Variable

In this study, we used data from northeastern Germany`s largest statutory health insurance provider (AOK Nordost) for 2012, which covers roughly 1.79 million persons (approximately one quarter of the population) of which 361 thousand persons are diagnosed with Type 2 Diabetes.

Persons diagnosed with T2DM were defined in our study as having a confirmed diagnosis of T2DM (ICD-10: E11.-). As long as the insurant is treated for T2DM, this diagnosis will remain in the insurant`s personal medical file as the diagnosis is renewed with each GP visit associated with T2DM. To ensure that each insurant and diabetic is included only once in the analysis, the unique insurant number was used to exclude possible double entries within the database from the analysis.

The data was anonymized and was geocoded based on exact street-level data using the ESRI ArcGIS geocoder. The data included only age in broad age categories (0 – 5, 6 – 11, 12 – 17, 18 – 24, 25 – 44, 45 – 64, 65 – 79 and 80 and older) and the address coordinates. We used a step-wise geocoding process where the data was first geocoded based on the exact street address where possible (90.2%). Of the remaining data, 6.7% were matched to the centroids of the street and 3.1% were matched to the postal code centroids. The address coordinates for Berlin were obtained from the Senatsverwaltung für Stadtverwaltung Berlin; the address coordinates for Brandenburg were obtained from the Landesvermessungsamt und Geobasisinformation Brandenburg (Geobasisdaten © GeoBasis-DE/LGB 2016, GB-D 13/16 ) and the coordinates for Mecklenburg-Vorpommern were obtained from the Landesamt für Innere Verwaltung, Amt für Geoinformation, Vermessungs- und Katasterwesen (Geobasisdaten © GeoBasis-DE/M-V 2016).

**Explanatory Variables**

In this study, we assessed a wide range of demographic, socioeconomic and variables related to the physical environment for their association with T2DM. Demographic variables were calculated based on the proportion of AOK insurants per demographic group. Socioeconomic variables included the proportion of unemployed persons in different age groups, information on taxation, land use, household composition and a wide range of other indicators. Variables related to the physical environment included the proportion of green spaces, recreational spaces and built surfaces. The data were obtained for the year 2012 from the INKAR database of the Federal Agency of Building and Urban Development (BBSR). Data on marital status, household and family composition were obtained from the census 2011 for Germany. All data were available on the spatial scale of the association of municipalities. Additionally, we included data on the spatial distribution of GPs in our analysis to examine whether the availability of healthcare influences the prevalence of T2DM. We included two variables: The proportion of inhabitants to GPs and the average distance to GPs. The average distance to GPs was calculated based on the driving distance of each insurant to the closest GP and was then aggregated to match the association of municipalities. The street network dataset was downloaded from OpenStreetMap (27). The association of municipalities in Germany was chosen as the unit of analysis as this is the smallest spatial scale, for which a wide range of indicators is available without areas being

omitted due to privacy protection as it would be the case for municipalities. However, this scale does not allow an analysis of intra-urban differences as the indicators of BBSR are not available for a smaller administrative unit than the association of municipalities.

## Statistical Analysis

### Bivariate Kernel Density Estimation

In this study, we used a bivariate, adaptive kernel density estimation (KDE) to display the spatial heterogeneity of T2DM independent of administrative boundaries. In most epidemiological studies, disease and population data are only available for aggregated data such as postal codes, municipalities, counties or districts (10, 16, 21, 28). However, problems arise in the detection of local clusters and associations to socio-demographic exposure factors due to the relatively arbitrary shape and quantity of spatial units, which is often referred to as the "modifiable area unit problem" (29). This may be especially misleading in rural areas where administrative boundaries are very large. As a consequence, a cartographic visualization of disease risk without the restrictions of artificially created boundaries is favorable.

Bivariate kernel density estimation has been previously applied in small-scale studies for HIV (30, 31), cancer (32, 33), Alzheimer (34) and crime intensity (35) and thus seems useful for a small-scale analysis of T2DM as well.

A major concern when applying a bivariate KDE is the choice of bandwidth. If the bandwidth is too small, rates become highly unstable and spatial patterns are difficult to detect. If the bandwidth is too large, the map appears to be over smoothed and local extremes are smoothed away (33). Although several statistical models exist to calculate the "optimal" bandwidth, such as the Likelihood Cross Validation (33, 36, 37) , Least Squares Cross Validation (33, 38), Biased Cross Validation (33, 39), Smoothed Cross Validation (33, 40), or the direct plug-in method (33, 41), these aforementioned bandwidth selection models are generally only available for fixed bandwidth types (33). As our study area comprises highly densely populated urban areas such as Berlin, Potsdam or Schwerin while at the same time comprising a large proportion of very sparsely populated rural areas, a KDE employing a fixed bandwidth would deliver no stable results. We therefore favored an adaptive bandwidth, which accounts for the varying population densities within our study area (32, 33).

Although a wide range of selection methods exist for a fixed bandwidth, automated procedures to select an optimal number of points to be included in an

adaptive bandwidth for a bivariate KDE are scarce and are not yet fully satisfactory (33). As there are no definite recommendations to define a bandwidth for a bivariate KDE, we therefore visually evaluated several possible combinations of minimum sample points (42, 43). Including at least 0.1% of T2DM cases and 0.1% of insurants delivered the most useful results. The T2DM prevalence was therefore calculated as the ratio of at least 361 T2DM cases per $km^2$ to 1,791 insurants per $km^2$. Given the varying population densities, the kernel was thus smaller in highly populated areas and larger in sparsely populated rural areas. In this study, we used a Gaussian kernel as it tends to produce more robust results than a kernel type with a definite boundary (43).

The calculation of the bivariate KDE was carried out using the CrimeStat IV software (43). The results were then imported in ESRI ArcGIS 10.3.

**Sex- and Age-Standardization of Prevalence Rates**

The bivariate, adaptive kernel density estimation allows a visualization of T2DM prevalence without the limitations of administrative areas but has the disadvantage of not being able to incorporate sex- and age-standardization.

To further facilitate interpretation of the spatial variations in T2DM prevalence, we directly adjusted for sex and age using the WHO standard population from 1976 (44) based on the five-digits postal codes of our study area. As the number of insurants between the five-digits postal code varies considerably, we applied spatial empirical Bayesian smoothing to borrow strength from neighboring postal codes to estimate more stable prevalence rates (45). Neighboring areas were defined as postal codes sharing a common edge or boundary (46). The computation was carried out in GeoDa 1.2.0 and the results were then imported in ESRI ArcGIS 10.3.

**Cluster Detection**

The aim of cluster detection in our study was to evaluate whether a statistically significant elevated risk exists in certain areas. A purely visual inspection of the KDE and the adjusted rates would be misleading, as it is not possible to examine the number of cases behind the estimated rates alone. Applying a local cluster test on health data is important to prioritize areas for future public health interventions (30, 47) and has been previously shown useful to locate new clinics for chronically ill patients for diabetic kidney patients (48).

In this study, we used the spatial scan statistic (SaTScan). The spatial scan statistic is a local cluster test, which determines the location and significance of local clusters. This is achieved by a circular scanning window, which moves over the coordinates of the study area and evaluates all possible cluster locations and cluster sizes up to either a user defined maximum or the default settings of including up to 50% of the population at risk inside a cluster (30, 49). The statistical significance is calculated using 999 Monte-Carlo replications (50). We applied a purely spatial Poisson model, where the T2DM cases per coordinate / sex- and age-adjusted number of T2DM cases per postal code were assigned as cases and all insurants per coordinate / postal code were assigned as population (30, 49, 50). The maximum cluster size was restricted to a maximum radius of 10km. This was done as (a) the standard setting of including up to 50% of the population at risk often produces results of no practical use (51) and (b), we defined 10km as the maximum reasonable driving distance to GPs in rural areas of northeastern Germany. For the analysis of the point data, we used the exact street-level coordinates and for the cluster analysis of the sex- and age-adjusted rates we used the centroid coordinates of the postal codes. The analysis was carried out using SaTScan v9.4.2.

**Spatial Regression Modelling**

**Ordinary Least Squares Regression Modelling**

To create a meaningful and correct specified geographically weighted regression model (GWR), we first aimed to identify all possible explanatory variables through the global ordinary least squares (OLS) regression model. To achieve this, we first performed a natural log-transformation of the T2DM prevalence to satisfy the assumption of the OLS model that the dependent variable has to be normally distributed (52). We used the raw rate instead of the age-adjusted T2DM prevalence as we specifically wanted to model the effect of older age groups on the T2DM prevalence.

We then compared the association between each potential explanatory variable and T2DM prevalence through univariate OLS regression models. As a large number of explanatory variables were found to be significantly associated to T2DM, we used a data-mining tool called “exploratory regression” in ESRI ArcGIS 10.3 to determine all possible variable combinations. This tool is comparable to a step-wise regression. It evaluates all possible variable combinations based on four criteria: (i): the coefficients are statistically significant; (ii): the explanatory variables are free from multicollinearity;

(iii): the residuals are normally distributed and (iv): the residuals are not spatially autocorrelated (52-54).

We then determined overall model significance, autocorrelation of the residuals, the presence of heteroscedasticity and a wide range of other diagnostics by creating an OLS model in ESRI ArcGIS 10.3. with the same explanatory variables as suggested by the exploratory regression that were found to deliver a plausible explanation of the T2DM prevalence.

**Geographically Weighted Regression Modelling**

The OLS model is a global model, it therefore estimates only one single coefficient per explanatory variable averaged over the entire study area. However, the socio-demographic composition of the population in northeastern Germany varies strongly at the local level. It is therefore unlikely that the association between socio-demographic explanatory variables and T2DM is realistically reflected by a global regression model. Previous studies applying GWR on Diabetes (4, 15) as well as on a wide range of other diseases (18, 55, 56) pointed out that the correlations between explanatory variables and T2DM vary strongly across space. We therefore hypothesize that this applies to our study area as well. The GWR methodology is an extension to the standard regression models and estimates a wide range of local parameters to reflect changes over space in the association between an epidemiological outcome and explanatory variables (57).

Similar to the OLS model, we used the log-transformed T2DM prevalence as the dependent variable with the same explanatory variables that were found to be significant in the OLS model.

We used the centroids of the association of municipalities as the input coordinates. Similarly to the KDE, the GWR methodology uses a circular kernel to calculate the local estimates. The GWR model fits for each coordinate a regression equation where the coordinates in the center of the kernel are the regression points. The data points inside the kernel are then weighted with decreasing weights from the center towards the edge of the kernel. The bandwidth of the kernel can be either fixed or adaptive and the shape of the kernel can follow a Gaussian or a bi-square distribution. The optimization of the bandwidth can be based on one of the four available criteria: (i) Akaikes Information Criterion (AIC); (ii) Akaikes corrected Information Criterion (AICc); (iii) Bayesian Information Criterion (BIC) and (iv) Cross Validation (CV) (57, 58). We thus evaluated all 14 possible combinations of kernel shape, bandwidth type and

bandwidth optimization method. The models without clustered residuals were further considered and out of those, the model with the lowest AICc value and highest adjusted $R^2$ was then chosen as the final model. The calculation of the GWR model was carried out in the GWR4 software. To enhance visualization of the spatially varying coefficients, we used the software`s "prediction at non-sample points" function and calculated the predicted values for a grid of northeastern Germany based on a cell size of 100m x 100m. The obtained values were then interpolated using ordinary kriging in ESRI ArcGIS 10.3.

**Ethics Statement**

The data and results used in this study were anonymized and do not contain any personal information. The use of anonymized data for research purposes does not require a vote by an ethics committee or an institutional research board.

**Results**

**Spatial Distribution of T2DM**

The overall raw prevalence of T2DM was 20.0% and the sex- and age-adjusted prevalence was 14.2%. However, the prevalence varied widely within the study area (Fig. 1). Generally, the prevalence was relatively low in the center of larger villages or urban areas and increased towards remote, rural areas. The highest prevalence and clusters with most cases could be observed in a ring in Brandenburg, surrounding Berlin. In Mecklenburg-Vorpommern, the number of clusters as well as the number of cases inside local clusters was lower than in Brandenburg.

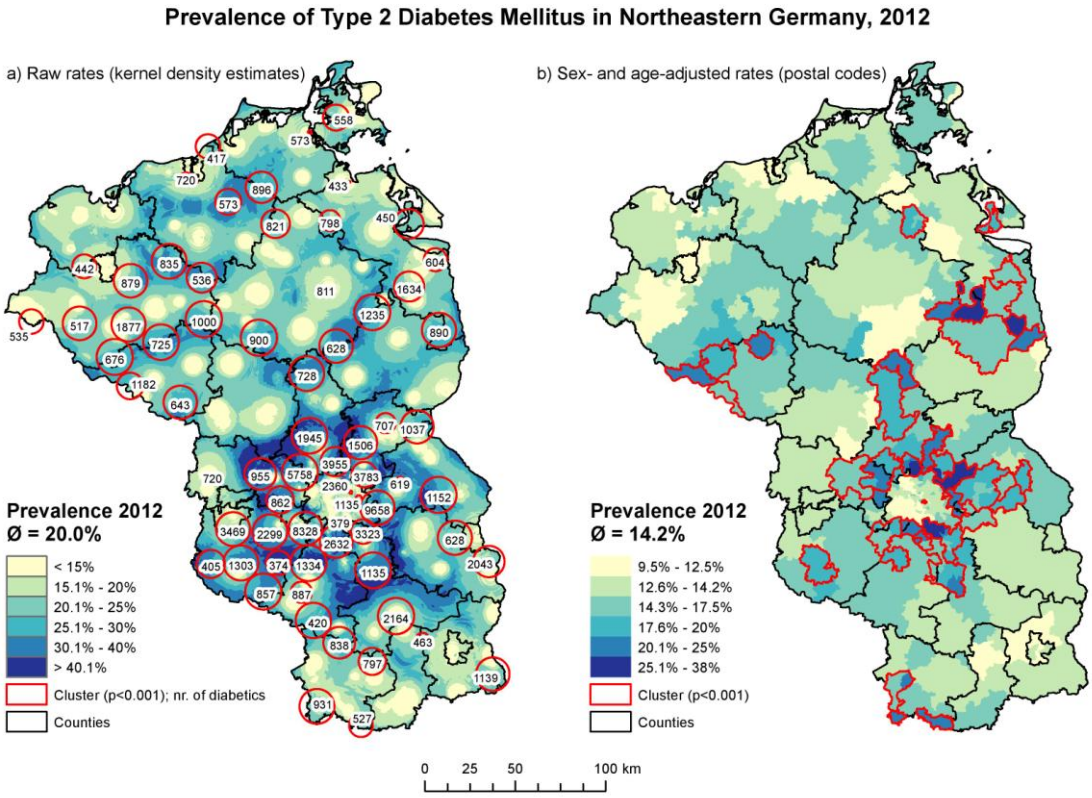



**Fig. 1:** The spatial distribution of T2DM in northeastern Germany represented as a) KDE estimates of the raw rate and b) sex- and age-adjusted rates based on the five-digit postal codes

## Socio-demographic Risk Factors of T2DM

Six variables were identified as significant predictors for T2DM in northeastern Germany (Table. 1): (i) proportion of persons aged 65 – 79, (ii) proportion of persons aged 80 and older, (iii) proportion of unemployed persons aged 55 – 65; (iv) proportion of employed persons which are subject to social insurance contribution, (v) mean income tax and (vi) proportion of non-married couples, which live together in the same household. These six variables explained 44% of the variation in T2DM prevalence (Table 1). However, the residuals were clustered, reflecting that a global OLS model is not suitable to model the prevalence of T2DM.

**Table 1:** Results of the global OLS regression model. Significance levels: * ≤ 0.05; ** ≤ 0.01: *** ≤ 0.001.

| Variable | Coefficient | VIF |
|---|---|---|
| Intercept | 2.259540*** | |
| Persons aged 65 – 79 (%) | 0.027251*** | 1.656689 |
| Persons aged 80 and older (%) | 0.010704** | 1.650654 |
| Unemployed persons aged 55 – 65 (%) | 0.013354*** | 2.593295 |
| Employed persons (%) | -0.006181** | 1.602619 |
| Mean Income tax | 0.000780** | 2.272369 |
| Non-married couples (%) | 0.014524* | 1.452730 |
| | | |
| Adjusted $R^2$ | 0.44 | |
| AICc | -313 | |
| Global Moran`s I of residuals | I = 0.26 (p<0.001) | |

**Spatially-Varying Risk Factors of T2DM**

By comparing all 14 possible combinations of bandwidth type, kernel shape and optimization methods in terms of their AICc value, adjusted $R^2$ and Moran`s I of the residuals (Table 2), the model using an adaptive bandwidth with a bi-square kernel shape and an AIC optimized bandwidth selection method fulfils the requirements of the residuals not being clustered and has the best model fit, both, in terms of the AICc value and adjusted $R^2$. This model explains 66% of the spatial variations of T2DM prevalence and has a much better fit (AICc: -374) than the global OLS model (AICc: -313). This suggests that a local model is more suitable to model the socio-demographic risk factors for T2DM than a global model.

**Table 2:** Comparison of bandwidth types, kernel shapes and bandwidth optimization methods

| Model (bandwidth type, kernel shape, optimization method) | AICc | Adjusted $R^2$ | Moran`s I of Residuals |
|---|---|---|---|
| Adaptive, Gaussian, AICc | -347 | 0.51 | p<0.001 |
| Adaptive, Gaussian, AIC | -347 | 0.51 | p<0.001 |
| Adaptive, Gaussian, BIC | -315 | 0.44 | p<0.001 |
| Adaptive, Gaussian, CV | -347 | 0.51 | p<0.001 |
| Fixed, Gaussian, AICc | -385 | 0.62 | p< 0.05 |
| Fixed, Gaussian, AIC | -265 | 0.66 | p>0.05 |
| Fixed, Gaussian, BIC | -316 | 0.44 | p<0.001 |
| Fixed, Gaussian, CV | -370 | 0.64 | p>0.05 |
| Adaptive, bi-square, AICc | -394 | 0.63 | p<0.001 |
| Adaptive, bi-square, AIC | -374 | 0.66 | p>0.05 |
| Adaptive, bi-square, BIC | -320 | 0.45 | p<0.001 |
| Fixed, bi-square, AICc | -385 | 0.62 | p<0.01 |
| Fixed, bi-square, AIC | 40 | 0.68 | p>0.05 |
| Fixed, bi-square, BIC | -316 | 0.44 | p<0.001 |

The cartographic visualization of the GWR regression coefficients revealed strong regional differences of the association between the examined socio-demographic variables and T2DM prevalence (Fig. 2).

The impact of proportion of persons aged 65 – 79 was strongest in the areas north of Berlin in Brandenburg and two districts in the western part of Mecklenburg-Vorpommern. In these areas, 1% increase in persons aged 65 – 79 will increase the prevalence of T2DM between 3.2% and 5.4%. The association between persons aged 65 – 79 and T2DM prevalence was not significant in several districts west of Berlin and the northeastern districts in Mecklenburg-Vorpommern.

The association to proportion of persons aged 80 and older was significant in those areas where persons aged 65 – 79 were not significant with the exception of the islands Rügen and Usedom. The strongest impact could be observed in parts of the districts Vorpommern-Greifswald, Mecklenburgische Seenplatte and Prignitz. In these

areas, 1% increase in persons aged 80 and older will increase the T2DM prevalence between 2.3% and 4%.

Unemployment rate among persons aged 55 – 65 was a significant positive predictor in several districts north of Berlin in Brandenburg and Mecklenburg-Vorpommern. In these areas, 1% increase in unemployment among the 55 – 65 year olds will increase the prevalence of T2DM between 3.8% and 6.6%. A significant negative association could only be observed in a small part of the districts Oder-Spree and Dahme-Spreewald. 1% decrease of unemployment among the 55 – 65 year olds will increase the T2DM prevalence between 1.3% and 6.4%.

The association between proportion of employed persons, which are subject to social insurance contribution, and T2DM changed sign across the study area. In the areas, where the proportion of employed persons was significant positively associated, 1% increase in employed persons was associated with 1.5% to 3.5% increase in T2DM prevalence. In the areas where the proportion of employed persons was significant negatively associated, 1% decrease of employed persons was associated with a 0.5% to 3.2% increase in T2DM prevalence. However, the association between employed persons and T2DM was only significant in a fraction of areas.

Similar to proportion of employed persons, the association between mean income tax and T2DM changed sign across the study area. In several districts north of Berlin, where the association between income tax and T2DM prevalence was positive, 10 Euro income tax per person per year will increase the T2DM prevalence by 0.1% to 3.2%. In the areas where the association to income tax was significant negative, 10 Euro less income tax per person per year will increase the T2DM prevalence between 1.6% and 3%.

The proportion of non-married couples sharing a common flat was only significant in several small parts of the districts Dahme-Spreewald and Teltow-Fläming. In these areas, 1% increase in non-married couples will increase the T2DM prevalence between 2.2% to 6.3%.

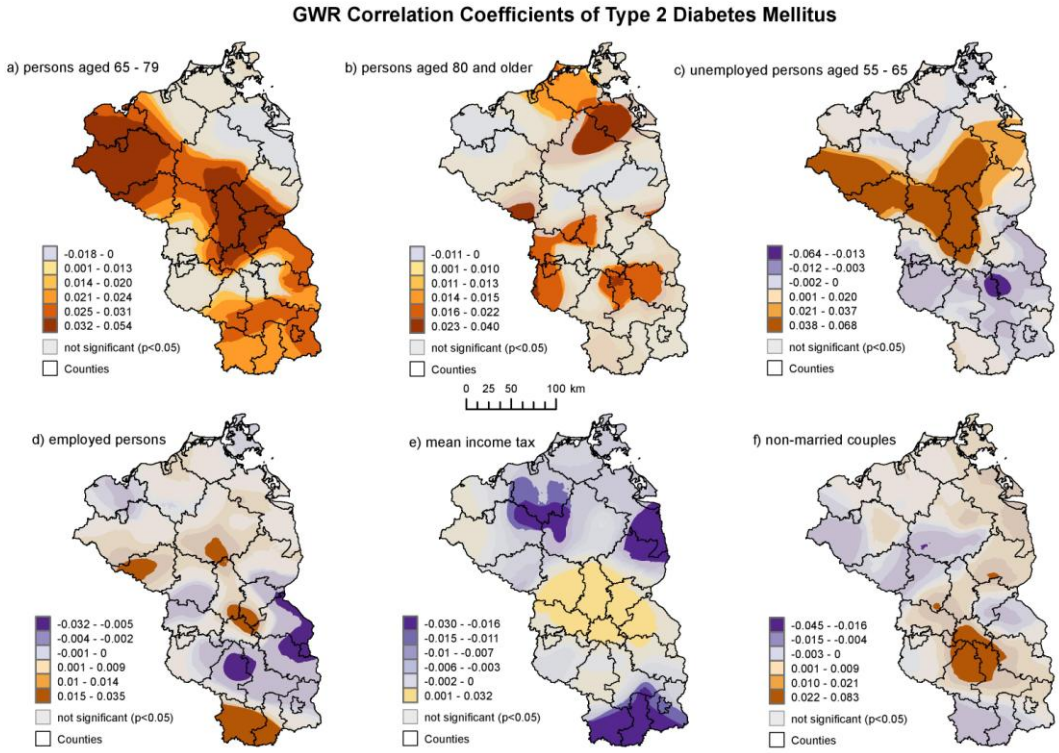



**Fig. 2:** GWR correlation coefficients of Type 2 Diabetes Mellitus

## Discussion

The prevalence of T2DM varies strongly at the very local level and clusters especially in rural areas in Brandenburg and Mecklenburg-Vorpommern. Socio-demographic risk factors consisted of proportion of persons aged 65 – 79, proportion of persons aged 80 and older, unemployment rate among the 55 to 65 year olds, proportion of employed persons, which are subject to social insurance contribution, mean income tax and proportion of non-married couples sharing a common flat. However, all associations displayed strong regional differences.

The overall prevalence of T2DM was 20%. After adjusting for sex and age, the prevalence of 14.2 % was still higher than national estimates based on data derived from the telephone survey of the Robert-Koch-Institute (GEDA), which estimated the prevalence of known Diabetes to be at 8.8% among adults in Germany (3). However, estimates derived from surveys such as the GEDA study are rather underestimated as healthy participants are more likely to respond than chronically ill patients (20, 21). In this study, the estimated prevalence exceeds these previous estimates by far. As our study area comprises the most deprived areas in Germany (28), it is not surprising that

our estimates exceed those of the GEDA study. Additionally, the proportion of older inhabitants, persons with low levels of education and unemployed persons among the local AOK health insurances is generally higher than in other statutory health insurances. As a logical consequence, the prevalence of chronic diseases is higher in our population sample than in the rest of the population (25).

The spatial distribution of T2DM varied strongly and formed clusters on small geographic scales. This was reflected by the results of the bivariate kernel density estimation and the results of the spatial scan statistic. Spatial heterogeneity and local clustering is typical for a wide range of chronic diseases (12, 59-62). Our results are therefore in line with other studies but add an important level of spatial detail to previous research. The combination of the bivariate KDE and the spatial scan statistic complimented each other fairly well using the settings chosen in this study. However, we had to use a very conservative p-value for the cluster analysis, as the number of clusters using a p-value of 0.05 was simply too high to allow a detailed investigation.

We identified six risk factors for T2DM in northeastern Germany: (i) proportion of persons aged 65 – 79, (ii) proportion of persons aged 80 and older, (iii) proportion of unemployed persons aged 55 – 65; (iv) proportion of employed persons which are subject to social insurance contribution, (v) income tax and (vi) proportion of non-married couples, which live together in the same household.

The association of T2DM to older age groups was expected as T2DM displays a strong association to older age groups (3, 4, 22). The association of T2DM to the proportion of persons aged 65 – 79 and persons aged 80 and older is therefore in line with these studies although these associations were not in the entire study area significant.

Several studies pointed out that T2DM is associated with a lower socio-economic status (4, 14-16). This is reflected by the strong association of unemployed persons aged 55 – 65 to T2DM. Given the high proportion of older persons among the AOK insurants, it is not surprising that specifically the unemployment rate among persons aged 55 – 65 was significant, but not unemployment rate in general. Additionally, this reflects the value of stratified socio-economic data as these findings could allow a more targeted prevention strategy among the at-risk population group.

The association to employed persons, which are subject to social insurance contribution, has to be seen in the context of income tax. Employed persons were positively associated in the areas, where income tax was negatively (but not significant)

associated with T2DM prevalence. This reflects the association of T2DM to the lower-income groups (4, 15) and thus highlights the importance of determining location-specific association for T2DM. The negative association of employed persons to T2DM in specific areas can in part be explained by the exclusion criteria of employed persons in Germany. Excluded under this definition are for example persons working in marginal employment, soldiers, self-employed persons, non-working family members and government officials (63). Given the association of T2DM to lower socio-economic status, these results might indicate that in areas where the association to employed persons is negative, persons working in marginally employment and non-working family members are at major risk for T2DM.

Although income tax was overall positively associated to T2DM, the results of GWR point out that income tax was in several areas significant negatively associated, confirming the results of previous studies (4, 15). The positive association of income tax to T2DM prevalence is very specific to the area surrounding Berlin, which is often referred to as the commuter belt. This positive association reflects that in specific areas, a higher income may pose a risk factor for T2DM as well.

Several studies have shown that marital status has an effect on the overall health of the population. An unmarried status is often associated with a higher prevalence of chronic diseases and premature death (64), although not all studies can confirm this association (65). The positive association of non-married couples sharing a common flat to T2DM can therefore be considered as very specific to the commuting belt around Berlin. Further research on an individual level is necessary to confirm this association.

Although several studies found an association between land-use, built environment and obesity and T2DM (66, 67), we found only a very moderate association between the proportion of built surfaces and T2DM. After carefully reviewing the results of a GWR model including the proportion of built surfaces as independent variable, we concluded that this association was misleading in our study area as it was only significant in the most sparsely populated area in Brandenburg. This seems implausible as villages in this area are generally very small and green spaces are widely available and accessible in walking distance. We thus excluded the proportion of built areas as independent variable from our analysis. However, this highlights the value of local regression models over global regression models to question the plausibility of possible associations.

We found no associations between availability of GPs and the prevalence of T2DM. Thus, access to and availability of GPs has no influence on the diagnosis of T2DM in our study area. Since the majority of T2DM is detected among persons in their 40s and older (68), and diabetics in rural areas consulting GPs less frequently than diabetics in urban areas (69), it seems reasonable to assume that a substantial amount of diabetics in our study area only sought medical attention when symptoms of T2DM persisted as our population sample is older than the rest of northeastern Germany`s population. As a consequence, the number of undiagnosed diabetics in rural areas is potentially higher among middle-aged persons, which do not display any symptoms yet.

## Strengths and Limitations

### Strengths

In this study, we used a large database, consisting of 1.8 million insurants. Our results clearly demonstrate that a spatial analysis using "big data" of health insurance providers is feasible and can be used to provide a finer spatial resolution for prevalence estimates of T2DM than it is currently possible with survey data.

Several spatial-epidemiological studies highlight the benefits of performing a cluster test based on point data over administrative data (30, 70, 71). Detailed cluster detection based on point data could not only enhance prevention strategies (17, 30) but could also be used for a demand-driven allocation of healthcare facilities where they are needed most (48). In northeastern Germany, this is of particular importance as the population is very unevenly distributed and the smallest administrative unit – municipalities – vary strongly in size and population among the states (72). Further, Germany`s largest city Berlin counts as only one municipality. Five-digit postal codes were thus used for the sex- and age standardization to highlight intra-urban differences. German postal codes have the disadvantage of - specifically in predominantly rural regions - covering very large areas and are thus not very suitable for the allocation of future healthcare. As a consequence, our approach of combining a bivariate KDE with a cluster analysis may serve as an alternative and relative exact prioritization for allocating new GP resources in the near future.

### Limitations

First, our study was based on health insurance claims of northeastern Germany`s largest statutory health insurance provider. Although the AOK Nordost covers

approximately one quarter of the population, the results cannot be assumed to sufficiently reflect the prevalence of T2DM for the whole population. Large socio-demographic differences exist between the insurants of the various statutory health insurance providers with the AOK having the largest proportion of persons with low income, low educational level and thus the highest prevalence of chronic diseases (25).

Second, we included all persons that were insured in 2012 with the AOK Nordost, irrespective of the length of insurance. We therefore did not exclude persons who died in 2012 from the analysis or persons being insured for short time-periods as these persons still contributed to the overall prevalence of T2DM.

Third, it is clear that the results of the bivariate KDE for T2DM represent the demographic distribution of insurants to a certain extent, given the strong association of T2DM to older age groups (3, 4, 22). However, age-standardization is currently not available for a bivariate KDE in the CrimeStat IV software. As a consequence, the combined results of the bivariate KDE and the spatial scan statistic are more relevant for immediate allocation of GPs than for long-term planning of future healthcare.

Fourth, although most clusters were concentrated in areas with above-average prevalence estimates of the KDE, a small proportion of clusters was also concentrated in areas with below-average prevalence estimates. This is attributable to the different settings used in this study for the bivariate KDE and the spatial scan statistic. As we used an adaptive kernel for the KDE and a fixed radius of 10km for the spatial scan statistic, higher prevalences cannot be sufficiently visualized if several hundred cases are concentrated in a very small location. This may occur for example with adjacent multi-story apartment blocks, which still constitute a significant cluster as detected by the spatial scan statistic but are smaller than the resolution offered by the KDE. When using fixed bandwidths of the same size for KDE and the spatial scan statistic simultaneously, this problem becomes less prominent (30).

Fifth, the associations examined in this study are based on aggregated data. Although our results generally reflect the results of other spatial-epidemiological studies on T2DM, it is necessary to review whether the associations revealed in this study at the ecological level are also valid associations on an individual level.

**Implications for Future Planning of Healthcare**

Our results clearly demonstrate that the prevalence of T2DM varies at very fine geographic scales. The small-scale spatial variability of T2DM thus challenges the

applicability of the spatial scale of central areas (Mittelbereiche) at which the allocation of GPs is currently planned (7, 73). Based on our results, a planning on smaller scales such as the association of municipalities would be more suitable to reflect the strong spatial variability of T2DM. It has been argued that the current provision of GPs – based on the ratio of 1 GP per 1671 inhabitants (7) – is too simplified and also outdated (8, 74). The association of T2DM to location-specific socio-demographic population characteristics demands a strong deviation from these ratios and calls for a stronger acknowledgement of increased medical needs among the elderly and socially underprivileged populations. The revised planning guidelines of the federal association of statutory physicians in 2013 would allow deviations from the current ratio for areas with a particular high prevalence of diseases or specific socio-economic characteristics (75). However, these revised planning guidelines still remain unspecific on how exactly a particular high prevalence or specific socio-economic characteristics can be translated into additional GP positions for a particular area. As a consequence, our analysis can only point out areas with a currently high medical demand and location-specific associations between T2DM and socio-demographic population characteristics.

Given that the spatial variability of T2DM is not only determined by current socio-demographic factors but also by the change of these factors over time (4), the results of our GWR analysis could serve as a first basis in developing approaches to model the expected, long-term future burden of T2DM to assist in allocating future GPs where they will be needed most.

**Conclusion**

This is to date the largest small-scale spatial-epidemiological study of T2DM in northeastern Germany. Our results clearly show that T2DM varies at the very local level and that a large variation of T2DM prevalence can be explained by location-specific, socio-demographic population characteristics. Future planning of healthcare would greatly benefit from smaller spatial scales and need to deviate from simple inhabitants to GP ratios to reflect the increased prevalence of chronic diseases in older and socially underprivileged population groups. These results are therefore valuable for the future planning of healthcare in northeastern Germany. Our approach of analyzing the spatial distribution of chronic diseases at the very local level and geographically weighted regression is not only useful for northeastern Germany, but could be an effective way of

targeting location-specific population groups with increased medical needs as precisely as possible in all countries, where chronic diseases are on the rise.

**References:**


1. Glynn LG, Valderas JM, Healy P, Burke E, Newell J, Gillespie P, et al. The prevalence of multimorbidity in primary care and its effect on health care utilization and cost. Family practice. 2011;28(5):516-23.
2. Dalstra JA, Kunst AE, Borrell C, Breeze E, Cambois E, Costa G, et al. Socioeconomic differences in the prevalence of common chronic diseases: an overview of eight European countries. International journal of epidemiology. 2005;34(2):316-26.
3. Heidemann C, Du Y, Scheidt-Nave C. Diabetes mellitus in Deutschland. 2011.
4. Dijkstra A, Janssen F, De Bakker M, Bos J, Lub R, Van Wissen LJ, et al. Using spatial analysis to predict health care use at the local level: a case study of type 2 diabetes medication use and its association with demograpHic change and socioeconomic status. PloS one. 2013;8(8):e72730.
5. Kanjilal S, Gregg EW, Cheng YJ, Zhang P, Nelson DE, Mensah G, et al. Socioeconomic status and trends in disparities in 4 major risk factors for cardiovascular disease among US adults, 1971-2002. Archives of Internal Medicine. 2006;166(21):2348-55.
6. Avendano M, Kunst AE, Huisman M, Lenthe FV, Bopp M, Regidor E, et al. Socioeconomic status and ischaemic heart disease mortality in 10 western European populations during the 1990s. Heart. 2006;92(4):461-7.
7. Bundesausschuss G. Bedarfsplanungs - Richtlinie Stand: 15. Oktober 2015 des Gemeinsamen Bundesausschusses über die Bedarfsplanung sowie die Maßstäbe zur Feststellung von Überversorgung und Unterversorgung in der vertragsärztlichen Versorgung: Gemeinsamer Bundesausschuss; 2012 [cited 2016 17th May]. Available from: https://http://www.g-ba.de/downloads/62-492-1109/BPL-RL_2015-10-15_iK-2016-01-06.pdf.
8. Ozegowski S, Sundmacher L. Wie „bedarfsgerecht "ist die Bedarfsplanung? Eine Analyse der regionalen Verteilung der vertragsärztlichen Versorgung. Gesundheitswesen. 2012;74(10):618-26.
9. Swart E, von Stillfried DG, Koch-Gromus U. Wo sich Wissenschaft, Praxis und Politik treffen. 2014.
10. Barker LE, Kirtland KA, Gregg EW, Geiss LS, Thompson TJ. Geographic distribution of diagnosed diabetes in the US: a diabetes belt. American journal of preventive medicine. 2011;40(4):434-9.
11. Wild S, Roglic G, Green A, Sicree R, King H. Global prevalence of diabetes estimates for the year 2000 and projections for 2030. Diabetes care. 2004;27(5):1047-53.
12. Margolis DJ, Hoffstad O, Nafash J, Leonard CE, Freeman CP, Hennessy S, et al. Location, location, location: geographic clustering of lower-extremity amputation among Medicare beneficiaries with diabetes. Diabetes care. 2011;34(11):2363-7.
13. Espeland M. Reduction in weight and cardiovascular disease risk factors in individuals with type 2 diabetes: one-year results of the look AHEAD trial. Diabetes care. 2007.
14. Siordia C, Saenz J, Tom SE. An introduction to macro-level spatial nonstationarity: a geographically weighted regression analysis of diabetes and poverty. Human geographies. 2012;6(2):5.
15. Hipp JA, Chalise N. Peer Reviewed: Spatial Analysis and Correlates of County-Level Diabetes Prevalence, 2009–2010. Preventing chronic disease. 2015;12.
16. Maier W, Scheidt-Nave C, Holle R, Kroll LE, Lampert T, Du Y, et al. Area level deprivation is an independent determinant of prevalent type 2 diabetes and obesity at the national level in Germany. Results from the National Telephone

Health Interview Surveys 'German Health Update'GEDA 2009 and 2010. PloS one. 2014;9(2):e89661.
17. Kauhl B, Heil J, Hoebe CJ, Schweikart J, Krafft T, Dukers-Muijrers NH. The Spatial Distribution of Hepatitis C Virus Infections and Associated Determinants—An Application of a Geographically Weighted Poisson Regression for Evidence-Based Screening Interventions in Hotspots. PloS one. 2015;10(9):e0135656.
18. Weisent J, Rohrbach B, Dunn JR. Socioeconomic determinants of geographic disparities in campylobacteriosis risk: a comparison of global and local modeling approaches. International journal of health geographics. 2012;11(1):1.
19. Wittchen H-U, Pieper L, Eichler T, Klotsche J. Prävalenz und Versorgung von Diabetes mellitus und Herz-Kreislauf-Erkrankungen: DETECT—eine bundesweite Versorgungsstudie an über 55.000 Hausarztpatienten. Prävention und Versorgungsforschung: Springer; 2008. p. 315-28.
20. Grundmann N, Mielck A, Siegel M, Maier W. Area deprivation and the prevalence of type 2 diabetes and obesity: analysis at the municipality level in Germany. BMC public health. 2014;14(1):1.
21. Kroll LE, Lampert T. Regionale Unterschiede in der Gesundheit am Beispiel von Adipositas und Diabetes mellitus. Robert Koch-Institut, editor Daten und Fakten: Ergebnisse der Studie» Gesundheit in Deutschland aktuell. 2010:51-9.
22. Erhart M, Herring R, Schulz M, Stillfried Dv. Morbiditätsatlas Hamburg. Gutachten zum kleinräumigen Versorgungsbedarf in Hamburg–erstellt durch das Zentralinstitut für die kassenärztliche Versorgung in Deutschland, im Auftrag der Behörde für Gesundheit und Verbraucherschutz Hamburg «Hamburg. 2013;7.
23. Schipf S, Werner A, Tamayo T, Holle R, Schunk M, Maier W, et al. Regional differences in the prevalence of known Type 2 diabetes mellitus in 45–74 years old individuals: Results from six population-based studies in Germany (DIAB-CORE Consortium). Diabetic medicine. 2012;29(7):e88-e95.
24. Ziegler U, Doblhammer G. Prävalenz und Inzidenz von Demenz in Deutschland–Eine Studie auf Basis von Daten der gesetzlichen Krankenversicherungen von 2002. Das Gesundheitswesen. 2009;71(05):281-90.
25. Schnee M. Sozioökonomische Strukturen und Morbidität in den gesetzlichen Krankenkassen. Gesundheitsmonitor. 2008:88-104.
26. Schubert I, Köster I, Küpper-Nybelen J, Ihle P. Versorgungsforschung mit GKV-Routinedaten. Bundesgesundheitsblatt-Gesundheitsforschung-Gesundheitsschutz. 2008;51(10):1095-105.
27. OpenStreetMap. [cited 2016 17. Mai]. Available from: https://download.geofabrik.de/.
28. Schäfer T, Pritzkuleit R, Jeszenszky C, Malzahn J, Maier W, Günther K, et al. Trends and geographical variation of primary hip and knee joint replacement in Germany. Osteoarthritis and Cartilage. 2013;21(2):279-88.
29. Fotheringham AS, Wong DW. The modifiable areal unit problem in multivariate statistical analysis. Environment and planning A. 1991;23(7):1025-44.
30. Tanser F, Bärnighausen T, Cooke GS, Newell M-L. Localized spatial clustering of HIV infections in a widely disseminated rural South African epidemic. International Journal of Epidemiology. 2009:dyp148.
31. Larmarange J, Vallo R, Yaro S, Msellati P, Méda N. Methods for mapping regional trends of HIV prevalence from Demographic and Health Surveys (DHS). CyberGeo: European Journal of Geography. 2011.

32. Shi X. Selection of bandwidth type and adjustment side in kernel density estimation over inhomogeneous backgrounds. International Journal of Geographical Information Science. 2010;24(5):643-60.
33. Lemke D, Mattauch V, Heidinger O, Pebesma E, Hense H-W. Comparing adaptive and fixed bandwidth-based kernel density estimates in spatial cancer epidemiology. International journal of health geographics. 2015;14(1):1.
34. Almeida MCdS, Gomes CdMS, Nascimento LFC. Spatial distribution of deaths due to Alzheimer's disease in the state of São Paulo, Brazil. Sao Paulo Medical Journal. 2014;132(4):199-204.
35. Oberwittler D, Wiesenhütter M. The Risk of Violent Incidents Relative to Population Density in Cologne Using the Dual Kernel Density Routine. Levine, N, CrimeStat II: A Spatial Statistics Program for the Analysis of Crime Incident Locations, Program Manual, Washington, district fédéral de Columbia National Institute of Justice. 2002:332.
36. Duin RPW. On the choice of smoothing parameters for Parzen estimators of probability density functions. IEEE Transactions on Computers. 1976 (11):1175-9.
37. Habbema JD, editor A stepwise discriminant analysis program using density estimetion. Compstat; 1974: Physica-Verlag.
38. Rudemo M. Empirical choice of histograms and kernel density estimators. Scandinavian Journal of Statistics. 1982:65-78.
39. Scott DW, Terrell GR. Biased and unbiased cross-validation in density estimation. Journal of the American Statistical Association. 1987;82(400):1131-46.
40. Sheather SJ, Jones MC. A reliable data-based bandwidth selection method for kernel density estimation. Journal of the Royal Statistical Society Series B (Methodological). 1991:683-90.
41. Hall P, Sheather SJ, Jones M, Marron JS. On optimal data-based bandwidth selection in kernel density estimation. Biometrika. 1991;78(2):263-9.
42. Lai P-C, So F-M, Chan K-W. Spatial epidemiological approaches in disease mapping and analysis: CRC Press; 2008.
43. Levine N. CrimeStat III: a spatial statistics program for the analysis of crime incident locations (version 3.0). Houston (TX): Ned Levine & Associates/Washington, DC: National Institute of Justice. 2004.
44. Ahmad OB, Boschi-Pinto C, Lopez AD, Murray CJ, Lozano R, Inoue M. Age standardization of rates: a new WHO standard. Geneva: World Health Organization. 2001;9.
45. Lawson A, Biggeri A, Böhning D, Lesaffre E, Viel J-F, Bertollini R. Disease mapping and risk assessment for public health: John Wiley & Sons; 1999.
46. Anselin L. Exploring spatial data with Geoda: A workbook, spatial analysis laboratory department of geography. University of Illinois, center for spatially Integrated social science. 2006.
47. Coleman M, Coleman M, Mabuza AM, Kok G, Coetzee M, Durrheim DN. Using the SaTScan method to detect local malaria clusters for guiding malaria control programmes. Malaria Journal. 2009;8(1):1-6.
48. Faruque LI, Ayyalasomayajula B, Pelletier R, Klarenbach S, Hemmelgarn BR, Tonelli M. Spatial analysis to locate new clinics for diabetic kidney patients in the underserved communities in Alberta. Nephrology Dialysis Transplantation. 2012;27(11):4102-9.
49. Kulldorff M. SaTScan user guide for version 9.4. 2015. 2016.

50. Kulldorff M. A spatial scan statistic. Communications in Statistics-Theory and methods. 1997;26(6):1481-96.
51. Chen J, Roth RE, Naito AT, Lengerich EJ, MacEachren AM. Geovisual analytics to enhance spatial scan statistic interpretation: an analysis of US cervical cancer mortality. International journal of health geographics. 2008;7(1):1.
52. Poole MA, O'Farrell PN. The assumptions of the linear regression model. Transactions of the Institute of British Geographers. 1971:145-58.
53. Haque U, Scott LM, Hashizume M, Fisher E, Haque R, Yamamoto T, et al. Modelling malaria treatment practices in Bangladesh using spatial statistics. Malar J. 2012;11(63):10.1186.
54. ESRI. How Exploratory Regression works [cited 2016 17. Mai]. Available from: http://desktop.arcgis.com/en/arcmap/10.3/tools/spatial-statistics-toolbox/how-exploratory-regression-works.htm.
55. Hu M, Li Z, Wang J, Jia L, Liao Y, Lai S, et al. Determinants of the incidence of hand, foot and mouth disease in China using geographically weighted regression models. PloS one. 2012;7(6):e38978.
56. Gebreab SY, Roux AVD. Exploring racial disparities in CHD mortality between blacks and whites across the United States: a geographically weighted regression approach. Health & place. 2012;18(5):1006-14.
57. Fotheringham AS, Brunsdon C, Charlton M. Geographically weighted regression: John Wiley & Sons, Limited; 2003.
58. Nakaya T. GWR4 user manual. WWW document, http://www st-andrews ac uk/geoinformatics/wp-content/uploads/GWR4manual_201311 pdf. 2012.
59. Curtis AJ, Lee W-AA. Spatial patterns of diabetes related health problems for vulnerable populations in Los Angeles. International journal of health geographics. 2010;9(1):1.
60. Fukuda Y, Umezaki M, Nakamura K, Takano T. Variations in societal characteristics of spatial disease clusters: examples of colon, lung and breast cancer in Japan. International Journal of Health Geographics. 2005;4(1):1.
61. Schmiedel S, Jacquez GM, Blettner M, Schüz J. Spatial clustering of leukemia and type 1 diabetes in children in Denmark. Cancer Causes & Control. 2011;22(6):849-57.
62. Schlundt DG, Hargreaves MK, McClellan L. Geographic clustering of obesity, diabetes, and hypertension in Nashville, Tennessee. The Journal of ambulatory care management. 2006;29(2):125-32.
63. Arbeit Bf. Methodische Hinweise zu sozialversicherungspflichtig und geringfügig Beschäftigten 2013 [cited 2016 May 17th]. Available from: https://statistik.arbeitsagentur.de/nn_280848/Statischer-Content/Grundlagen/Methodische-Hinweise/BST-MethHinweise/SvB-und-GB-meth-Hinweise.html.
64. Kaplan RM, Kronick RG. Marital status and longevity in the United States population. Journal of epidemiology and community health. 2006;60(9):760-5.
65. Azimi-Nezhad M, Ghayour-Mobarhan M, Parizadeh M, Safarian M, Esmaeili H, Parizadeh S, et al. Prevalence of type 2 diabetes mellitus in Iran and its relationship with gender, urbanisation, education, marital status and occupation. Singapore medical journal. 2008;49(7):571.
66. Salois MJ. Obesity and diabetes, the built environment, and the 'local'food economy in the United States, 2007. Economics & Human Biology. 2012;10(1):35-42.

67. Papas MA, Alberg AJ, Ewing R, Helzlsouer KJ, Gary TL, Klassen AC. The built environment and obesity. Epidemiologic reviews. 2007;29(1):129-43.
68. Koopman RJ, Mainous AG, Diaz VA, Geesey ME. Changes in age at diagnosis of type 2 diabetes mellitus in the United States, 1988 to 2000. The Annals of Family Medicine. 2005;3(1):60-3.
69. Dansky KH, Dirani R. The use of health care services by people with diabetes in rural areas. The Journal of Rural Health. 1998;14(2):129-37.
70. Warden CR. Comparison of Poisson and Bernoulli spatial cluster analyses of pediatric injuries in a fire district. International Journal of Health Geographics. 2008;7(1):1.
71. Olson KL, Grannis SJ, Mandl KD. Privacy protection versus cluster detection in spatial epidemiology. American Journal of Public Health. 2006;96(11):2002-8.
72. Maier W, Fairburn J, Mielck A. [Regional deprivation and mortality in Bavaria. Development of a community-based index of multiple deprivation]. Gesundheitswesen (Bundesverband der Arzte des Offentlichen Gesundheitsdienstes (Germany)). 2012;74(7):416-25.
73. Gerlach F, Greiner W, Haubitz M. Bedarfsgerechte Versorgung-Perspektiven für ländliche Regionen und ausgewählte Leistungsbereiche. Gutachten; 2014.
74. Kucharska W, Pieper J, Schweikart J. Zugang zur Kindergesundheit in Brandenburg–eine Untersuchung auf der Grundlage freier Geodaten. Angewandte Geoinformatik. 2014:282-91.
75. Bundesvereinigung K. Die neue Bedarfsplanung Grundlagen, Instrumente und regionale Möglichkeiten: Kassenärztliche Bundesvereinigung; [cited 2016 May 17th]. Available from: http://www.kbv.de/media/sp/Instrumente_Bedarfsplanung_Broschuere.pdf.

# CHAPTER 5

## Case study on Pertussis in the Netherlands

published as:

Kauhl B., Heil J., Hoebe CJPA, Schweikart J., Krafft T., Dukers-Muijrers NHTM (2017) Is the current pertussis incidence only the results of testing? A spatial and space-time analysis of pertussis surveillance data using cluster detection methods and geographically weighted regression modelling. PLoS ONE 12(3):e0172383.doi:10.1371/journal.pone.0172383

**Abstract**

**Background:** Despite high vaccination coverage, pertussis incidence in the Netherlands is amongst the highest in Europe with a shifting tendency towards adults and elderly. Early detection of outbreaks and preventive actions are necessary to prevent severe complications in infants. Efficient pertussis control requires additional background knowledge about the determinants of testing and possible determinants of the current pertussis incidence. Therefore, the aim of our study is to examine the possibility of locating possible pertussis outbreaks using space-time cluster detection and to examine the determinants of pertussis testing and incidence using geographically weighted regression models.

**Methods** :We analysed laboratory registry data including all geocoded pertussis tests in the southern area of the Netherlands between 2007 and 2013. Socio-demographic and infrastructure-related population data were matched to the geo-coded laboratory data. The spatial scan statistic was applied to detect spatial and space-time clusters of testing, incidence and test-positivity. Geographically weighted Poisson regression (GWPR) models were then constructed to model the associations between the age-specific rates of testing and incidence and possible population-based determinants.

**Results**: Space-time clusters for pertussis incidence overlapped with space-time clusters for testing, reflecting a strong relationship between testing and incidence, irrespective of the examined age group. Testing for pertussis itself was overall associated with lower socio-economic status, multi-person-households, proximity to primary school and availability of healthcare. The current incidence in contradiction is mainly determined by testing and is not associated with a lower socioeconomic status.

**Discussion**: Testing for pertussis follows to an extent the general healthcare seeking behaviour for common respiratory infections, whereas the current pertussis incidence is largely the result of testing. More testing would thus not necessarily improve pertussis control. Detecting outbreaks using space-time cluster detection is feasible but needs to adjust for the strong impact of testing on the detection of pertussis cases.

## Introduction

Pertussis is a highly infectious respiratory disease caused by *Bordetella Pertussis* and is especially severe in unvaccinated and incomplete vaccinated children (1). Despite the implementation of extensive vaccination schemes, the incidence of pertussis is increasing in many countries with a shifting tendency towards adults and elderly (2-7). In fully vaccinated children and adults with waning immunity, the symptoms are often mild and indistinguishable from other respiratory diseases (5). The clinical diagnosis of pertussis is challenging, not only because symptoms are often unspecific, but also because co-infection with respiratory diseases complicates diagnosis (5, 8, 9). Additionally, sensitivity and specificity of the applied laboratory tests are influenced by vaccination coverage, frequency of mild cases within the population, exposure to pertussis and age of the patient so that no single laboratory test can be considered as "gold standard" for confirming pertussis cases (10). The lack of universal standards to confirm pertussis infections thus further facilitates the spread of undiagnosed infections.

This is problematic, as transmission through infected, but undiagnosed members of the same household are held responsible for most transmissions to not or incomplete vaccinated infants (11). To further reduce transmission, several countries such as France, USA and Australia have incorporated adult booster doses in their respective vaccination schemes (12-15) and the Dutch health council recently recommended the introduction of maternal vaccination to the national vaccination program (16).

In the Netherlands, the pertussis incidence is amongst the highest in Europe and rates have increased since 1996 (17). The underlying reasons of this increase are not fully conclusive. Several studies attribute the increase of pertussis to a waning immunity in adults (2, 17) and new emerging strains of *Bordetella Pertussis* (18, 19). Other studies suggest that an increase of detected pertussis infections occurs mainly because of an increased awareness of the population and general practitioners (GPs) (20-22) and enhanced notification systems (21-23).

According to current general practitioner guidelines in the Netherlands, a clinical pertussis diagnosis is considered in patients having typical symptoms such as severe coughing who had contact with a proven pertussis case. Additional testing for pertussis is only recommended for patients in a household with an unvaccinated or incomplete vaccinated child younger than one year old and in households with a woman, which is

more than 34 weeks pregnant (24). For all other groups, testing is rather induced by the patient than the GP (25).

As pertussis is a notifiable disease in the Netherlands (26) and many other countries, the resulting surveillance data on testing and infections is used to monitor changes over space and time (27-29). Despite previous findings that pertussis is highly heterogeneously distributed in space as well as in space-time (30, 31), a substantial amount of current surveillance activities on pertussis is still restricted to a temporal analysis only (7, 28, 29, 32), masking important regional variations and thus complicating an effective public health response.

Geographic Information Systems (GIS) and cluster detection methods – both, purely spatial as well as in time and space have proven useful to locate possible outbreaks of infectious diseases (33-35), including pertussis (30), resulting in a timely and effective response in affected areas. Such an approach might ultimately help to minimize the spread of pertussis at an early stage when the risk of transmission is highest (36).

Efficient pertussis control however, requires additional background knowledge about the determinants of pertussis testing and the determinants of the current pertussis incidence. However, personal patient information such as occurrence of infections within the same household, household composition, vaccination status and socio-demographic characteristics are mostly unavailable on an individual level due to privacy restrictions of surveillance data (37). Detailed population information on household composition, socio-economic variables and information on available healthcare and infrastructure is in the Netherlands only available on an aggregated level such as neighbourhoods or municipalities (38, 39).

In this context, spatial regression models at the ecological level have been increasingly applied in epidemiological studies of infectious diseases in recent years as regression modelling based on aggregated population data allows an analysis of possible risk factors that are unavailable on an individual level (39-41).

Geographically weighted regression modelling (GWR) is an extension of traditional global spatial regression models and measures how the association between disease risk and socio-demographic population characteristics varies over space. This approach often led to the conclusion that the key populations for certain diseases depend largely on the place of residence, resulting in more cost-effective, targeted public

health interventions aimed at those groups who are most at risk in specific locations (39, 42, 43).

This approach has shown to be effective in revealing associated determinants of several infectious diseases such as Hepatitis C (39), HIV (44) and Japanese encephalitis (42) but has also been useful to examine determinants of treatment seeking behaviour i.e. for Malaria (45) and could thus provide a feasible basis to examine the determinants of pertussis testing and the associated determinants of the current pertussis incidence.

The aim of our article is therefore (i) to examine spatial and space-time clustering of pertussis testing, incidence and test-positivity and (ii) to model the associations between socio-demographic, healthcare and infrastructure related determinants and pertussis testing and incidence using geographically weighted regression models.

## Data and Methods

### Ethics Statement

The medical ethics committee of the Maastricht University Medical Centre (Maastricht, the Netherlands) approved the study (11-4-136) and waived the need for consent to be collected from participants. Since retrospective data originated from standard care (in which one can opt-out for the use of their data for scientific research) and were analysed anonymously, no further informed consent for data analysis was obtained.

### Laboratory Data

Pertussis laboratory data were collected between Jan. 1st 2007 and Dec. 31st 2013 in the province Limburg, the Netherlands, comprising a population of 1,121,820 inhabitants (46). Testing for pertussis was performed by GPs and hospitals in the area and the test-samples were then sent to the six laboratories in the region, which are capable of analysing the collected samples. The data for this study were therefore retrieved from the registries of these six laboratories in the province and included all pertussis tests requested by health care providers. In total, the data consisted of 15,429 tested persons of which 3,312 (21.5%) persons were tested positive. Positivity was either based on the test result of the PCR (5.5% of tests), culture (2.4% of tests) or serology (IgG, 92.2% of tests); for the latter the international standard cut-off value of IgG≥62.5 IU/ml or IgG≥13 VU/ml was used to be a sensitive and specific indicator of a pertussis infection in the past year (47, 48). This standardisation was made, instead of

using the laboratory interpretation, because laboratories used test cut-offs that differed between the laboratories and over time. When multiple serology was applied to test a person and test results were inconsistent, the standardised test result was positive when seroconversion occurred from a negative test to a positive test result. In 8.8% (1,361) of the tested persons, standardisation was not possible because IgG-titres were unavailable for the serological test. We were able to filter the laboratory result for most of these tests, but for a total of 1.0% (151) the test result remained missing.

Besides the results of the laboratory diagnosis, the available information included the four-digits postal code, sex, age and date of testing. In total 14,810 tested persons (96.0%) and 3,150 positive persons (95.1%) had valid postal codes assigned and were therefore included in our analyses.

**Outcomes**

As outcomes, we examined three different rates: (i) the proportion of tested persons per inhabitants (testing); (ii) the incidence of pertussis expressed as proportion of positive tested persons per inhabitants (incidence) and (iii) the proportion of positive to tested persons (test-positivity). Due to the different vulnerability between age groups (7), four demographic strata were used to calculate the rates outlined above: (i) children aged 0 – 14, (ii) adults between 15 – 64, (iii) persons aged 65 and older and (iv), total population. Although infants display the highest vulnerability to pertussis (1), an analysis of pertussis testing and incidence among 0-5 year olds was not advisable due to the low number of tests and infections in this age group.

**Explanatory Variables**

We assessed several socio-demographic, infrastructure and healthcare-related variables for their association with the proportion of tested persons and pertussis incidence among the different age strata. The data and map sources were available from Statistics Netherlands (49). The data were available per neighbourhood and had to be aggregated to the four-digits postal codes to match the pertussis laboratory data. In the Netherlands, a neighbourhood is an administrative area within a municipality with a homogenous socio-economic structure (38, 39, 49). Due to privacy restrictions of statistics Netherlands, data for each variable is only reported for a minimum of inhabitants or households. For example, the number of persons in a specific age group is only reported for neighbourhoods with more than 50 persons, while average income is

only reported for neighbourhoods with more than 100 persons (38, 39, 49). The population-weighted aggregation was therefore only based on the neighbourhoods, for which data were made available. The age stratified population data, which were used for the calculation of the proportion of tested persons and the Pertussis incidence, was only available for 2013 from the Central Bureau for statistics Netherlands. As transmission occurs mainly between members of the same households (11), testing between different demographic strata is expected to follow a comparable pattern. We therefore included testing among the other age groups as independent variables in a regression analysis for testing in a specific age group. To analyse the determinants of the incidence, we included the rate of tested persons within the same age group and the rates of infected persons in the other age groups as independent variables. This should in a later stage confirm through GWR in which areas the effect of testing and potential intra-household transmission is stronger than in other areas.

**Exploratory Disease Mapping**

As our study aimed to highlight geographic heterogeneity among different demographic strata and different epidemiological outcomes on a small spatial scale, the numbers in the numerator for each examined demographic strata and epidemiological outcome can be considered as fairly low. This leads to a large variance of the respective rate simply due to varying population densities of relatively arbitrary administrative boundaries. As pertussis testing and infections are highly depending on local characteristics, neighbouring areas can therefore be expected to display similar rates and abrupt changes are more likely to occur due to the effect of relatively arbitrary administrative boundaries - also referred to as the modifiable area unit Problem (MAUP) (50). We therefore applied a local empirical Bayesian smoothing approach where the respective rates are smoothed towards a local mean. The neighbours were defined as an area sharing a common edge or boundary (51). The analysis was carried out in GeoDa 1.2.0 (52). The resulting rates were then imported in ESRI ArcGIS 10.2.

**Local Cluster Detection**

**Purely Spatial Cluster Detection**

The spatial scan statistic is a local cluster test, which identifies the geographic location and statistical significance of local clusters (39, 53, 54). The rationale behind a cluster analysis in our study was to detect significant local clusters of high rates within

one demographic stratum and compare the location of clusters within the other demographic strata. In this study, we used two different models in SaTScan: For the proportion of tested persons and the incidence of pertussis, we used a purely spatial Poisson model (39, 55). The input data for this model consisted of the number of tested persons / positive tested persons and the population per demographic stratum as well as the centroid coordinates of each postal code (56). For the proportion of positive tested persons, we used a purely spatial Bernoulli model (54, 57). For this model, the input data consisted of the number of positive tested persons, the number of negative tested persons per demographic stratum and the centroid coordinates for each postal code (56). The spatial scan statistic then uses a circular scanning window, which is flexible in size up to a user-specified maximum or the standard setting of including up to 50% of the population inside a cluster. The scanning window gradually moves over the coordinates over the study area, evaluating all possible cluster locations and sizes. The statistical significance is evaluated by computing 999 Monte-Carlo replications (58). In our study, we set the maximum population at risk to be included in a possible cluster not to exceed 5%. This was done since the default settings in SaTScan are more likely to produce very large clusters and therefore contain locations of low relative risk simply because of the circular scanning window (59). The value of 5% of the maximum population at risk was based on the experience of a previous study in the area, which met our criterion of including only locations of elevated risk in a cluster (39).

**Space-time Cluster Detection**

A space-time cluster analysis was employed in this study to evaluate whether testing, incidence and test-positivity occur at the same geographical locations and in the same time periods, providing background information whether space-time clusters of pertussis incidence are correlated in space-time with testing.

The space-time cluster analysis in SaTScan is comparable to a purely spatial model, except that the scanning window may be represented as a cylinder, where the base of the cylinder represents the geographic location and the height of the cylinder represents the time component of the scanning window. The scanning window then moves over all centroid coordinates across the study area and evaluates all possible space-time clusters within the study area and study period (60). Similar to the purely spatial cluster analysis, we used for both, proportion of tested persons and incidence a space-time Poisson model and for the proportion of positive to tested persons, a space-

time Bernoulli model (61). In this study, we used a scanning window that may contain up to 10% of the background population and up to 12 months of the study period. The calculation of purely spatial and space-time clusters was carried out using SaTScan v9.4.1.

**Selection of Explanatory Variables**

To select a meaningful set of explanatory variables for the regression analysis, we used a data-mining tool called "exploratory regression" in ESRI ArcGIS 10.2. This tool is comparable to a forward step-wise regression. In each step, one additional variable is added to the regression equation and evaluated based on following criteria in our analysis: (i) The regression coefficients are statistically significant ($p<0,05$) and (ii) do not display multicollinearity (Variance Inflation Factor < 7.5) (62). We then chose a set of statistically significant explanatory variables as suggested by the exploratory regression that delivered a plausible explanation of the respective outcome (the proportion of tested persons or the pertussis incidence in the respective age group).

**Geographically Weighted Poisson Regression**

We constructed spatial regression models based on the variables suggested by the exploratory regression, which delivered a plausible explanation of the respective outcome. Global spatial regression models are often applied to determine the strength of the association between the dependent variable and a set of explanatory variables, but the obtained coefficients are averaged over the whole study area (51, 63). Our study area however, consisted of 258 postal codes and the socio-demographic composition and available infrastructure varies at local level. It is therefore unlikely that one single coefficient per explanatory variable would be a good estimator of the strength of the association for the whole study area. We therefore favoured a geographically weighted regression (GWR) approach over a global approach. Geographically weighted regression modelling measures how the relationship between a set of explanatory variables and an epidemiological outcome varies over space, resulting in more detailed understanding of the spatially varying key populations and local characteristics of the study area for different epidemiological outcomes (39, 41, 43, 45, 64).

The Poisson distribution among the available GWR models is most suitable for diseases, especially if observed counts of cases are low in certain areas (39, 65-67). The dependent variable was in the analysis for testing specified as the number of tests and in

the analysis for incidence as positive cases per postal code. The offset variable was specified as the number of inhabitants per postal code. The centroids of each postal code area were used as input coordinates. The geographically weighted Poisson regression (GWPR) calculates an additional global Poisson model to allow a comparison between a global and a local approach. The GWPR uses a kernel function to fit a regression equation for each postal code area, where the centre of the kernel is the regression point. The kernel function assigns decreasing weights to the observations, depending on the distance (bandwidth) of the respective observation to the centre. The bandwidth of the kernel in GWPR can be either fixed or adaptive and the shape of the kernel can follow a Gaussian or a bi-square distribution. The optimization of the bandwidth in a GWPR model can be based on one of the three available criteria: (i) Akaikes Information Criterion (AIC); (ii) Akaikes corrected Information Criterion (AICc) and (iii) Bayesian Information Criterion (BIC) (63, 68). We thus evaluated all 12 possible combinations of kernel shape, bandwidth type and bandwidth optimization method for the eight different dependent variables. The models without clustered residuals were further considered and out of those, the models with the lowest AICc value and highest adjusted $R^2$ were then chosen as the final models. The statistical significance of each coefficient per postal code was calculated using pseudo t-values (63). The statistic behind GWPR is described in detail elsewhere (63).

We assessed clustering of the residuals of the GWPR using the global Moran`s I test in ESRI ArcGIS 10.2. The computation of GWPR was carried out in the GWR4 software (68). The coefficients were standardized to allow a direct comparison of the strength of association among the examined explanatory variables. To enhance visualization of the spatially varying coefficients, we used the software`s "prediction at non-sample points" function and calculated the predicted values for a grid of Limburg based on a cell size of 100m x 100m. The obtained values were then interpolated using ordinary kriging in ESRI ArcGIS 10.2.

## Results

### Purely spatial analysis

### Testing

The spatial distribution of the proportion of tested persons among the different demographic strata displayed strong local variations and local clustering (Fig 1a). The

proportion of tested persons differed widely between the examined age groups and is highest in children (Table 1).

All demographic groups displayed strong local clustering. Despite the differences in the proportion of tested persons, local clustering displayed similar patterns among the different demographic strata. Local clusters were observed especially in the central and northern parts of the study area. In the southern part in contradiction, no clusters could be observed.

**Incidence**

The incidence of pertussis varied between the different demographic groups and was also highest in children (Table 1). Overall, the clusters of the pertussis incidence followed closely the locations of the observed clusters for testing (Fig 1b), reflecting a strong spatial correlation to the patterns of testing. The clusters of pertussis incidence among children partially overlapped with clusters among adults and the clusters among adults overlapped in certain areas with those for seniors.

**Test-positivity**

Test-positivity was again highest in children, but the positivity rate did not differ much between adults and seniors (Table 1). Test-positivity in children was the only rate that did not display any local clusters (Fig 1c). Among adults, seniors and the total tested persons, only a small number of clusters were observed. Except for one cluster in the southwestern part of the study area for the total tested persons, the clusters observed for test-positivity overlapped with the clusters for testing and incidence.

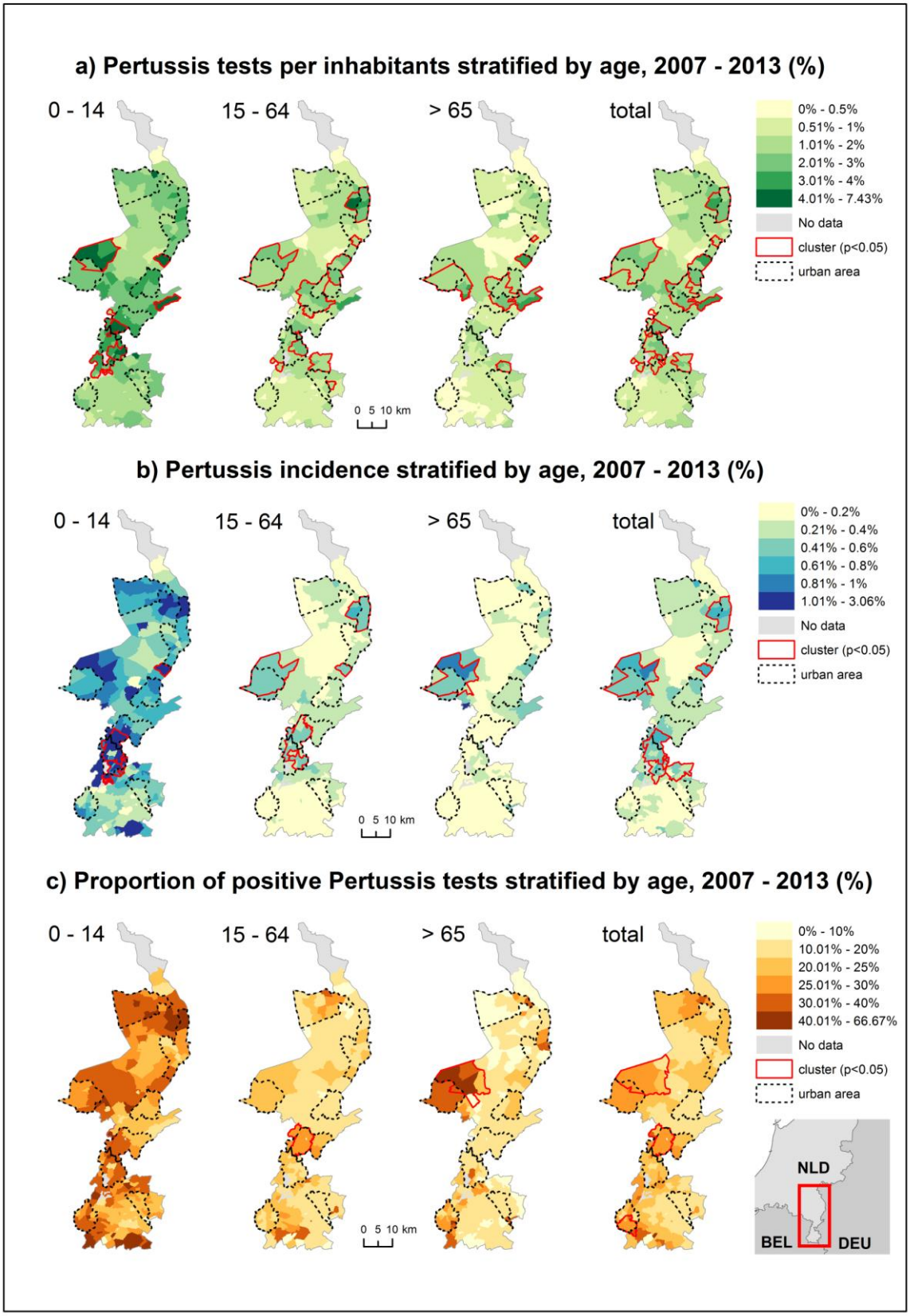



**Fig 1:** Spatial distribution of a) pertussis testing, b) incidence and c) test-positivity, 2007 - 2013

**Table 1:** Rates of testing, test positivity and population incidence. SD = standard deviation

| Age | Tested (%) | | Positive tested (%) | | Incidence (%) | |
|---|---|---|---|---|---|---|
| | Mean | SD | Mean | SD | Mean | SD |
| 0 – 14 | 2.39 | 1.52 | 29.77 | 22.83 | 0.72 | 0.70 |
| 15 -64 | 1.27 | 0.75 | 18.45 | 13.20 | 0.23 | 0.16 |
| >65 | 0.92 | 0.93 | 18.22 | 17.49 | 0.16 | 0.24 |
| All ages | 1.36 | 0.72 | 21.27 | 14.09 | 0.29 | 0.22 |

**Space-time Analysis**

**Testing**

In all demographic groups, space-time clustering started generally in the beginning of 2012 and lasted partially until the beginning of 2013 (Fig 2a). Only in children, one cluster in 2007 and one cluster starting in 2011 could be observed. Space-time clustering for testing thus started relatively uniform across the study area in the beginning of 2012.

**Incidence**

In adults, seniors and among the total population, the space-time distribution of clusters for the pertussis incidence followed closely the space-time distribution of clusters for testing (Fig 2b). The majority of clusters were observed in the beginning of 2012, lasting partially until the beginning of 2013 and were located in the same locations as space-time clusters for testing. In children however, space-time clusters were also observed in 2007 and 2009, for which no space-time clusters for testing were observed.

**Test-positivity**

The distribution of space-time clusters for positivity differed strongly from the observed clusters from testing and pertussis incidence (Fig 2c). In children, a cluster observed in 2009 in the centre of the study area overlapped with a cluster for pertussis incidence. In seniors, a cluster observed in 2012 in the northwestern part overlapped

with the clusters for testing and pertussis incidence. The other clusters were scattered over the entire study area and study period.

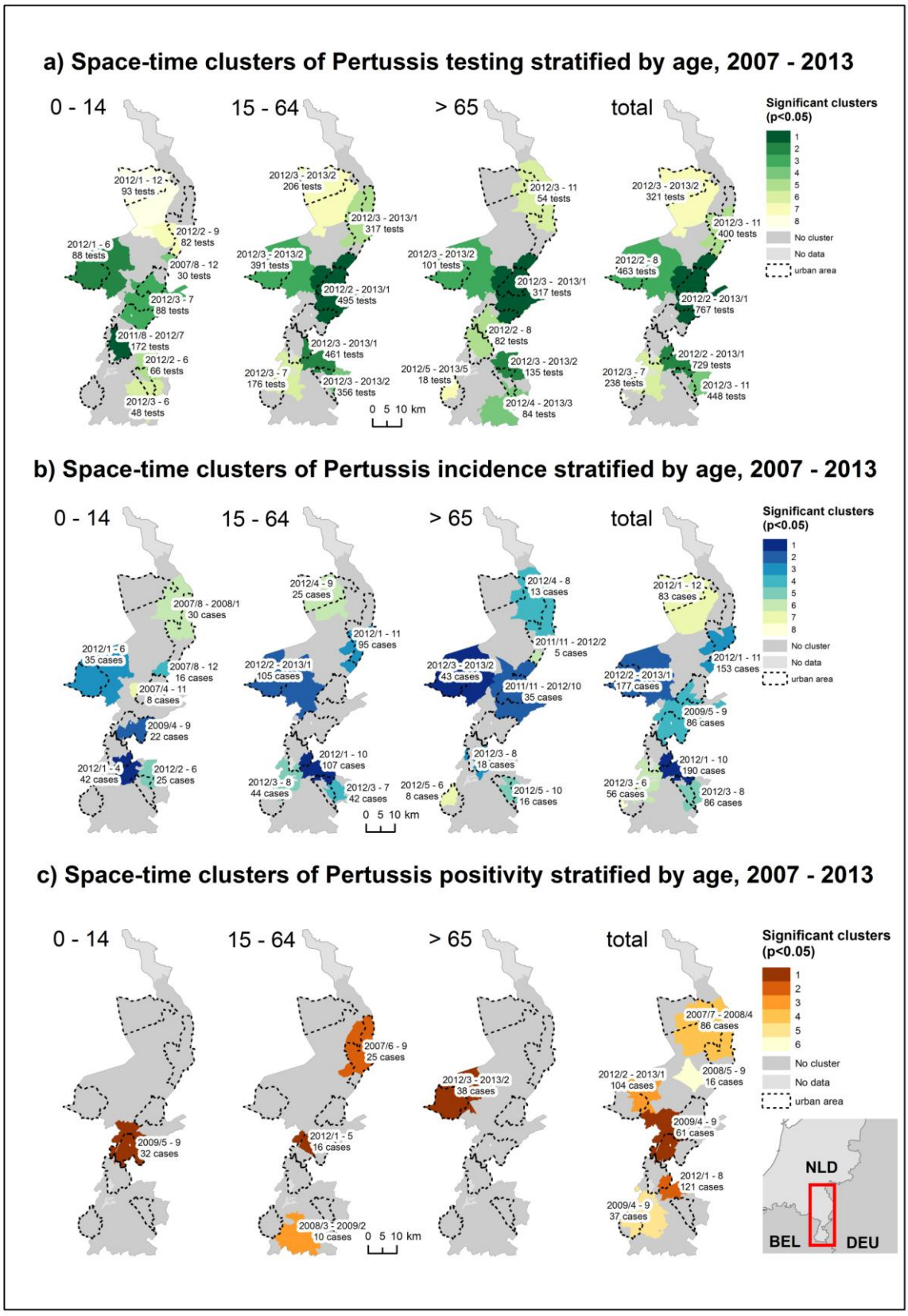



**Fig 2:** Space-time clusters of a) testing, b) incidence and c) test-positivity, 2007 - 2013

## Geographically Weighted Poisson Regression

### Determinants of Testing

For pertussis testing, the Gaussian kernel type using a fixed, AICc optimized bandwidth fulfilled the requirements of the residuals not displaying spatial autocorrelation among all evaluated demographic groups (Table 2). The local models generally outperformed the global models, reflecting important local differences in the associations between testing and the examined explanatory variables.

**Testing in children**: For testing in children, testing in adults was the strongest predictor, followed by a moderate association to testing in seniors. The negative association to unmarried persons reflects that parents, which have ever been married, are a determinant for testing in children. The negative association to mean property value indicates that children in deprived neighbourhoods are more likely to get tested for pertussis.

**Testing in adults:** Testing in children and seniors had the strongest impact on testing in adults. The positive association to household size indicates that adults in multi-person households are more likely to get tested. Similar to testing in children, mean property value was negatively associated. Additionally, testing in adults was associated with proximity to primary schools and GPs, but was also associated with increasing distance to pharmacies.

**Testing in seniors:** Testing in adults had the strongest impact on testing in seniors. The strength of the association to testing in children was relatively weak. Additionally, testing in seniors was negatively associated with proportion of households with high income.

**Testing in the total population:** The strongest predictor for testing among the total population was the positive association to household size, followed by the negative associations to mean property value, proximity to GPs and primary schools. The proportion of children and distance to pharmacies were overall positively associated. However, none of the predictors was significant in the entire study area. The results of the GWPR model additionally point out important local variations of the association between testing among the total population and the examined explanatory variables (Fig 3).

**Table 2:** Poisson regression models for pertussis testing, stratified by age. Significance levels: * = 0.05; ** = 0.01; *** = 0.001. Only significant predictors are reported

| | **Standardized coefficients of pertussis testing** | | | |
|---|---|---|---|---|
| **Variable** | **0 - 14** | **15 - 64** | **>65** | **Total** |
| Tested 0 – 14 (%) | | 0.2033*** | 0.0818** | |
| Tested 15 – 64 (%) | 0.2655*** | | 0.3874*** | |
| Tested seniors (%) | 0.0659** | 0.2360*** | | |
| Children (%) | | 0.0610*** | | 0.0525*** |
| Unmarried (%) | -0.0418* | | | |
| One person households (%) | | -0.0769*** | | -0.1567*** |
| property value (Euro) | -0.0576*** | -0.0419*** | | -0.1380*** |
| HH with high income | | | -0.0750** | |
| Dist. Prim. school (km) | | -0.0354* | | -0.0973*** |
| Dist. GP (km) | | -0.1930*** | | -0.1367*** |
| Dist. Pharmacy (km) | | 0.0700*** | | 0.0441** |
| | | | | |
| AICc global | 608 | 1151 | 523 | 2722 |
| Deviance expl. global | 0.35 | 0.51 | 0.40 | 0.17 |
| AICc local | 543 | 697 | 469 | 925 |
| Deviance expl. local | 0.44 | 0.83 | 0.49 | 0.81 |
| Moran`s I of residuals | I=0.06; p>0.05 | I=-0,00; p>0,05 | I=0.03; p>0.05 | I=-0.01; p>0.05 |

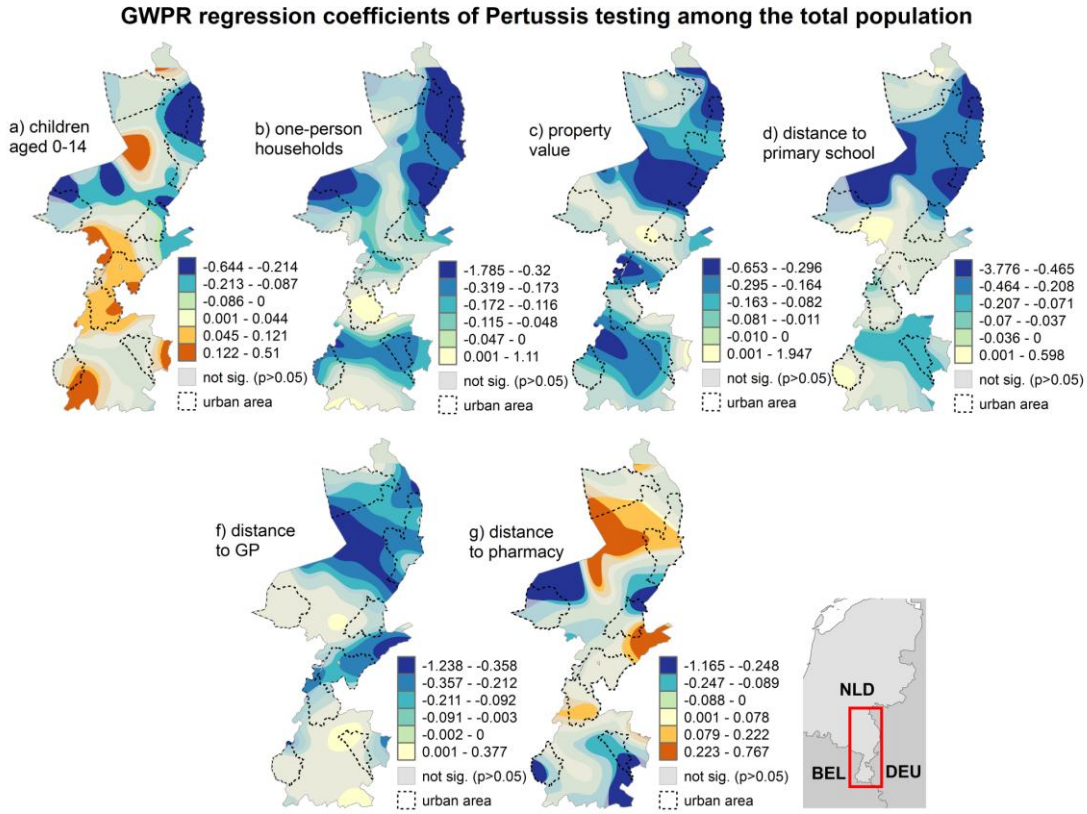



**Fig 3:** Results of the geographically weighted Poisson regression of pertussis testing among the total population

## Determinants of Pertussis Incidence

For modelling the pertussis incidence, the Gaussian kernel type using a fixed, AIC optimized bandwidth fulfilled the requirements of the residuals not displaying spatial autocorrelation among all evaluated demographic groups (Table 3). The local models outperformed the global models, although the difference was not as pronounced as the difference between global and local models in testing.

**Incidence in children:** In children, testing was the only significant predictor for the pertussis incidence.

**Incidence in adults:** In adults, testing was by far the strongest predictor. The strength of the association to the incidence in children and seniors was comparably small. One-person households were – similar to testing in adults – negatively associated. A significant association to proximity to hospitals could also be observed.

**Incidence in seniors:** Testing was also the strongest predictor in seniors, followed by the incidence in adults.

**Incidence in the total population:** Among the incidence for the total population, testing had again the strongest impact on the pertussis incidence. The strength of the negative association to households with low income and proximity to hospitals was comparatively weak. The visualized results of the GWPR model point out that testing was significant in the total study area, despite strong local differences in the strength of this association. The negative association to households with low income and proximity to hospitals was only significant in specific areas (Fig 4).

**Table 3:** Poisson regression models of pertussis incidence stratified by age. Significance levels: * = 0.05; ** = 0.01; *** = 0.001. Only significant predictors are reported

| | **Standardized coefficients of pertussis incidence** | | | |
|---|---|---|---|---|
| **Variable** | **0 - 14** | **15 - 64** | **>65** | **Total** |
| Tested 0 – 14 (%) | 0.5191*** | | | |
| Tested 15 – 64 (%) | | 0.4425*** | | |
| Tested seniors (%) | | | 0.5591*** | |
| Tested total (%) | | | | 0.4692*** |
| Incidence 0 – 14 (%) | | 0.0819** | | |
| Incidence 15 – 64 (%) | | | 0.3093*** | |
| Incidence senior (%) | | 0.0654** | | |
| Children (%) | | | - | |
| One pers. hh (%) | | -0.0975*** | | |
| Immigrants (%) | | | | |
| property value (Euro) | | | | |
| HH with low income (%) | | | | -0.0546** |
| Dist. Kinder garden (km) | | | | |
| Dist. Pharmacy (km) | | | | |
| Dist. Hospital (%) | | -0.1481*** | | -0.0676* |
| | | | | |
| AIC global | 273 | 285 | 275 | 375 |
| Deviance expl. global | 0.53 | 0.60 | 0.36 | 0.65 |
| AIC local | 268 | 280 | 249 | 309 |
| Deviance expl. local | 0.55 | 0.65 | 0.47 | 0.78 |

| | | | | |
|---|---|---|---|---|
| Moran`s I of residuals | I=0.02; p>0.05 | I=-0.00; p>0.05 | I=-0.04; p>0.05 | I=-0.01; p>0.05 |

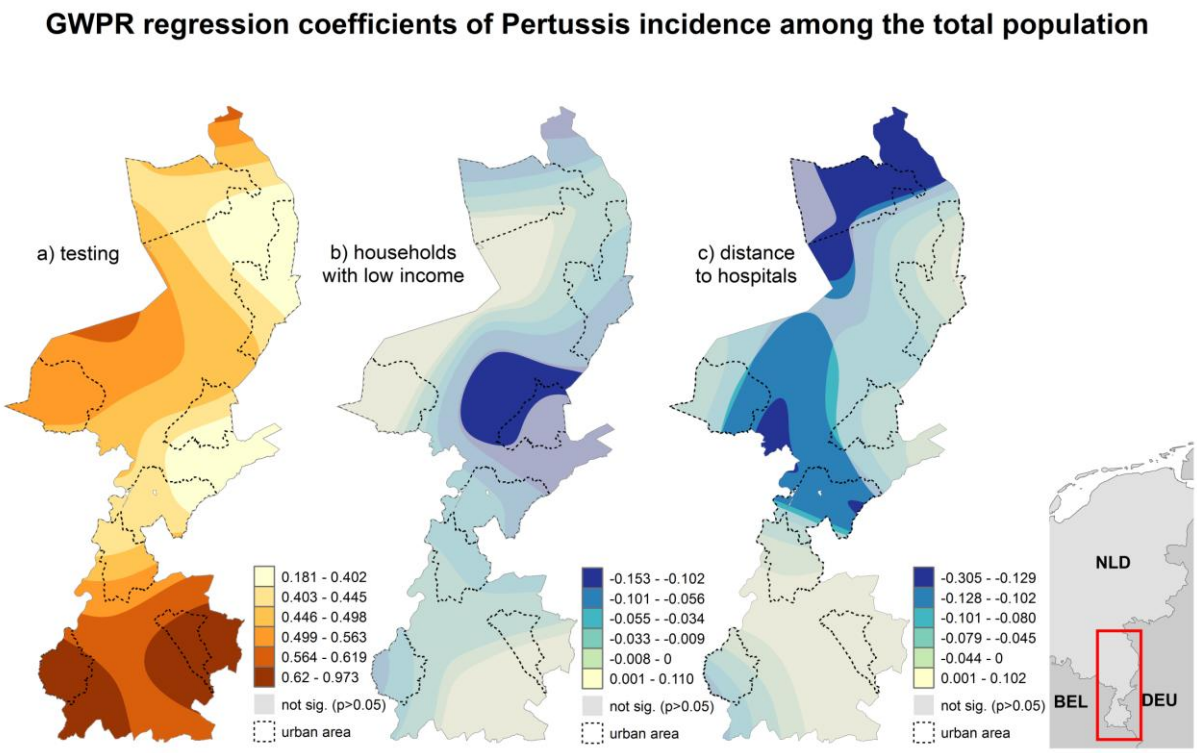



**Fig 4:** Results of the geographically weighted Poisson regression of pertussis incidence among the total population

## Discussion

The main findings of this study are: (i) The determinants of pertussis testing reflect to a certain extent the healthcare seeking behaviour of the general population for common respiratory infections; (ii) the current pertussis incidence is mainly determined by testing and (iii) outbreak detection for pertussis is feasible using surveillance data but needs to adjust for the strong impact of testing.

### Determinants of Pertussis Testing

Pertussis testing follows to an extent the healthcare seeking behaviour of the general population for common respiratory infections. This can be seen by the associations of proximity to GPs and distance to pharmacies as both compliment each other. Especially in areas where distance to GPs was negatively associated, distance to

pharmacies was positively associated. Self-medication for common respiratory infections resulting in coughing is typical in countries where over the counter medication is available (69). Our results thus could indicate that in areas, where a pharmacy is easily accessible, self-medication for coughing is the first option to choose, whereas in areas where a physician is close, patients with symptoms related to coughing are more likely to consult a GP (70). As a result, testing for pertussis is to an extent related to treatment-seeking behaviour for common respiratory infections. The association to proximity to primary schools in certain areas indicates potentially an increased awareness of pertussis vulnerability among parents and GPs for children enrolled in primary schools (71). An overall positive association of pertussis testing among the general population to the proportion of children within the study area was expected as the incidence of pertussis in children is generally higher than in other age groups (72). The negative association to unmarried persons and mean property value both display a comparable spatial pattern. This could indicate that in families, where the parents have ever been married and are located in a more deprived area are tested more frequently for pertussis, at least in the areas where both associations were significant. Previous studies noted a lower participation rate in childhood immunization schemes among persons with a lower socio-economic status (73). These associations could therefore be seen as an indicator for an increased awareness of a higher vulnerability to pertussis among this population group, although it has to be noted that the vaccination coverage in the Netherlands is generally very high with approximately 95% (74). Additionally, higher healthcare use is typical for persons living in deprived areas (75).

When analysing the determinants of pertussis testing in specific age groups, we found that testing in the adjacent age group had the strongest impact on testing in the examined age group. This strong association between the examined age groups corresponds well to previous studies suggesting that intra-household transmission is the most likely route of infection (11, 76) and thus logically leads to testing among the same households.

**Determinants of Pertussis Incidence**

The spatial and space-time distribution of clusters for pertussis incidence followed closely the spatial and space-time distribution of pertussis testing, reflecting a strong spatial and space-time correlation between testing and incidence.

We found evidence that space-time clustering of testing for pertussis and the pertussis incidence itself increased across the total study area in the first quarter of 2012, irrespective of the examined demographic group. An increase of diagnosed pertussis cases between 2011 and 2012 has also been noted by the US (77), Great Britain (78) and Spain (79). When we searched for newspaper articles related to pertussis in the Netherlands within the NexisLexis database (www.nexislexis.com), we found that the number of newspaper articles related to pertussis increased sharply in 2012 as compared to the previous years.

Other studies also suggest a "positive feedback loop" where an increase of diagnosed pertussis cases subsequently causes an increase in testing, which in turn results in an increase of diagnosed pertussis cases again (80, 81). In our study area, the strong increase of testing and pertussis incidence in 2012 could be the result of a combination of several factors; an increased awareness of GPs due to the noted increase in other countries (77-79), stronger media attention and the noted "positive feedback loop" (80, 81). The effect of the "positive feedback loop" is indicated by the fact that space-time clusters for testing and incidence occurred mostly in the same areas and time periods across all evaluated age groups. These findings highlight the importance of monitoring changes in pertussis testing in addition to monitoring the pertussis incidence.

The GWPR analyses confirmed what became already apparent during the spatial and space-time analyses. Regardless of age group, testing was the most important predictor of the pertussis incidence. The second most important predictor for the pertussis incidence in a specific age group was the incidence of pertussis in the adjacent age groups. As our analysis is based on aggregated data, we see these associations as an indicator for intra-household transmission (11, 76).

In children, testing was the only predictor for pertussis. No other variables were found to be significant in a regression analysis. Thus, the current incidence among the most vulnerable group does not exhibit any socio-economic associations and can be explained by testing only. In adults, we found an additional association to proximity to hospitals. Increased exposure to pertussis and outbreaks among persons employed in hospitals has been previously noted in several countries such as the US (82) and France (83). The results of our analysis for adults thus indicate a possible exposure to pertussis in our study area. The positive association to household-size indicates, that for adults living in multi-person households, intra-household transmission might be an important

risk factor as well, despite the findings of previous studies that intra-household transmission occurs mainly from parents to children (11, 76).

The results of the GWPR model for the pertussis incidence among the total population confirm that testing is the most important predictor for pertussis, despite local differences in the strength of this association. The negative association to households with low income in a small part of the study area was the only significant variable related to socio-economic status. An overall assumption that pertussis is related to specific socio-economic characteristics is therefore not possible. Similar to adults, an association to proximity to hospitals was observed in a small part of the study area. However, further research on an individual level would be necessary to confirm whether an increased pertussis exposure in hospitals exists in this area.

**Outbreak Detection using Space-time Cluster Detection**

Our results clearly demonstrate that space-time cluster detection is feasible using surveillance data and the methods described in this study. However, in light of the similar increase in testing and incidence in 2012 and its potential association to an increased awareness of the population and GPs, it is necessary to adjust for the impact of testing. This is reflected by the fact that we mostly detected space-time clusters for pertussis incidence in 2012, where the majority of clusters overlap with those for testing. We found only in children clusters in earlier years in 2007 – 2009, which did not overlap with clusters for testing. It is clear that the incidence of pertussis has to be strongly correlated to testing, no matter which algorithm is used. Logically, pertussis incidence is not suitable in our study area to detect outbreaks. By using test-positivity as indicator however, we could locate space-time clusters within the whole study area and study period, which were not the results of testing only. Test-positivity as indicator is to a certain extent adjusted for testing in a space-time analysis as a similar increase in positive cases and tested persons would not increase the ratio of positive cases to tested persons in the three-dimensional cylindrical scanning window used by the spatial scan statistic. Our results clearly demonstrate that test-positivity could be a better indicator for locating outbreaks, which are not the results of testing only. This conclusion is supported by previous studies, which evaluated the temporal correlation of pertussis incidence to testing behaviour over time (80, 81).

When surveillance data – as in our study - are used to locate outbreaks, a prospective space-time cluster analysis using daily updated surveillance data would

allow a detection of recent outbreaks in an automated surveillance system as early as possible. Prospective space-time cluster detection has been widely applied in syndromic surveillance (84, 85) and an automated, space-time analysis could be implemented for pertussis as well (86).

### Implications for Pertussis Surveillance

The surveillance data for pertussis in our study were collected by the respective local laboratories and had to merged manually for our study area to allow a retrospective analysis for a whole province. Surveillance of pertussis would thus greatly benefit from an automated near real-time data transfer of each respective laboratory to a centralized institution to allow a detailed, prospective analysis of possible disease outbreaks as early as possible. However, such an approach should focus on a day-wise analysis rather than a month-wise analysis as employed in this study to capture possible outbreaks as precisely as possible (35).

### Strengths and Limitations of this Study

#### Strengths

First, a major strength of this study is that the design of this study can be repeated for the whole country to confirm whether testing is also in the whole of the Netherlands the most important predictor of pertussis.

Second, the majority of studies on pertussis focus on a purely temporal analysis (7, 28, 29, 32). Our approach of analysing the spatial and space-time distribution of pertussis therefore provides a novel level of detail as our approach of detecting clusters in space as well as in space-time could allow a more focused detection of outbreaks, resulting in a more timely and cost-effective response.

Third, we analysed the spatial distribution of pertussis testing and incidence at the smallest possible spatial scale, for which surveillance data and population data can be combined without violating privacy restrictions of surveillance data. The four-digit postal code areas are very suitable for small-scale spatial-epidemiological analyses in the Netherlands (38) and thus allow a very detailed analysis of the determinants of pertussis testing and incidence.

Fourth, the use of GWPR allowed us to examine the spatially varying associations between the evaluated outcomes (testing and incidence) and the examined predictor variables. We could therefore clearly see in which areas several predictors such as

distance to GPs, pharmacies and hospitals were significant. Our results can therefore be used to allow more targeted investigations - for example - if working in hospitals in the areas outlined in our study comprises an exposure factor on an individual level as well.

Fifth, our study highlighted that the detection of outbreaks using space-time cluster detection is feasible, when using test-positivity as indicator.

**Limitations**

First, we used the international standard IgG cut-off value of 62.5 to consider a diagnosed pertussis infection. As a consequence, the incidence in our study does not necessarily reflect the notification data, which is based on laboratory interpretation.

Second, our analysis was based on aggregated data. The associations detected in our study may not necessarily reflect determinants of pertussis testing and incidence on an individual level. Given the current privacy protection of surveillance data (37), the ecological analysis employed here allowed us to analyse potential determinants that are unavailable on an individual level. Therefore, further research is necessary to confirm whether the associations captured in the ecological analysis really reflect associations on an individual level as well.

Third, we may have missed associations simply because several variables were not available within the data of Statistics Netherlands such as educational level. Lower educational level has been shown to be an important risk factor for poorer health outcomes (87) and could thus also be a possible risk factor for pertussis as well.

Fourth, the GWR4 software currently allows only a purely spatial geographically weighted regression but does not account for space-time differences (63). Given the strong temporal component of Pertussis testing and infections, a geographically weighted temporal regression might yield more detailed results. However, given the mathematical complexity of such a model (88), this approach could not be implemented in this study.

Fifth, we could not account for the role of vaccinations in our study, as this information is not available in the examined surveillance data. Potentially, data on vaccination among children and booster doses among adults could have a strong impact on the serological test results presented here and could be of additional value for future analyses.

Sixth, we could not verify whether the space-time clusters detected for pertussis incidence or test-positivity capture true outbreaks. However, we think it is reasonable to

assume that space-time clusters for the pertussis incidence, which are the results of testing only, are not of interest for the detection of outbreaks. Ultimately, verifying the detected space-time clusters with known outbreaks should be discussed with public health officials residing in the area.

Seventh, we performed a month-wise space-time cluster analysis because a day-wise analysis did consume considerable computation time (35). Our results are therefore not as temporally precise as it would be technically possible with the surveillance data of our study.

Eight, it would be useful to compare both approaches to detect outbreaks – purely temporal algorithms with the results of the space-time cluster analysis - to evaluate (i) which method allows a more focused outbreak detection and (ii), to evaluate whether the purely temporal algorithms or the space-time cluster detection delivers fewer false alarms. However, such a comparison was beyond the scope of this paper.

## Conclusion

We found empirical evidence that testing for pertussis and the current pertussis incidence in our study area increased similar to other countries in 2012. Testing was higher in deprived areas and was also associated with proximity to primary schools and availability of healthcare. Testing for pertussis thus reflects to an extent the overall healthcare seeking behaviour for common respiratory infections. The pertussis incidence in our study area is largely the result of testing and does not display a specific socio-economic association. The main population at risk for pertussis remains therefore still unknown. Given the strong association of detected pertussis cases to testing, it is questionable whether more testing would enhance pertussis control.

The detection of outbreaks using space-time cluster detection is feasible and could help to facilitate an early and appropriate public health response. However, this approach has to be adjusted for the strong dependency to testing and is probably most efficient when using test-positivity as indicator.

## Acknowledgements

We wish to thank Peter Jacobs (North Limburg Public Health Service, Venlo), Inge van Loo (Maastricht University Medical Centre, Maastricht), Jos Bus (Zuyderland Medical Centre, Heerlen), Dick van Dam (Zuyderland Medical Centre, Sittard-Geleen), Thera Trienekens (VieCuri Medical Centre, Venlo) and Trix van Dijke (Laurentius

Hospital, Roermond and St. Jans Gasthuis, Weert) for involvement in data collection. We would like to thank Inge van Loo (Maastricht University Medical Centre, Maastricht), Alexandra Ziemann (Maastricht University), Jochen Cals (Maastricht University) and Henriëtte ter Waarbeek (South Limburg Public Health Service, Geleen) for their contribution to the study proposal

**References:**


1. Parkhill J, Sebaihia M, Preston A, Murphy LD, Thomson N, Harris DE, et al. Comparative analysis of the genome sequences of Bordetella pertussis, Bordetella parapertussis and Bordetella bronchiseptica. Nature genetics. 2003;35(1):32-40.
2. Sizaire V, Garrido-Estepa M, Masa-Calles J, Martinez de Aragon MV. Increase of pertussis incidence in 2010 to 2012 after 12 years of low circulation in Spain. Euro surveillance : bulletin Europeen sur les maladies transmissibles = European communicable disease bulletin. 2014;19(32). PubMed PMID: 25139074.
3. Jenkinson D. Increase in pertussis may be due to increased recognition and diagnosis. Bmj. 2012;345:e5463. PubMed PMID: 22915721.
4. Vitek CR, Pascual FB, Baughman AL, Murphy TV. Increase in deaths from pertussis among young infants in the United States in the 1990s. The Pediatric infectious disease journal. 2003 Jul;22(7):628-34. PubMed PMID: 12867839.
5. Barlow RS, Reynolds LE, Cieslak PR, Sullivan AD. Vaccinated children and adolescents with pertussis infections experience reduced illness severity and duration, Oregon, 2010–2012. Clinical Infectious Diseases. 2014;58(11):1523-9.
6. Rozenbaum MH, De Cao E, Postma MJ. Cost-effectiveness of pertussis booster vaccination in the Netherlands. Vaccine. 2012 Nov 26;30(50):7327-31. PubMed PMID: 22749838.
7. van der Maas NA, Mooi FR, de Greeff SC, Berbers GA, Spaendonck MA, de Melker HE. Pertussis in the Netherlands, is the current vaccination strategy sufficient to reduce disease burden in young infants? Vaccine. 2013 Sep 23;31(41):4541-7. PubMed PMID: 23933365.
8. Zouari A, Smaoui H, Kechrid A. The diagnosis of pertussis: which method to choose? Critical reviews in microbiology. 2012;38(2):111-21.
9. Nuolivirta K, Koponen P, He Q, Halkosalo A, Korppi M, Vesikari T, et al. Bordetella pertussis infection is common in nonvaccinated infants admitted for bronchiolitis. The Pediatric infectious disease journal. 2010;29(11):1013-5.
10. Tozzi AE, Celentano LP, degli Atti MLC, Salmaso S. Diagnosis and management of pertussis. Canadian Medical Association Journal. 2005;172(4):509-15.
11. Te Beest DE, Henderson D, van der Maas NA, de Greeff SC, Wallinga J, Mooi FR, et al. Estimation of the serial interval of pertussis in Dutch households. Epidemics. 2014 Jun;7:1-6. PubMed PMID: 24928663.
12. Rozenbaum MH, De Vries R, Le HH, Postma MJ. Modelling the impact of extended vaccination strategies on the epidemiology of pertussis. Epidemiology and infection. 2012 Aug;140(8):1503-14. PubMed PMID: 22115361. Pubmed Central PMCID: 3404482.
13. Wendelboe AM, Njamkepo E, Bourillon A, Floret DD, Gaudelus J, Gerber M, et al. Transmission of Bordetella pertussis to young infants. The Pediatric infectious disease journal. 2007;26(4):293-9.
14. de Greeff SC, Mooi FR, Westerhof A, Verbakel J, Peeters MF, Heuvelman C, et al. Pertussis disease burden in the household: how to protect young infants. Clinical Infectious Diseases. 2010;50(10):1339-45.
15. Coudeville L, Van Rie A, Andre P. Adult pertussis vaccination strategies and their impact on pertussis in the United States: evaluation of routine and targeted (cocoon) strategies. Epidemiology and infection. 2008;136(05):604-20.
16. Gezondheidsraad. Vaccinatie tegen kinkhoest: doel en strategie. The Hague2015.
17. De Greeff SC, De Melker HE, Van Gageldonk PG, Schellekens JF, van der Klis FR, Mollema L, et al. Seroprevalence of pertussis in The Netherlands: evidence for increased circulation of Bordetella pertussis. PloS one. 2010;5(12):e14183.

18. He Q, Mertsola J. Factors contributing to pertussis resurgence. 2008.
19. Mooi FR. Bordetella pertussis and vaccination: the persistence of a genetically monomorphic pathogen. Infection, Genetics and Evolution. 2010;10(1):36-49.
20. Cherry JD. The science and fiction of the "resurgence" of pertussis. Pediatrics. 2003;112(2):405-6.
21. Greenberg DP. Pertussis in adolescents: increasing incidence brings attention to the need for booster immunization of adolescents. The Pediatric infectious disease journal. 2005;24(8):721-8.
22. Cherry JD. Epidemic pertussis in 2012—the resurgence of a vaccine-preventable disease. New England Journal of Medicine. 2012;367(9):785-7.
23. Galanis E, King AS, Varughese P, Halperin SA. Changing epidemiology and emerging risk groups for pertussis. Canadian Medical Association Journal. 2006;174(4):451-2.
24. Verheij TJM, Hopstaken RM, Prins JM, Salomé PL, Bindels PJ, Ponsioen BP, et al. Acuut hoesten Samenvattingskaart M78. Huisarts Wet 2011;2011(54):2.
25. Heil JLtWJH, C; HA Jacobs, P; W van Dam, D; AM Trienekens, T; HM van Loo, I; HTM Dukers-Muijerers, N. Pertussis surveillance and control in the Netherlands, 2010-2013: exploring variations and delays in GPs' diagnoses, laboratories' testing and public health services' notifications. 2016.
26. de Melker HE, Schellekens J, Neppelenbroek S, Mooi F, Rümke H, Conyn-van Spaendonck M. Reemergence of pertussis in the highly vaccinated population of the Netherlands: observations on surveillance data. Emerging infectious diseases. 2000;6(4):348.
27. Guimaraes LM, Carneiro EL, Carvalho-Costa FA. Increasing incidence of pertussis in Brazil: a retrospective study using surveillance data. BMC infectious diseases. 2015;15:442. PubMed PMID: 26498058. Pubmed Central PMCID: 4619034.
28. Choe YJ, Park YJ, Jung C, Bae GR, Lee DH. National pertussis surveillance in South Korea 1955-2011: epidemiological and clinical trends. International journal of infectious diseases : IJID : official publication of the International Society for Infectious Diseases. 2012 Dec;16(12):e850-4. PubMed PMID: 22921258.
29. Wymann MN, Richard JL, Vidondo B, Heininger U. Prospective pertussis surveillance in Switzerland, 1991-2006. Vaccine. 2011 Mar 3;29(11):2058-65. PubMed PMID: 21251904.
30. Omer SB, Enger KS, Moulton LH, Halsey NA, Stokley S, Salmon DA. Geographic clustering of nonmedical exemptions to school immunization requirements and associations with geographic clustering of pertussis. American journal of epidemiology. 2008 Dec 15;168(12):1389-96. PubMed PMID: 18922998.
31. Choisy M, Rohani P. Changing spatial epidemiology of pertussis in continental USA. Proceedings Biological sciences / The Royal Society. 2012 Nov 22;279(1747):4574-81. PubMed PMID: 23015623. Pubmed Central PMCID: 3479730.
32. Fathima S, Ferrato C, Lee BE, Simmonds K, Yan L, Mukhi SN, et al. Bordetella pertussis in sporadic and outbreak settings in Alberta, Canada, July 2004-December 2012. BMC infectious diseases. 2014;14:48. PubMed PMID: 24476570. Pubmed Central PMCID: 3931923.
33. Dhewantara PW, Ruliansyah A, Fuadiyah ME, Astuti EP, Widawati M. Space-time scan statistics of 2007-2013 dengue incidence in Cimahi city, Indonesia. Geospatial health. 2015;10(2):373. PubMed PMID: 26618319.
34. Ndiath M, Faye B, Cisse B, Ndiaye JL, Gomis JF, Dia AT, et al. Identifying malaria hotspots in Keur Soce health and demographic surveillance site in context of low

transmission. Malaria journal. 2014;13:453. PubMed PMID: 25418476. Pubmed Central PMCID: 4251691.
35. Van Den Wijngaard CC, Van Asten L, Van Pelt W, Doornbos G, Nagelkerke NJ, Donker GA, et al. Syndromic surveillance for local outbreaks of lower-respiratory infections: would it work? PloS one. 2010;5(4):e10406.
36. Control CfD, Prevention. Outbreaks of respiratory illness mistakenly attributed to pertussis--New Hampshire, Massachusetts, and Tennessee, 2004-2006. MMWR Morbidity and mortality weekly report. 2007;56(33):837.
37. Regidor E. The use of personal data from medical records and biological materials: ethical perspectives and the basis for legal restrictions in health research. Social science & medicine. 2004;59(9):1975-84.
38. Dijkstra A, Janssen F, De Bakker M, Bos J, Lub R, Van Wissen LJ, et al. Using spatial analysis to predict health care use at the local level: a case study of type 2 diabetes medication use and its association with demographic change and socioeconomic status. PloS one. 2013;8(8):e72730. PubMed PMID: 24023636. Pubmed Central PMCID: 3758350.
39. Kauhl B, Heil J, Hoebe CJ, Schweikart J, Krafft T, Dukers-Muijrers NH. The Spatial Distribution of Hepatitis C Virus Infections and Associated Determinants-An Application of a Geographically Weighted Poisson Regression for Evidence-Based Screening Interventions in Hotspots. PloS one. 2015;10(9):e0135656. PubMed PMID: 26352611. Pubmed Central PMCID: 4564162.
40. Kauhl B, Pilot E, Rao R, Gruebner O, Schweikart J, Krafft T. Estimating the spatial distribution of acute undifferentiated fever (AUF) and associated risk factors using emergency call data in India. A symptom-based approach for public health surveillance. Health & place. 2015;31:111-9.
41. Shoff C, Yang TC. Spatially varying predictors of teenage birth rates among counties in the United States. Demographic research. 2012 Sep 11;27(14):377-418. PubMed PMID: 23144587. Pubmed Central PMCID: 3493119.
42. Robertson C, Pant DK, Joshi DD, Sharma M, Dahal M, Stephen C. Comparative spatial dynamics of Japanese encephalitis and acute encephalitis syndrome in Nepal. PloS one. 2013;8(7):e66168. PubMed PMID: 23894277. Pubmed Central PMCID: 3718805.
43. Weisent J, Rohrbach B, Dunn JR, Odoi A. Socioeconomic determinants of geographic disparities in campylobacteriosis risk: a comparison of global and local modeling approaches. International journal of health geographics. 2012 Oct 13;11(1):45. PubMed PMID: 23061540. Pubmed Central PMCID: 3528622.
44. Feldacker C, Emch M, Ennett S. The who and where of HIV in rural Malawi: Exploring the effects of person and place on individual HIV status. Health & place. 2010 Sep;16(5):996-1006. PubMed PMID: 20598623. Pubmed Central PMCID: 3454472.
45. Haque U, Scott LM, Hashizume M, Fisher E, Haque R, Yamamoto T, et al. Modelling malaria treatment practices in Bangladesh using spatial statistics. Malaria journal. 2012 Mar 05;11:63. PubMed PMID: 22390636. Pubmed Central PMCID: 3350424.
46. Netherlands S. http://www.cbs.nl/nl-NL/menu/themas/dossiers/nederland-regionaal/publicaties/geografische-data/archief/2014/2013-wijk-en-buurtkaart-art.htm 2015 [cited 2015 Oct. 5].
47. Giammanco A, Chiarini A, Maple P, Andrews N, Pebody R, Gay N, et al. European Sero-Epidemiology Network: standardisation of the assay results for pertussis. Vaccine. 2003;22(1):112-20.

48. De Melker H, Versteegh F, Conyn-van Spaendonck M, Elvers L, Berbers G, van Der Zee A, et al. Specificity and sensitivity of high levels of immunoglobulin G antibodies against pertussis toxin in a single serum sample for diagnosis of infection with Bordetella pertussis. Journal of clinical microbiology. 2000;38(2):800-6.
49. Statistiek CBvd. Toelichting Wijk- en Buurtkaart 2010. Update 2 Den Haag/Heerlen,: Centraal Bureau voor de Statistiek; 2010 [cited 2015 Oct. 5]. Available from: http://www.cbs.nl/nl-NL/menu/themas/dossiers/nederland-regionaal/publicaties/geografische-data/archief/2011/2011-wijk-en-buurtkaart-2010-art.htm.
50. Lawson AB, Biggeri AB, Boehning D, Lesaffre E, Viel JF, Clark A, et al. Disease mapping models: an empirical evaluation. Disease Mapping Collaborative Group. Statistics in medicine. 2000 Sep 15-30;19(17-18):2217-41. PubMed PMID: 10960849.
51. Waller LA, Gotway CA. Applied spatial statistics for public health data: John Wiley & Sons; 2004.
52. Anselin L. Exploring Spatial Data with GeoDaTM : A Workbook Urbana, Illinois, USA: Spatial Analysis Laboratory, Department of Geography, University of Illinois at Urbana-Champaign; 2005 [cited 2014 2 March]. Available from: https://geodacenter.asu.edu/system/files/geodaworkbook.pdf.
53. Kulldorff M, Feuer EJ, Miller BA, Freedman LS. Breast cancer clusters in the northeast United States: a geographic analysis. American journal of epidemiology. 1997 Jul 15;146(2):161-70. PubMed PMID: 9230778. Epub 1997/07/15. eng.
54. Tanser F, Barnighausen T, Cooke GS, Newell ML. Localized spatial clustering of HIV infections in a widely disseminated rural South African epidemic. International journal of epidemiology. 2009 Aug;38(4):1008-16. PubMed PMID: 19261659. Pubmed Central PMCID: 2720393. Epub 2009/03/06. eng.
55. Wang T, Xue F, Chen Y, Ma Y, Liu Y. The spatial epidemiology of tuberculosis in Linyi City, China, 2005-2010. BMC public health. 2012;12:885. PubMed PMID: 23083352. Pubmed Central PMCID: 3487863. Epub 2012/10/23. eng.
56. Kulldorff M. SaTScan user guide for version 9.0. 2011.
57. Coleman M, Coleman M, Mabuza AM, Kok G, Coetzee M, Durrheim DN. Using the SaTScan method to detect local malaria clusters for guiding malaria control programmes. Malaria journal. 2009;8(68):10.1186.
58. Kulldorff M. SaTScanTM User Guide for version 9.2 2013 [cited 2013 8 September].
59. Chen J, Roth RE, Naito AT, Lengerich EJ, Maceachren AM. Geovisual analytics to enhance spatial scan statistic interpretation: an analysis of U.S. cervical cancer mortality. International journal of health geographics. 2008;7:57. PubMed PMID: 18992163. Pubmed Central PMCID: 2596098.
60. Kulldorff M, Heffernan R, Hartman J, Assunçao R, Mostashari F. A space-time permutation scan statistic for disease outbreak detection. PLoS medicine. 2005;2(3):216.
61. Kulldorff M. SaTScan-Software for the spatial, temporal, and space-time scan statistics. Boston: Harvard Medical School and Harvard Pilgrim Health Care. 2010.
62. ESRI. Interpreting Exploratory Regression results: ESRI; [cited 2016 November 21st]. Available from: http://desktop.arcgis.com/en/arcmap/10.3/tools/spatial-statistics-toolbox/interpreting-exploratory-regression-results.htm.
63. Fotheringham AS, Brunsdon C, Charlton M. Geographically weighted regression: the analysis of spatially varying relationships: John Wiley & Sons; 2003.

64. Hu M, Li Z, Wang J, Jia L, Liao Y, Lai S, et al. Determinants of the incidence of hand, foot and mouth disease in China using geographically weighted regression models. PloS one. 2012;7(6):e38978. PubMed PMID: 22723913. Pubmed Central PMCID: 3377651.
65. Nakaya T, Fotheringham AS, Brunsdon C, Charlton M. Geographically weighted Poisson regression for disease association mapping. Statistics in medicine. 2005 Sep 15;24(17):2695-717. PubMed PMID: 16118814.
66. Lovett AA, Bentham C, Flowerdew R. Analysing geographic variations in mortality using poisson regression: the example of ischaemic heart disease in England and Wales 1969–1973. Social science & medicine. 1986;23(10):935-43.
67. Lovett A, Flowerdew R. Analysis of count data using poisson regression*. The Professional Geographer. 1989;41(2):190-8.
68. Nakaya T. GWR4 user manual 2012. Available from: http://www. st-andrews. ac. uk/geoinformatics/wp-content/uploads/GWR4manual_201311. pdf.
69. Smith SM, Schroeder K, Fahey T. Over-the-counter (OTC) medications for acute cough in children and adults in ambulatory settings. Cochrane Database Syst Rev. 2008;1.
70. Campbell SM, Roland MO. Why do people consult the doctor? Family practice. 1996;13(1):75-83.
71. Centers for Disease C, Prevention. School-associated pertussis outbreak--Yavapai County, Arizona, September 2002-February 2003. MMWR Morbidity and mortality weekly report. 2004 Mar 19;53(10):216-9. PubMed PMID: 15029116.
72. Chuk LM, Lambert SB, May ML, Beard FH, Sloots TP, Selvey CE, et al. Pertussis in infants: how to protect the vulnerable? Communicable diseases intelligence quarterly report. 2008 Dec;32(4):449-56. PubMed PMID: 19374274.
73. van Lier A, van de Kassteele J, de Hoogh P, Drijfhout I, de Melker H. Vaccine uptake determinants in The Netherlands. The European Journal of Public Health. 2013:ckt042.
74. Van Lier E, Oomen P, Oostenbrug M, Zwakhals S, Drijfhout I, de Hoogh P, et al. [High vaccination coverage of the National Immunization Programme in the Netherlands]. Nederlands tijdschrift voor geneeskunde. 2009;153(20):950-7.
75. Barnett K, Mercer SW, Norbury M, Watt G, Wyke S, Guthrie B. Epidemiology of multimorbidity and implications for health care, research, and medical education: a cross-sectional study. The Lancet. 2012;380(9836):37-43.
76. Terry JB, Flatley CJ, van den Berg DJ, Morgan GG, Trent M, Turahui JA, et al. A field study of household attack rates and the effectiveness of macrolide antibiotics in reducing household transmission of pertussis. Communicable diseases intelligence quarterly report. 2015 Mar;39(1):E27-33. PubMed PMID: 26063095.
77. From C. Pertussis epidemic—Washington, 2012. Morbidity and Mortality Weekly Report (MMWR). 2012;61(28):517-22.
78. Service NH. Sharp rise in whooping cough cases: NHS; 2012 [cited 2016 January 25]. Available from: http://www.nhs.uk/news/2012/04april/Pages/whooping-cough-pertussis-warning.aspx.
79. Sizaire V, Garrido-Estepa M, Masa-Calles J, Martinez de Aragon M. Increase of pertussis incidence in 2010 to 2012 after 12 years of low circulation in Spain. Euro surveillance : bulletin Europeen sur les maladies transmissibles = European communicable disease bulletin. 2014;19(2).
80. Fisman DN, Tang P, Hauck T, Richardson S, Drews SJ, Low DE, et al. Pertussis resurgence in Toronto, Canada: a population-based study including test-incidence feedback modeling. BMC public health. 2011;11(1):1.

81. Kaczmarek MC, Valenti L, Kelly HA, Ware RS, Britt HC, Lambert SB. Sevenfold rise in likelihood of pertussis test requests in a stable set of Australian general practice encounters, 2000–2011. The Medical journal of Australia. 2013;198(11):624-8.
82. Calugar A, Ortega-Sánchez IR, Tiwari T, Oakes L, Jahre JA, Murphy TV. Nosocomial pertussis: costs of an outbreak and benefits of vaccinating health care workers. Clinical infectious diseases. 2006;42(7):981-8.
83. Ward A, Caro J, Bassinet L, Housset B, O'Brien JA, Guiso N. Health and economic consequences of an outbreak of pertussis among healthcare workers in a hospital in France. Infection Control. 2005;26(03):288-92.
84. Heffernan R, Mostashari F, Das D, Karpati A, Kulldorff M, Weiss D. Syndromic surveillance in public health practice, New York City. Emerging infectious diseases. 2004;10(5):858-64.
85. Ziemann A, Riesgo LG-C, Boris K, Schrell S, Rosenkötter N, Fischer M, et al. Added value of routine emergency medical data for detecting clusters of acute gastrointestinal illness in Europe. Resuscitation. 2012;83:e30.
86. Robertson C, Nelson TA. Review of software for space-time disease surveillance. International journal of health geographics. 2010;9(1):1.
87. Cutler DM, Lleras-Muney A. Education and health: evaluating theories and evidence. National Bureau of Economic Research, 2006.
88. Huang B, Wu B, Barry M. Geographically and temporally weighted regression for modeling spatio-temporal variation in house prices. International Journal of Geographical Information Science. 2010;24(3):383-401.

# CHAPTER 6

# General discussion

## Main Findings

This thesis assessed the value of Geographic Information Systems (GIS) and spatial epidemiological methods for demand-based planning and allocation of healthcare and targeted prevention strategies. The chosen approach in this thesis could successfully locate areas of high-risk for interventions and identify location-specific populations at risk for the provided case studies. GIS and spatial epidemiological methods have thus proven to be useful tools to inform evidence-based strategies for a demand-based planning and allocation of healthcare and targeted prevention strategies. The main finding of this thesis is that a one-size-fits-all approach is not very effective for both, demand-based planning and allocation of healthcare, but also for targeted prevention strategies. Cost-effective public health policies need to acknowledge that geographic aspects are important determinants of health and should therefore aim to shape future policies more towards local needs.

## Aims of this thesis

This thesis aimed to analyse how GIS and spatial epidemiological methods are useful tools to inform evidence-based strategies in public health. Three areas of application were examined where GIS and spatial epidemiological methods can provide an added value:

### 1. Demand-based planning and allocation of healthcare

Chapter 4 examined how spatial epidemiological methods can inform strategies to plan healthcare more effectively by analysing the spatial distribution as well as clustering of type 2 Diabetes Mellitus and associated population-based risk factors based on data of northeastern Germany`s largest statutory health insurance provider.

### 2. Evidence-based prevention strategies

Chapter 2, 3, 4 and 5 assessed how spatial regression modelling can identify the main population at risk for various diseases. The main population at risk for fever in rural areas of India was identified in chapter 2. Chapter 3, 4 and 5 expanded the spatial regression approach by examining how the association between Hepatitis C (chapter 3), type 2 Diabetes Mellitus (chapter 4), pertussis (chapter 5) and socio-demographic population characteristics varies within the respective regions.

3. **Detection of outbreaks for public health surveillance**

Chapter 5 analysed how space-time cluster detection can identify possible pertussis outbreaks based on laboratory data for pertussis testing.

**Research questions of this thesis**

1. **Is there an added value of the spatial scan statistic to identify areas for prevention strategies?**

This thesis assessed the use of the spatial scan statistic to identify and prioritize high-risk areas for interventions. The spatial scan statistic was favoured in this thesis over other local cluster tests such as the local indicator of spatial association (LISA) or the Besag-Newell-test (1) for following reasons: The spatial scan statistic facilitates the analysis of point as well as polygon data (2, 3), incorporates Bernoulli and Poisson distributions and allows a modification of the scanning window (3, 4). Additionally, the spatial scan statistic allows the identification of possible disease outbreaks in space and space-time. It is therefore currently the most flexible local cluster test and for those reasons the most widely used (1).

The application of the spatial scan statistic clearly identified areas with higher than expected disease risk in all four case studies. It is important to note however, that its usefulness should not be judged only by the ability to detect areas of significantly elevated risk. Especially for areas with few observations as in the case study on fever in India (n = 38) or the case study on Hepatitis C in the Netherlands (n = 126), spatial empirical Bayesian smoothing of disease rates would be sufficient to detect areas with elevated disease risk. When the goal is to prioritize areas for interventions however, the spatial scan statistic is a powerful method as it calculates for each cluster a likelihood rank, which can be used to prioritize specific areas (4). This was demonstrated in the case study on Hepatitis C. For large-scale address-based geocoded point data as it was used in the case study on type 2 Diabetes Mellitus, the identification of high-risk areas using the spatial scan statistic is inevitable as it is not possible to estimate the number of cases behind the surface of the kernel density estimation alone. Even when spatial empirical Bayesian smoothing is applied to polygon data consisting of a large number of observations, the application of a local cluster test is superior to a purely visual inspection alone as smaller administrative units in urban areas might be overlooked while larger administrative units with fewer inhabitants in rural areas would dominate the map. This was especially evident for the analysis of the sex- and age-adjusted rates

of type 2 Diabetes Mellitus on postal code level (n = 598). A major limitation of the spatial scan statistic is the use of a circular scanning window (4). This can lead to areas of low risk being included inside a cluster and therefore increases the risk of detecting false positives. The default setting of including up to 50% of the population at risk is not advisable in most cases (5). For different diseases in different geographic contexts, several settings need to be evaluated to find a meaningful search radius, capable of minimizing the risk of detecting false positives (2, 5, 6). To account for this limitation, it is advisable to compare the results of a disease mapping approach to the results of the spatial scan statistic (5). This comparison has been conducted in all four case studies and helped to determine a useful maximum search radius.

2. **How can GIS and spatial regression modelling facilitate demand-based allocation of healthcare?**

Chapter 4 examined how spatial epidemiological methods can inform strategies to plan healthcare more effectively by displaying how type 2 Diabetes Mellitus varies at the very local level and where significant hotspots are located. These hotspots may serve as indicator for areas, where specialized medical care should be allocated. The additional knowledge gained from the spatial regression approach that type 2 Diabetes Mellitus is associated with a lower socio-economic status contributes to the on-going discussion in Germany to include measures of area deprivation into planning of healthcare (7).

3. **Is geographically weighted regression a suitable method to identify location-specific risk groups for targeted prevention strategies?**

While there is an abundance of global spatial regression models in public health (1), only few spatial regression modelling approaches are available to analyse spatially varying associations. The main approaches used in spatial epidemiological research remain geographically weighted regression (GWR) (1) and Bayesian spatially varying coefficient models (SVC) (8, 9). GWR was favoured in this research as the application of SVC still remains computationally challenging and difficult to implement (8), while the use of GWR is facilitated through the free available GWR4 software (10) and the R package "GWmodel" (11). For these reasons, GWR is more popular than the SVC model. In three case studies, GWR was successfully used to identify location-specific determinants of Hepatitis C in the Netherlands, type 2 Diabetes Mellitus in Germany and pertussis testing and infections in the Netherlands. Based on the experience of applying

GWR on different datasets in different geographic contexts, GWR has proven useful to identify location-specific risk groups for infectious and chronic diseases. Nonetheless, several methodological issues need to be discussed in the subsequent paragraphs.

### 3.1. Which statistical properties of geographically weighted regression modelling have to be considered to obtain useful results?

First, the dataset needs to include a sufficient amount of observations. Based on a simulation study conducted in 2011, Paéz et al. point out that the results of GWR are unreliable for datasets with fewer than 160 observations. A low number of observations poses the risk of extreme coefficients although the "true" underlying spatial process may not vary as much as indicated by GWR (12). It is important to consider that the statistical methodology of GWR has continuously improved since the publication of this simulation study and now includes diagnostics to assess whether the underlying process is stationary or not (11). Additionally, GWR provides goodness-of-fit-statistics to compare its performance to a global baseline model (10). Comparing the goodness-of-fit statistics of a global model to those of a local model is an important step, which should always be conducted (11, 13). In three case studies, the local approach provided a better fit than the global approach. The case study on acute undifferentiated fever was the only case study where a local approach did not provide a better fit than a global approach. This was due to the low number of observations of this dataset (n = 38). Although the case study on Hepatitis C also included less than 160 administrative units (n = 126), the local approach provided a better fit than the global approach. Several other studies similarly applied GWR on datasets with fewer than 160 administrative units (14, 15). Based on the experience from the case studies in this thesis, it may be suggested that GWR can be successfully applied also on smaller datasets, assured that the goodness-of-fit statistics point out that a local model provides a better fit than a global model. This can be assessed by comparing Akaike`s corrected information criterion (AICc) and the explained variance (adjusted $R^2$) of the competing global and local models (11, 13).

Second, the choice of kernel distribution and size of bandwidth play an important role in how the associations between disease risk and possible risk factors vary over space in the model. Both, the GWR4 software and the R package "GWmodel" provide a vast amount of possible settings for the kernel distribution and size of the bandwidth (11, 16). A major quality criterion of a spatial regression model is the absence of residual spatial autocorrelation (13). This is an important criterion, which should always be

compared – not only between a global model and a local model – but also between possible local candidate models. In the case studies on type 2 Diabetes Mellitus and pertussis, several combinations of kernel distribution and bandwidth size had to be evaluated to find a model capable of eliminating residual spatial autocorrelation. Additionally, although one combination may be able to eliminate residual spatial autocorrelation, it may not necessarily be the best model according to the goodness-of-fit statistics. This became especially apparent in the case study of type 2 Diabetes Mellitus. It is therefore highly recommendable to examine all possible combinations of kernel distributions and bandwidth sizes to find the best fitting model. During the time of this thesis, the "GWmodel" package in R was introduced. Not only does it provide a larger amount of kernel distributions and further statistical diagnosis tests than the GWR4 software (11), the R software allows writing a script to automatically compare all different settings of a GWR model and to test for residual spatial autocorrelation within one software, speeding up computation time and minimizing the amount of tasks, which have to be performed manually (11).

Third, a major critique about GWR is the unreliability of the coefficient estimates, especially in situations where the coefficients change sign across the study area. This makes the interpretation of the coefficient estimates counter-intuitive in certain areas (12, 17, 18). This problem is not only apparent in simulation studies, but is also evident in several spatial epidemiological studies based on real-world data (19, 20). The change of sign of coefficient estimates also became apparent in all three case studies using GWR. This issue was alleviated when the coefficient estimates were displayed together with significance thresholds, as areas with counter-intuitive coefficient estimates were in most cases insignificant. However, not all studies report statistical significance of coefficient estimates of a GWR model, making the interpretation of the results challenging (19, 20). GWR coefficient estimates are logically only meaningful when they are reported with corresponding significance thresholds to minimize the risk of possible false conclusions.

### 3.2. What are current limitations of geographically weighted regression modelling?

Despite being a very flexible spatial regression modelling approach with an extensive amount of adjustment options, several studies see the use of GWR only as exploratory

but inappropriate for statistical inference (21, 22). GWR has several methodological limitations, which need to be addressed in this context:

First, GWR calculates the spatially varying coefficients within a circular kernel (23). As a result, the coefficient estimates depend highly upon the bandwidth of the kernel. If the bandwidth is too small, the coefficient estimates are exaggerated and unstable. If the bandwidth is too large, there will be only little variation in the coefficient estimates and local processes remain hidden. The subjective choice of the bandwidth and its impact on the resulting coefficient estimates is currently the main core argument why GWR can not be considered as an inferential method (21). This problem became apparent in all case studies as the choice of bandwidth largely determined the amount of variability of the coefficient estimates. To account for this issue, the optimal bandwidth and kernel function was based on objective goodness-of-fit-statistics of the respective kernel and bandwidth rather than a subjective choice. In addition, a circular form of the kernel does not necessarily represent the true form of the underlying spatial process. Most spatial processes are irregularly formed and the use of a circular kernel poses the risk of over-generalization of coefficient estimates (8). This limitation is inherent to most local spatial statistics including kernel density estimation (24), kriging (25) and the spatial scan statistics (4) and is thus not a limitation of GWR alone.

Second, unlike global spatial regression modelling approaches, GWR does not calculate one single regression equation over the entire dataset, but rather divides the dataset into as much regression equations as there are observations. This approach in turn challenges the calculation of significance of the local regression coefficients, as multiple testing is unavailable due to problems arising from these multiple local regression equations. As a result, significance values of the regression coefficients can be considered only as an approximation, but not as exact significance values (21).

Third, the application of a GWR model requires substantial knowledge about the study area to question the plausibility of the results. This became apparent during the analysis of type 2 Diabetes Mellitus. During the initial analysis, the variable “proportion of built surfaces” was included as possible explanatory variable. While an association between built environment and type 2 Diabetes Mellitus has been noted in the literature (26), this association was only significant in the most sparsely populated regions of the study area, where green spaces are easily accessible. Based on knowledge about this area, the variable was then removed from the analysis. If only the results of a global model would have been reported, this association would be seen as in line with previous

research (26). The local approach in contradiction showed that this association would be misleading in the study area. This limitation can therefore be seen as a huge advantage over global regression models as local models facilitate questioning the plausibility of the results.

Fourth, excess zeros within an outcome variable are difficult to account for in GWR and lead to exaggerated coefficient estimates. This became apparent in the case study on pertussis as the coefficient estimates were highly exaggerated when interpreted as change in the number of cases with a one unit change of the explanatory variable. If a large amount of observations has zero cases, zero-inflated Poisson models would be more suitable than the standard Poisson framework employed within GWR (27). However, zero-inflated Poisson models are currently unavailable in the GWR framework (11). To account for this problem in the case study on pertussis, the explanatory variables were standardized. The resulting coefficient estimates thus allowed only an exploratory interpretation where the effect of an explanatory variable was stronger in comparison to other explanatory variables.

The question whether GWR is only an exploratory method or can be used as inferential method still remains difficult to answer after the case studies in this thesis because the only competing model with which it can be compared is modelled in a Bayesian framework. Unfortunately, the Bayesian SVC model remains challenging to implement in existing software (8, 9). It is important to consider that this question can also not be answered conclusively in existing literature (21, 23). Given the theoretical assumption that associations between disease occurrence and aggregated population characteristics vary over space due to cultural, social and environmental processes on an individual and ecological level (28), allowing associations to vary over space in a regression model seems more plausible than assuming that associations are constant across space. It is however also clear, that current methodological limitations prevent GWR from providing exact regression parameters. Despite this limitation, the knowledge who is where at risk is still more meaningful for public health policies than knowing only who is at risk, even when the estimates should be considered as approximations only. The possibility to target specific risk groups in specific locations based on the results of GWR is therefore of more practical relevance than knowing only which population group is at risk.

4. **Can geographically weighted regression and space-time cluster detection facilitate the detection of possible pertussis outbreaks in the Netherlands?**

Chapter 5 demonstrated that space-time cluster detection could be used to identify possible outbreaks in space and time for pertussis. However, the spatial regression approach clearly confirmed that the current incidence of pertussis is mainly the result of testing behaviour of the general population and general practitioners. As a result, this case study provided evidence that the incidence of pertussis based on the total population is not a useful indicator to detect outbreaks of pertussis, given the strong increase of pertussis cases with testing. To detect possible outbreaks of pertussis, which are not the result of testing behaviour only, test-positivity – defined as ratio of positive tested persons to all tested persons – is a more suitable indicator as input for the space-time cluster analysis. The space-time cluster analysis in the SaTScan software could successfully locate clusters in space and time and has thus proven useful to detect possible outbreaks of pertussis.

**Recommendations to implement GIS in public health departments**

This thesis has shown that GIS and spatial epidemiological methods are useful tools to inform evidence-based strategies at the local level. The implementation of GIS in public health departments should consider several recommendations arising from this dissertation:

First, spatial analyses are still not considered a standard methodology in public health studies. Specialized training is necessary to acquire the skills to handle spatial data and to be able to analyse them in an epidemiological context. Such training should include experienced professionals with similar research backgrounds.

Second, with the increasing availability of commercial or open-source GIS and statistical software, public health departments may find it difficult to choose appropriate software. Based on the case studies of this dissertation, following GIS software can be recommended: If the data is available aggregated only and requires only little data preparation, free open-source GIS software such as QGIS (29) is sufficient. If the data is available based on exact addresses and requires more sophisticated data preparation and analyses on transportation networks, commercial GIS software such as ESRI ArcGIS (30) is still considered as gold standard. Spatial statistical software is more challenging to choose from as the various programmes are often designed to fulfil only one specific task. This dissertation relied on SaTScan (4) for the local cluster analysis, GeoDa (31) for

spatial empirical Bayesian smoothing and global spatial regression modelling, GWR4 (16) for geographically weighted regression modelling and CrimeStat IV (24) for the kernel density estimation. It is important to recognize that the majority of spatial statistical analyses in this dissertation can also be carried out using the R programming language (32). Although the learning curve of R is steep, the time spent on learning how to use R pays off in settings where analyses have to be carried out repeatedly (33). Incorporating GIS at local public health departments therefore becomes feasible, even in settings with only limited financial resources.

Third, GIS has potentially more benefits at the local level rather than the national level. Local knowledge of limitations associated with data collection and characteristics of the study area, which may be unknown at the national level, have to be taken into consideration to provide realistic and practically useful results.

**Recommendations for public health professionals**

The general application of the methods used in this thesis is covered in detail in the accompanying manuals of the respective programmes. Spatial empirical Bayesian smoothing and global regression modelling are covered in the GeoDa workbook (31), the spatial scan statistic is covered in the SaTScan user manual (4) and geographically weighted regression modelling is described in the GWR4 user manual (16). Bivand`s book on spatial modelling in R provides tutorials of these methods in the R programming language (33). While the application of spatial empirical Bayesian smoothing and global spatial regression modelling is relatively straightforward, the use of the spatial scan statistic and geographically weighted regression modelling is very sensitive to the choice of settings used to conduct the analysis. Several practical recommendations based on the four case studies can be provided:

1. Spatial scan statistic: As the usefulness of the detected clusters depends upon the size of the scanning window, it is necessary to compare different possible scanning window sizes. In the four case studies, smaller window sizes tended to deliver results, which are of more practical use than the standard settings. A cluster can be defined as practically useful if it does not include areas with a relative risk below average. Several iterations with different scanning window sizes may be necessary to find only clusters containing areas with a relative risk above average. To examine, whether the detected clusters meet this criterion, it is

recommended to visualize the clusters in addition to the cartographic analysis of disease risk.

2. Limitations of data collection: Several diseases such as pertussis are rather driven by testing behaviour than real outbreaks. To examine risk factors of similar diseases, the proportion of tested persons should be included in a spatial regression model to account for the impact of testing. This approach provides more realistic estimates of the association to possible risk factor as the important role of testing is accounted for in this approach.
3. Geographically weighted regression modelling: A global spatial regression model always precedes a GWR to assess the requirements of an appropriate spatial regression model: All variables are significantly associated with the outcome, the explanatory variables are free from redundancy and the residuals are free from spatial autocorrelation (13). As a major critique of GWR is the subjectivity of bandwidth size and kernel distribution (21), all possible combinations of bandwidth size, kernel distribution and optimization method should be compared by Akaike`s corrected information criterion, deviance explained by the model and clustering of the residuals to objectively find the best fitting model.

**Limitations of this dissertation**

First, socio-demographic and socio-economic risk factors for diseases were analysed on an aggregated population level. Possible risk factors derived from an ecological analysis do not necessarily constitute risk factors on an individual level. This problem is termed “ecological bias” and is inherent to all studies based on aggregated population characteristics (34, 35). The ecological bias is not a limitation for planning and allocation of healthcare as healthcare is typically planned based on larger spatial scales where broader processes are of interest (14, 34). However, targeting individuals with specific socio-demographic characteristics based on the results of an ecological analysis remains challenging if these associations were derived from large-scale administrative units with considerable in-area variation (35). Small administrative units with homogenous population characteristics are preferable to identify demographic and socio-economic population characteristics. The effect of the ecological bias is usually small in small-scale spatial epidemiological studies with a few hundred or thousand inhabitants, but increases with size of the administrative units (35). It can therefore be assumed that this bias is relatively small for analyses based on postal codes in the

Netherlands and analyses based on the association of municipalities in Germany. In the case study based on counties in India however, this effect could be relatively large. The strength of the ecological bias could however not be assessed in this thesis and still remains unknown.

Second, the identified associations depend on the available population characteristics. It was not possible to evaluate whether the diseases of the four case studies are related to lower levels of education, as this information was not available for the four case studies.

Third, the usefulness of an ecological study design depends on the quality and spatial resolution of population data. This became especially apparent in the case study of fever in India. The relatively large counties, on which the analysis was based, constitute the smallest spatial scale for which population data were available during the time of the analysis. Additionally, population data in India are updated only every ten years (36). In how far population data from the census of India for 2001 are still meaningful for spatial epidemiological data of 2009 could not be evaluated.

Fourth, causality based on an ecological study design is difficult to assess. Although there is an association between chronic diseases and lower socio-economic status at the ecological level (19), an ecological study design is unable to assess whether chronic diseases are the result of a lower socio-economic status (e.g. through lower levels of education, resulting in an unhealthy lifestyle) or whether chronic diseases led to a lower socio-economic status through work impairment or loss of workplace (37).

Despite these limitations, spatial ecological studies allow the analysis of large, anonymised epidemiological data and readily available population characteristics. Spatial ecological studies therefore allow inference about possible demographic and socio-economic risk factors, which are unavailable on an individual level. This critical information would often remain unknown without time-consuming study designs relying on voluntary participation of individuals. As spatial ecological studies facilitate the prioritization of areas for interventions, can be conducted in a very time- and cost-effective manner and can be adapted to most available epidemiological datasets, their importance for public health policies is likely to increase in the near future.

## Conclusions

GIS and spatial regression modelling empower decision-makers with important background knowledge about high-risk areas and the main population, which is most at

risk in specific locations. This knowledge is important to allocate financial resources for demand-based, cost-effective planning and allocation of healthcare and targeted prevention strategies.

This thesis clearly demonstrated that a one-size-fits-all approach of public health policies is often inappropriate. Future public health policies need to acknowledge the importance of geographic aspects of health and disease and should aim to shape policies more towards local needs.

**References:**


1. Auchincloss AH, Gebreab SY, Mair C, Roux AVD. A review of spatial methods in epidemiology, 2000–2010. Annual review of public health. 2012;33:107.
2. Tanser F, Bärnighausen T, Cooke GS, Newell M-L. Localized spatial clustering of HIV infections in a widely disseminated rural South African epidemic. International Journal of Epidemiology. 2009:dyp148.
3. Warden CR. Comparison of Poisson and Bernoulli spatial cluster analyses of pediatric injuries in a fire district. International journal of health geographics. 2008;7(1):51.
4. Kulldorff M. SaTScan user guide for version 9.0. 2010.
5. Chen J, Roth RE, Naito AT, Lengerich EJ, MacEachren AM. Geovisual analytics to enhance spatial scan statistic interpretation: an analysis of US cervical cancer mortality. International journal of health geographics. 2008;7(1):57.
6. Martin SW, Michel P, Middleton D, Holt J, Wilson J. Investigation of clusters of giardiasis using GIS and a spatial scan statistic. International journal of health geographics. 2004;3(1):11.
7. Maier W, Fairburn J, Mielck A. Regionale Deprivation und Mortalität in Bayern. Entwicklung eines, Index Multipler Deprivation 'auf Gemeindeebene. Das Gesundheitswesen. 2012;74(07):416-25.
8. Wheeler DC, Waller LA. Comparing spatially varying coefficient models: a case study examining violent crime rates and their relationships to alcohol outlets and illegal drug arrests. Journal of Geographical Systems. 2009;11(1):1-22.
9. Waller LA, Zhu L, Gotway CA, Gorman DM, Gruenewald PJ. Quantifying geographic variations in associations between alcohol distribution and violence: a comparison of geographically weighted regression and spatially varying coefficient models. Stochastic Environmental Research and Risk Assessment. 2007;21(5):573-88.
10. Nakaya T. GWR4 user manual. WWW Document Available online: http://www st-andrews ac uk/geoinformatics/wp-content/uploads/GWR4manual_201311 pdf (accessed on 4 November 2013). 2014.
11. Lu B, Harris P, Charlton M, Brunsdon C. The GWmodel R package: further topics for exploring spatial heterogeneity using geographically weighted models. Geo-spatial Information Science. 2014;17(2):85-101.
12. Páez A, Farber S, Wheeler D. A simulation-based study of geographically weighted regression as a method for investigating spatially varying relationships. Environment and Planning A. 2011;43(12):2992-3010.
13. ESRI. Regression analysis basics: ESRI; 2016 [cited 2017 Jan. 30th]. Available from: http://desktop.arcgis.com/en/arcmap/10.3/tools/spatial-statistics-toolbox/regression-analysis-basics.htm.
14. Dijkstra A, Janssen F, De Bakker M, Bos J, Lub R, Van Wissen LJ, et al. Using spatial analysis to predict health care use at the local level: a case study of type 2 diabetes medication use and its association with demograpHic change and socioeconomic status. PloS one. 2013;8(8):e72730.
15. Robertson C, Pant DK, Joshi DD, Sharma M, Dahal M, Stephen C. Comparative spatial dynamics of Japanese encephalitis and acute encephalitis syndrome in Nepal. PloS one. 2013;8(7):e66168.
16. Nakaya T. GWR4 user manual. WWW Document Available online: http://www st-andrews ac uk/geoinformatics/wp-content/uploads/GWR4manual_201311 pdf (accessed on 4 November 2013). 2009.

17. Farber S, Páez A. A systematic investigation of cross-validation in GWR model estimation: empirical analysis and Monte Carlo simulations. Journal of Geographical Systems. 2007;9(4):371-96.
18. Farber S, Yeates M. A comparison of localized regression models in a hedonic house price context. Canadian Journal of Regional Science. 2006;29(3):405-20.
19. Ford MM, Highfield LD. Exploring the Spatial Association between Social Deprivation and Cardiovascular Disease Mortality at the Neighborhood Level. PloS one. 2016;11(1):e0146085.
20. Weisent J, Rohrbach B, Dunn JR. Socioeconomic determinants of geographic disparities in campylobacteriosis risk: a comparison of global and local modeling approaches. International journal of health geographics. 2012;11(1):45.
21. Wheeler DC. Geographically weighted regression. Handbook of Regional Science: Springer; 2014. p. 1435-59.
22. Wheeler D, Tiefelsdorf M. Multicollinearity and correlation among local regression coefficients in geographically weighted regression. Journal of Geographical Systems. 2005;7(2):161-87.
23. Fotheringham AS, Brunsdon C, Charlton M. Geographically weighted regression: John Wiley & Sons, Limited; 2003.
24. Levine N. CrimeStat III: a spatial statistics program for the analysis of crime incident locations (version 3.0). Houston (TX): Ned Levine & Associates/Washington, DC: National Institute of Justice. 2004.
25. Schabenberger O, Gotway CA. Statistical methods for spatial data analysis: CRC press; 2004.
26. Salois MJ. Obesity and diabetes, the built environment, and the 'local'food economy in the United States, 2007. Economics & Human Biology. 2012;10(1):35-42.
27. Lambert D. Zero-inflated Poisson regression, with an application to defects in manufacturing. Technometrics. 1992;34(1):1-14.
28. Cromley EK, McLafferty SL. GIS and public health: Guilford Press; 2011.
29. QGIS D. QGIS geographic information System. Open source geospatial Foundation project. 2015.
30. Desktop EA. Release 10. Redlands, CA: Environmental Systems Research Institute. 2011.
31. Anselin L. Exploring spatial data with GeoDaTM: a workbook. Urbana. 2004;51:61801.
32. Team RDC. R: A language and environment for statistical computing. : R Foundation for Statistical Computing, Vienna, Austria.; 2008 [cited 2017 Sep. 25th]. Available from: http://www.r-project.org.
33. Bivand RP, E., Gomez-Rubio, V. Applied Spatial Data Analysis with R. 2 ed. New York: Springer; 2013.
34. Anselin L. Under the hood issues in the specification and interpretation of spatial regression models. Agricultural economics. 2002;27(3):247-67.
35. Wakefield J. Ecologic studies revisited. Annu Rev Public Health. 2008;29:75-90.
36. General R. census Commissioner. Census of India. 2001;2003.
37. Kessler RC, Greenberg PE, Mickelson KD, Meneades LM, Wang PS. The effects of chronic medical conditions on work loss and work cutback. Journal of Occupational and Environmental Medicine. 2001;43(3):218-25.

# CHAPTER 7

Summary

Nederlandse samenvatting

Acknowledgement

Curriculum Vitae

Publications

Valorisation of this thesis

# Summary

The case studies in this thesis describe the use of Geographic Information Systems (GIS) in public health. The main focus lies in the assessment of GIS and spatial epidemiological methods for planning and allocation of healthcare and targeted prevention strategies.

**Chapter 2: Case study on Acute Undifferentiated Fever in India**

The **S**ystem for **E**arly-warning based on **E**mergency **D**ata (SEED) is a pilot project to evaluate the use of emergency call data with the main complaint acute undifferentiated fever (AUF) for syndromic surveillance in India. Although the main focus of syndromic surveillance lies in the detection of possible disease outbreaks, additional information about the main population at risk is necessary for targeted interventions and future preparedness strategies. By analysing the emergency calls of a small, remote area geographically and by merging the emergency call data with socio-economic population characteristics, we found that the incidence of fever was higher in rural areas and showed strong regional variation. The results of the spatial regression analysis clearly identified scheduled tribes and household industries as the main population at risk and proximity to forests as an environmental risk factor. These results are especially important in the Indian context, where laboratory confirmed disease data are scarce and emergency call data could provide a suitable alternative as proxy for infectious diseases.

**Chapter 3: Case study on Hepatitis C in the Netherlands**

Hepatitis C Virus (HCV) infections are a major cause for liver diseases. A large proportion of these infections remain hidden to care due to its mostly asymptomatic nature. Population-based screening and screening targeted on behavioural risk groups had not proven to be effective in revealing these hidden infections. By geocoding the HCV tests collected between 2002 and 2008 in the southern part of the Netherlands and merging them with socio-demographic population data at the four-digits postal code level, local hotspots of high HCV prevalence could be detected and the main population at risk in specific locations could be identified. The risk group for HCV consisted of persons living in one-person households, persons with low income, non-western immigrants and persons living in deprived areas. Combining the information where local hotspots are with the background knowledge which population group is most at risk in these hotspots provides a useful starting point for future screening interventions.

**Chapter 4: Case study on Type 2 Diabetes Mellitus in Germany**

The provision of general practitioners (GPs) in Germany still relies mainly on the ratio of inhabitants to GPs at relatively large scales and barely accounts for an increased prevalence of chronic diseases among the elderly and socially underprivileged populations. As health insurance is mandatory in Germany, geocoded health insurance claims can be used to analyse the spatial distribution of chronic diseases as indicator for the demand for primary care. Data from northeastern Germany`s largest statutory health insurance provider was used to analyse the spatial distribution of type 2 Diabetes Mellitus (T2DM). The results clearly show that T2DM varies at the very local level and is strongly clustered, especially in rural areas. The results of the spatial regression analysis emphasize that a lower socio-economic status is – at least in some areas – an important predictor of T2DM. The results of our analysis provide very detailed information, where hotspots of T2DM in northeastern Germany are located. The background knowledge, that T2DM is driven by lower socio-economic status further facilitates the recent political discussion to include measures of lower socio-economic status into the current guidelines of planning and allocation of primary care in Germany.

**Chapter 5: Case study on Pertussis in the Netherlands**

Despite high vaccination coverage, pertussis incidence in the Netherlands is amongst the highest in Europe with a shifting tendency towards adults and elderly. Early detection of outbreaks and preventive actions are necessary to prevent severe complications in infants. We therefore analysed geocoded laboratory registry data collected between 2007 and 2013 in the province South Limburg. We could successfully locate space-time clusters for pertussis testing, incidence and test-positivity. The space-time clusters for pertussis incidence largely overlapped with pertussis testing. The spatial regression approach further confirmed that the current pertussis incidence is largely the result of testing. More testing would therefore not necessarily improve pertussis control. Although the detection of pertussis outbreaks would be feasible using space-time cluster detection, such an approach should rather use test-positivity as indicator to account for the strong association between testing and the current pertussis incidence.

**Chapter 6: General Discussion**

In chapter 6, the main findings of this thesis are discussed. The main results of this thesis are that GIS and spatial epidemiological methods are suitable to locate high-risk areas and to identify the main populations at risk in specific locations for the analysed diseases. The results of the four case studies emphasize that a one-size-fits-all approach is not very effective for both, planning and allocation of healthcare and targeted prevention strategies. Future public health policies need to acknowledge that geographic aspects are important determinants of health and should aim future interventions more towards local needs.

# Nederlandse samenvatting

De case study's in deze thesis beschrijven het gebruik van Geografische Informatie Systemen (GIS) in de openbare gezondheidszorg. De hoofdfocus ligt op de evaluatie van inzet GIS en ruimtelijke epidemiologische methoden ten behoeve van planning en toewijzing van gezondheidszorg en doelgerichte preventiestrategieën.

**Hoofdstuk 2: case study van Acute Ongedifferentieerde Koorts in India**

Het **S**ystem for **E**arly-warning based on **E**mergency **D**ata (SEED) is een pilotproject om de bruikbaarheid van data van noodoproepen van de belangrijkste syndroom 'acute undifferentiated fever (AUF)', ten behoeve van syndroomsurveillance in India te evalueren. Alhoewel het belangrijkste doel van syndroomsurveillance ligt in het opsporen van mogelijke ziekte-uitbraken is additionele informatie over de belangrijkste risicopopulatie noodzakelijk voor gerichte interventies en toekomstige preventiestrategieën. Door de noodoproepen van een klein, afgelegen gebied geografisch te analyseren en door de data van de noodoproepen samen te voegen met sociaal-economische bevolkingskarakteristieken ontdekten we dat de incidentie van koorts in plattelandsgebieden hoger was en het toonde grote regionale verschillen.
De resultaten van de ruimtelijke regressie-analyse wezen duidelijk bepaalde stammen en bedrijfjes aan huis als grootste risicopopulaties aan en de nabijheid tot bosgebieden als omgevingsrisicofactor.

Deze resultaten zijn vooral belangrijk in de Indiase context waar weinig laboratoriumbevestigde ziektedata aanwezig zijn en waar data over noodoproepen een geschikt alternatief zouden kunnen vormen als indicatie voor infectieziekten.

**Hoofdstuk 3: Case study van hepatitis C in Nederland**

Hepatitis C-virusinfecties (HCV) vormen een belangrijke oorzaak voor leverziekten. Een groot deel van deze infecties blijft verborgen voor de zorg omdat het meestal een asymptomtisch verloop heeft. Bevolkingsonderzoek en screening gericht op gedragsgerelateerde risicogroepen bleken niet effectief om deze verborgen infecties op te sporen. Door geocodering van de HCV-testen verzameld tussen 2002 en 2008 in het zuidelijk deel van Nederland en door deze samen te voegen met socio-demografische bevolkingsdata op vier-cijferig postcodeniveau, werden lokale hotspots met hoge HCV-prevalentie gevonden en kon de belangrijkste risicogroep op specifieke locaties worden geïdentificeerd. De risicogroep voor HCV bestond uit eenpersoonshuishoudens, personen met een laag inkomen, niet-westerse immigranten en personen uit

achterstandswijken. De informatie waar lokale hotspots zich bevinden, gecombineerd met de achtergrondinformatie welke populatiegroep het grootste risico loopt binnen deze hotspots, levert een goed uitgangspunt voor toekomstige screeninginterventies.

**Hoofdstuk 4: Case study van diabetes mellitus type 2 in Duitsland**

Het systeem voor de verdeling van huisartsen per inwoner in Duitsland hangt hoofdzakelijk af van een standaard aantal inwoners en houdt amper rekening met de verhoogde prevalentie van chronische ziekten onder ouderen en achtergestelde bevolkingsgroepen. Aangezien in Duitsland ziektekostenverzekering verplicht is, kan geocodering van ziektekostendeclaraties worden gebruikt om de ruimtelijke verdeling van chronische ziekten te analyseren als indicator voor zorgvraag. Data van de grootste (wettelijke) ziektekostenverzekeraar in Noordoost Duitsland werden gebruikt om de ruimtelijke verspreiding van diabetes mellitus type 2 (DM2) te analyseren. De resultaten tonen duidelijk aan dat DM2 op zeer lokaal niveau varieert en sterk geclusterd is, m.n. in landelijke gebieden. De resultaten van de ruimtelijke regressie analyse benadrukken dat een lagere sociaal economische status – tenminste in bepaalde gebieden – een belangrijke voorspeller voor DM2 is. De resultaten van onze analyse leveren zeer gedetailleerde informatie over waar hotspots van DM2 in Noordoost Duitsland zijn gelegen. De achtergrondkennis dat DM2 wordt aangedreven door een lage SES-status draagt bij aan de recente, politieke discussie om maatregelen bij een lagere SES-status te includeren in de huidige richtlijnen van plannen en toewijzen van primaire zorg in Duitsland.

**Hoofdstuk 5: Case study over pertussis in Nederland**

Ondanks de hoge vaccinatiegraad is de incidentie van pertussis in Nederland één van de hoogste in Europa met een neigende verschuiving naar volwassenen en ouderen. Vroege detectie van uitbraken en preventieve acties zijn nodig om ernstige complicaties bij zuigelingen te voorkomen. Hiervoor analyseerden wij geogecodeerde laboratoriumdata in de provincie Limburg, verzameld tussen 2007 en 2013. We konden met succes locatie-tijd-clusters voor het testen op pertussis, de incidentie en de testpositiviteit opsporen. De locatie-tijd-clusters voor pertussisincidentie vertoonden een grote overlap met het testen op pertussis. De ruimtelijke regressiebenadering bevestigde dat de huidige pertussisincidentie grotendeels het resultaat is van testen. Meer testen zou daarom niet noodzakelijkerwijs de bestrijding van pertussis verbeteren.

Alhoewel detectie van pertussisuitbraken door middel van plaats-tijd clusterdetectie mogelijk is, zou eigenlijk het gebruik van testpositiviteit als indicator dienen te worden gebruikt om de sterke associatie tussen testen en de huidige pertussisincidentie te verklaren.

**Hoofdstuk 6: Algemene discussie**

In hoofdstuk 6 worden de belangrijkste bevindingen van deze thesis besproken. De belangrijkste resultaten van deze thesis zijn dat GIS en ruimtelijke epidemiologische methodes geschikt zijn om hoog-risicogebieden te traceren en om de belangrijkste risicogroepen op specifieke locaties voor de geanalyseerde ziekten te identificeren. De resultaten van de 4 case study's benadrukken dat een algemene aanpak niet erg effectief is, zowel voor toewijzen van gezondheidszorg als voor gerichte preventiestrategieën. Toekomstig public health-beleid dient te onderkennen dat geografische aspecten belangrijke determinanten van gezondheid zijn en het zou toekomstige interventies meer moeten richten op lokale behoeften.

# Acknowledgements

After an exhaustive and exciting time conducting the research for this dissertation, now is finally the time to lay down the finishing touches and to thank everyone involved in this undertaking. It was an intensive and inspiring period of my life. The process of writing this dissertation was shaped by many persons, which I have met along the way:

First, I would like to thank my supervisors:

Prof. dr. Christian Hoebe for believing in me to do a PhD and always providing thoughtful and supporting guidance where it was needed. Dear Christian, the meetings we had were full of positive and inspiring discussions. I learned a lot about seeing the data we work with in their broader context. This helped me to understand the importance of the daily and practical context for epidemiological research.

Prof. dr. Thomas Krafft for offering me the opportunity to work in an interesting, international, and very inspiring environment. Dear Thomas, I have to express my gratitude to you for providing me the chance to see the many interesting areas of medical geography and expanding my academic horizon. Learning to be scientifically sound and critical was an important lesson, which is invaluable for my career. Also, I would like to thank you very much for providing me a place to stay at yours and Eva`s place during my visits in Maastricht.

Dr. Nicole Dukers-Muijrers for being a very inspiring co-author on many papers during this dissertation. Dear Nicole, your knowledge of all the many aspects of epidemiological research was truly inspiring and helped me to look at things from many different angles.

Prof. dr. Jürgen Schweikart for being a very supportive and interested supervisor. Dear Jürgen, having a person with such a deep history and knowledge of GIS as supervisor was a real enhancement of this thesis. I cannot express my gratitude for the great and many discussions and advices on a scientific, but also on a personal level.

Also, I would like to thank Marita Moskwyn and Andrea Keste from the AOK Nordost for their amazing support of my daily work and for the opportunity to use part of it in my dissertation.

Further, my sincere gratitude belongs to all co-authors, I have had the pleasure to work with: Dr. Alexandra Ziemann, Eva Pilot, Dr. Oliver Grübner, Dr. Biranchi Jena, Dr.

Ramana Rao, Jonas Pieper and Dr. Werner Maier. I hope we can work together again sometime in the future.

A large proportion of keeping up the spirit and work ethic during such a long undertaking stems from taking a break from complicated, scientific matter. I therefore owe a big thank you to all my friends that supported me indirectly through their love and happiness. The countless skateboarding sessions, workouts at the gym and fun evenings at the bar helped me not to lose my focus in the long run. Let`s do it again, my friends.

I struggle to find words to express my gratitude for my significant other and love of my life, as words will not do my gratefulness any justice. You have endured me through challenging times, made me laugh and smile, and walked along with me the path of life for already more than eight years. I love you for all of this and so much more.

Finally, I would like to thank my parents for giving me the gift of life and being not only parents but also friends. Being from a non-academic family and having your support to start a scientific career nonetheless was a true blessing. I thank you so much.

# Curriculum Vitae

Boris Kauhl was born in 1984 in Laupheim, Germany. He studied geography at the University of Cologne and received his diploma in Geography in 2012. During his studies, he found strong interest in medical geography and interned at the World Health Organization regional office for southeast Asia (WHO SEARO) in New Delhi, India. Inspired by this internship, he wrote his diploma thesis about the use of emergency calls for the surveillance of infectious diseases in Andhra Pradesh, India. During this time, he came in contact with the numerous applications of Geographic Information Systems (GIS) in public health.

From 2010 to 2012 he worked as student research assistant and later from 2012 to 2014, he worked as researcher for Maastricht University under the supervision of Prof. dr. Thomas Krafft. At Maastricht University, he closely collaborated with the emergency medical dispatch centre in Tyrol, Austria to establish an early warning system for unusual health threats and conducted GIS-based analyses of emergency calls. Further areas of work included close collaboration with the South Limburg Public Health Service (GGD ZL) for the spatial analysis of infectious diseases. He also held lectures and provided trainings of GIS for students at Maastricht University.

In 2015, he joined the AOK Nordost – northeastern Germany`s largest statutory health insurance provider – where he currently works as expert consultant for the department of medical care. His main areas of expertise at the AOK Nordost consist of GIS-based analyses of chronic diseases, access to and planning of healthcare. For one of his analyses, he was awarded with the science award 2017 for regional healthcare research from the central institute of statutory health insurance physicians in Germany.

# Publications

**2017**

**Boris Kauhl**, Werner Maier, Jürgen Schweikart, Jonas Pieper, Andrea Keste, Marita Moskwyn (2017). Anwendungsgebiete und Limitierungen der amtlichen Statistik für die regionale Versorgungsforschung. Ein Diskussionsbeitrag der AOK Nordost am Beispiel der koronaren Herzkrankheit. *Zeitschrift für amtliche Statistik 3/2017*

**Boris Kauhl,** Jeanne Heil, Christian JPA Hoebe, Jürgen Schweikart, Thomas Krafft, Nicole HTM Dukers-Muijrers (2017). Is the current pertussis incidence only the results of testing? A spatial and space-time analysis of pertussis surveillance data using cluster detection methods and geographically weighted regression modelling. *PLoS ONE 12(3):e0172383.doi:10.1371/journal.pone.0172383*

**Boris Kauhl,** Jonas Pieper, Jürgen Schweikart, Andrea Keste, Marita Moskwyn (2017). Die räumliche Verbreitung des Typ 2 Diabetes Mellitus in Berlin–Die Anwendung einer geografisch gewichteten Regressionsanalyse zur Identifikation ortsspezifischer Risikogruppen. *Das Gesundheitswesen*.

Jonas Pieper, Ulrike Dapp, **Boris Kauhl,** Jürgen Schweikart (2017). Ageing in the Spatial Context – GIS Analyzes of the Longitudinal Urban Cohort Ageing Study. *AGIT – Journal für Angewandte Geoinformatik· 3-2017*

Eva Pilot Ramana Rao, Biranchi Jena, **Boris Kauhl,** Thomas Krafft, Murthy VS Gudlavalleti (2017). Towards Sustainable Public Health Surveillance in India: Using Routinely Collected Electronic Emergency Medical Service Data for Early Warning of Infectious Diseases. *Sustainability*, *9*(4), 604.

**2016**

**Boris Kauhl,** Jürgen Schweikart, Thomas Krafft, Andrea Keste, Marita Moskwyn (2016). Do the risk factors for type 2 diabetes mellitus vary by location? A spatial analysis of health insurance claims in Northeastern Germany using kernel density estimation and geographically weighted regression. *International Journal of Health Geographics, 15(1), 38.*

**2015**

**Boris Kauhl,** Jeanne Heil, Christian JPA Hoebe, J., Jürgen Schweikart, Thomas Krafft, Nicole HTM Dukers-Muijrers (2015). The spatial distribution of hepatitis C virus infections and associated determinants—An application of a geographically weighted poisson regression for evidence-based screening interventions in hotspots. *PloS one*, *10*(9), e0135656.

**Boris Kauhl,** Eva Pilot, Ramana Rao, Oliver Grübner, Jürgen Schweikart, Thomas Krafft (2015). Estimating the spatial distribution of acute undifferentiated fever (AUF) and associated risk factors using emergency call data in India. A symptom-based approach for public health surveillance. *Health & place, 31, 111-119.*

**2014**

Alexandra Ziemann, Nicole Rosenkoetter, Luis Garcia-Castrillo Riesgo, Sabrina Schrell, **Boris Kauhl,** Gernot Vergeiner, Matthias Fischer, Freddy K. Lippert, Alexander Kramer, Helmut Brand, Thomas Krafft (2014). A Concept for Routine Emergency-care Date-based Syndromic Surveillance in Europe. *Epidemiology and Infection. Jan. 24: pp. 1-14*

**2012**

Alexandra Ziemann, Luis Garcia-Castrillo Riesgo, Boris, **Kauhl,** Sabrina Schrell, Nicole Rosenkötter, Matthias Fischer, Gernot Vergeiner, Jean-Bernard Gillet, Agnes Meulemanns, Thomas Krafft (2012). Added value of routine emergency medical data for detecting clusters of acute gastrointestinal illness in Europe. *Resuscitation*, *83*, e30.

**2010**

Nicole Rosenkötter, **Boris Kauhl,** Luis Garcilla-Castrillo Riesgo, Francisco Javier Llorca Diaz, Janneke Kraan, Alexandra Ziemann, Martina Schorbahn, Thomas Krafft, Helmut Brand (2010). Retrospective data analysis and simulation study as basis for an automated syndromic surveillance system - Results from the SIDARTHa project. *Bad Honnef.*

# Valorisation of this thesis

**Relevance of research results**

The innovative aspect of this research is the provision of new insights on the importance of geographic aspects for demand-based planning and allocation of healthcare and targeted prevention strategies. The insight that geographic aspects are important determinants of health and disease has broad implications that are useful beyond science.

There are three areas where these research results could improve public health policies beyond academia:

1. Demand-based planning and allocation of healthcare
2. Improvement of current prevention strategies
3. Improvement of public health surveillance through implementation of geographic information systems and spatial epidemiological methods

**Demand-based planning and allocation of healthcare**

The case study on type 2 Diabetes Mellitus in Germany is important for demand-based planning and allocation of healthcare in northeastern Germany. This case study is part of a larger and on-going project between the AOK Nordost health insurance and the Beuth University of Applied Sciences to enhance the current planning and allocation of primary healthcare in northeastern Germany. The current planning of general practitioners (GPs) is still based on a target radio of 1671 inhabitants per GP and does not acknowledge a higher prevalence of chronic diseases in socially disadvantaged areas nor the accessibility of GPs in rural areas. Since the planning and provision of GPs is planned between health insurance providers and the association of statutory health insurance physicians, health insurance providers have a high interest to detect an increased medical demand of their insurants to provide healthcare where it is needed most. This in turn decreases the chance of expensive and possibly avoidable complications. Logically, spatial analyses of chronic diseases are important for evidence-based negotiations where new GPs should be allocated. Providing background knowledge about geographic determinants of chronic diseases thus helps to model the expected demand for healthcare in the future. Detecting areas with increased medical demand and identifying associated determinants of chronic diseases is not only important for type 2 Diabetes Mellitus, but also for several other chronic diseases with high prevalence rates such as hypertension or cardiovascular diseases. The knowledge that a lower socio-economic status is a strong determinant for chronic diseases

facilitates the current debate about including additional population-based variables such as area deprivation into planning of healthcare. The discussion about the consideration of area deprivation as driving factor for healthcare needs is still relatively young in Germany when compared to other countries.

**Improvement of current prevention strategies**

The results of the four case studies are relevant for prevention strategies for different actors in the healthcare sector:

1. **General practitioners**

Disseminating the results of the case studies on hepatitis C and type 2 Diabetes Mellitus to general practitioners in the respective region could facilitate the implementation of preventive screenings. Although hepatitis C and type 2 Diabetes Mellitus are fundamentally different, these two diseases have one particular characteristic in common: Both diseases are often asymptomatic in the beginning and have a high probability of adverse health outcomes if they remain undetected. For both diseases, preventive screening could help to provide early diagnosis and necessary medical care and could thus minimize the risk of potentially preventable complications – such as liver cirrhosis in the case of hepatitis C and lower extremity amputations in the case of type 2 Diabetes Mellitus. The onset of type 2 Diabetes Mellitus could even be prevented or delayed if possible pre-diabetic conditions such as glucose intolerance are detected early enough. Providing GPs with background knowledge about local clusters and location-specific risk groups for these diseases could therefore enable GPs to offer free testing to patients belonging to local risk groups.

2. **Health insurance providers**

The AOK Nordost health insurance is the second key actor for whom the results of the case study on type 2 Diabetes Mellitus are relevant. Early diagnosis of Diabetes is likely to reduce the amount of potentially preventable complications. However, demographic and socio-economic risk factors remain largely unknown within the database. The AOK Nordost can benefit in several ways from the results of this case study: They provide important information about the location-specific risk groups. This information can be used to invite insurants with similar socio-demographic risk factors for preventive screenings. Additionally, health insurance providers in Germany have

intensive care programs for chronic diseases – the so-called disease management programs - where participating GPs provide intensive care for insurants with one or several chronic diseases. The areas highlighted as local clusters for type 2 Diabetes Mellitus can help the AOK Nordost to invite GPs to participate in disease management programmes for type 2 Diabetes Mellitus. This approach might ultimately benefit insurants by reducing the risk of potentially preventable complications through planning and allocation of more specialized and intensive medical care.

### 3. Local public health agencies

The results of the case study on hepatitis C have been used by the public health service South Limburg (GGD ZL) to facilitate preventive screening programmes in the areas highlighted as clusters in Maastricht. This shows the importance of a spatial ecological analysis for practical prevention strategies. By disseminating the results of the case studies for acute undifferentiated fever, hepatitis C and pertussis to other local public health agencies, public health departments in the Netherlands and India can be motivated to conduct similar studies. It is likely that the risk factors may be different in other regions. Enabling local public health agencies to assess location-specific risk groups for prevalent diseases within their jurisdiction could be an effective way to enhance preventive strategies, tailored towards local needs.

## Improvement of public health surveillance through implementation of geographic information systems and spatial epidemiological methods

The three case studies based on surveillance data in India and the Netherlands have clearly demonstrated the added value of analysing surveillance data geographically, ranging from an analysis of location-specific risk groups in the case of hepatitis C, the analysis of possible determinants of testing to the identification of possible disease outbreaks in space and space-time for pertussis. With geographic identifiers becoming increasingly available in datasets generated by surveillance systems, important opportunities for prevention and effective public health response emerge from the spatial analysis of surveillance data. All analytical methods used in this dissertation were available in sophisticated open-source software such as SaTScan, GeoDa, CrimeStat and GWR4. With the increasing additional availability of spatial statistical methods within the R programming language and enhancements in open source GIS software such as Quantum GIS, spatial epidemiological methods can be easily

implemented even in public health departments with limited financial resources. This is important for GGD ZL as it empowers GGD ZL to conduct analyses of different diseases and develop further prevention strategies. Communicating these possibilities to other public health departments also could be an effective way to make further use of the potential of spatially referenced surveillance data for effective public health strategies.