



VCU

Virginia Commonwealth University
VCU Scholars Compass

Biostatistics Publications

Dept. of Biostatistics

2015

Evaluation of the Performance of Smoothing Functions in Generalized Additive Models for Spatial Variation in Disease

Umaporn Siangphoe
Virginia Commonwealth University

David C. Wheeler
Virginia Commonwealth University, dcwheeler@vcu.edu

Follow this and additional works at: http://scholarscompass.vcu.edu/bios_pubs



Part of the [Medicine and Health Sciences Commons](#)

Copyright © 2015 the author(s), publisher and licensee Libertas Academica Ltd. This is an open-access article distributed under the terms of the Creative Commons CC-BY-NC 3.0 License.

Downloaded from

http://scholarscompass.vcu.edu/bios_pubs/32

This Article is brought to you for free and open access by the Dept. of Biostatistics at VCU Scholars Compass. It has been accepted for inclusion in Biostatistics Publications by an authorized administrator of VCU Scholars Compass. For more information, please contact libcompass@vcu.edu.

Evaluation of the Performance of Smoothing Functions in Generalized Additive Models for Spatial Variation in Disease

Umaporn Siangphoe¹ and David C. Wheeler¹

¹Department of Biostatistics, Virginia Commonwealth University, Richmond, VA, USA.

Supplementary Issue: Computer Simulation, Bioinformatics, and Statistical Analysis of Cancer Data and Processes

ABSTRACT: Generalized additive models (GAMs) with bivariate smoothing functions have been applied to estimate spatial variation in risk for many types of cancers. Only a handful of studies have evaluated the performance of smoothing functions applied in GAMs with regard to different geographical areas of elevated risk and different risk levels. This study evaluates the ability of different smoothing functions to detect overall spatial variation of risk and elevated risk in diverse geographical areas at various risk levels using a simulation study. We created five scenarios with different true risk area shapes (circle, triangle, linear) in a square study region. We applied four different smoothing functions in the GAMs, including two types of thin plate regression splines (TPRS) and two versions of locally weighted scatterplot smoothing (loess). We tested the null hypothesis of constant risk and detected areas of elevated risk using analysis of deviance with permutation methods and assessed the performance of the smoothing methods based on the spatial detection rate, sensitivity, accuracy, precision, power, and false-positive rate. The results showed that all methods had a higher sensitivity and a consistently moderate-to-high accuracy rate when the true disease risk was higher. The models generally performed better in detecting elevated risk areas than detecting overall spatial variation. One of the loess methods had the highest precision in detecting overall spatial variation across scenarios and outperformed the other methods in detecting a linear elevated risk area. The TPRS methods outperformed loess in detecting elevated risk in two circular areas.

KEYWORDS: generalized additive model (GAM), smoothing functions, spatial analysis, simulation, cancer, cluster

SUPPLEMENT: Computer Simulation, Bioinformatics, and Statistical Analysis of Cancer Data and Processes

CITATION: Siangphoe and Wheeler. Evaluation of the Performance of Smoothing Functions in Generalized Additive Models for Spatial Variation in Disease. *Cancer Informatics* 2015;14(S2) 107–116 doi: 10.4137/CIN.S17300.

RECEIVED: November 16, 2014. **RESUBMITTED:** January 14, 2015. **ACCEPTED FOR PUBLICATION:** January 22, 2015.

ACADEMIC EDITOR: JT Efrid, Editor in Chief

TYPE: Original Research

FUNDING: DCW was supported by the National Cancer Institute of the National Institutes of Health under Award Number R03CA173823. The authors confirm that the funder had no influence over the study design, content of the article, or selection of this journal.

COMPETING INTERESTS: Authors disclose no potential conflicts of interest.

CORRESPONDENCE: dcwheeler@vcu.edu

DISCLAIMER: The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

COPYRIGHT: © the authors, publisher and licensee Libertas Academica Limited. This is an open-access article distributed under the terms of the Creative Commons CC-BY-NC 3.0 License.

Paper subject to independent expert blind peer review by minimum of two reviewers. All editorial decisions made by independent academic editor. Upon submission manuscript was subject to anti-plagiarism scanning. Prior to publication all authors have given signed confirmation of agreement to article publication and compliance with all applicable ethical and legal requirements, including the accuracy of author and contributor information, disclosure of competing interests and funding sources, compliance with ethical requirements relating to human and animal study participants, and compliance with any copyright requirements of third parties. This journal is a member of the Committee on Publication Ethics (COPE).

Published by Libertas Academica. Learn more about this journal.

Introduction

The use of spatial analytic techniques has substantially advanced in the last two decades due to increasing access to specialized software and increased computing capacity.¹ Modern analytic techniques enable researchers to address public health concerns regarding environmental and other risk factors associated with spatial variation in a variety of diseases, including cancers.² According to a recent review for research related to geographical information system studies, cancers are the most common noninfectious diseases investigated for spatial variation in the literature.¹ The studies usually focus on spatial distribution of cancer cases and underlying risk factors influencing the spatial distribution. In our review, many previous studies have illustrated the applications of spatial analysis on cancer data such as female oral cancer mortality in the United States³; breast cancer mortality in Northeastern United States³; lung, colorectal, and breast cancers in upper Cape Cod, Massachusetts^{2,4}; lung cancer incidence in Kentucky, the United States⁵; cervical cancer mortality in United States⁶; colorectal

cancer survival in New Jersey, United States⁷; prostate cancer in relation to environmental carcinogens in Great Britain, United Kingdom⁸; childhood leukemia incidence in Ohio⁹; childhood acute leukemia incidence in France¹⁰; childhood acute lymphoblastic leukemia in Hungary¹¹; non-Hodgkin lymphoma (NHL) incidence related to distances of benzene release sites in Georgia¹²; oncological mortality in a Municipality of North-Western Italy Vercelli¹³; and spatial-temporal analysis of cancer risk such as female breast cancer mortality in Spain¹⁴; breast cancers in upper Cape Cod, Massachusetts^{15,16}; NHL in National Cancer Institute (NCI) – Surveillance, Epidemiology, and End Results (SEER) case-control study^{17,18}; and in a Danish case-control study.¹⁹

A number of different statistical approaches have been applied in the past decade to evaluate spatial clustering or detect geographic areas of elevated risk in cancer data.² For example, Bonetti and Pagano's M statistic is a nonparametric statistic that compares expected and observed values representing interpoint distances between cancer cases in the detection



of clustering.^{2,20–22} The Besag, York, and Mollié (BYM) model is a Bayesian area-level regression model with components for spatial correlation and nonspatial heterogeneity that has been used for detection high-risk areas, where posterior distributions for the area-level parameters are used to identify elevated risk areas.^{23–25} Tango's index, Moran's I, and Oden's I statistics are used in detecting clustering but are sensitive to outliers. Tango's MEET and Oden's I_{pop}^* perform well in detecting global clustering.²⁶ Kulldorff's spatial scan statistic is a widely used method for detecting areas of elevated risk and is implemented in the freely available software SaTScan.^{2,21,27} The local spatial scan statistic has been effective in epidemiologic studies at detecting areas of elevated disease risk, but it requires a stratified analysis to adjust for covariates.²¹

In addition to the previous statistical methods, generalized additive models (GAMs) with bivariate smoothing functions have been applied to evaluate spatial variation of disease risk and identify areas of elevated risk in many types of cancers.^{2,4,15,17,18,28} The GAM framework is commonly used to determine areas of increased and decreased risk for a response variable while adjusting for covariates, including spatial confounders.^{4,28} However, little is known in the literature regarding both absolute and relative performance of different smoothing functions applied in the spatial GAMs. The performance may differ due to different characteristics of elevated risk areas with respect to shape, size, location in a study area, and disease risk level or probability of disease. This study evaluates through a simulation study the performance of different smoothing functions applied in the GAM models for detecting overall spatial variation and elevated risk areas with different geographical characteristics that may realistically appear in spatial analysis studies of cancer.

Statistical Methods

Generalized additive models. A GAM is a semi-parametric method extended from a generalized linear model and has a general formula

$$g(\mu) = \alpha + \sum_{j=1}^p f_j(x_j) + \varepsilon, \quad (1)$$

where $g(\mu)$ is a link function, α is a model intercept, and $f_j(\cdot)$ are smoothing functions of covariates x_j for $j = 1, \dots, p$. GAMs can include component functions with two or more dimensions, categorical variable terms, or interaction terms with continuous variables. The model therefore allows nonlinear functions of covariates to be included in regression equations and avoid restrictions imposed by parametric assumptions.^{29,30} Specifically, for case-control data, GAMs can model a binary disease outcome through the log-odds of disease as a linear function of covariates z_i and a spatial smoothing over locations $\mathbf{x}_i = (x_{1i}, x_{2i})$ for $i = 1, \dots, n$, such as

$$\log[P(y_i = 1)/P(y_i = 0)] = \alpha + f(x_{1i}, x_{2i}) + \beta z_i, \quad (2)$$

where $y_i = 1$ for cases and $y_i = 0$ for controls, β is a vector of linear regression coefficients representing the effects of z_i , and $f(x_{1i}, x_{2i})$ is a bivariate smoothing function.²⁹ The model specified in equation (2) is also known as a generalized additive logistic model.

Smoothing functions. There are several options available for the functional form of the bivariate spatial smoother $f(\mathbf{x}_i)$.²⁸ The locally weighted scatterplot smoothing (loess) smoother has been applied previously in GAMs to estimate spatial variation in disease risk.^{2,4,15,17,18,21,24,28,31} In loess regression, the outcome variable y_i at location \mathbf{x}_i is regressed on a function of the data values of the locations within a neighborhood of \mathbf{x}_i , where the size of the neighborhood is controlled by the span parameter b , which is defined as the proportion of the total data points that are used to estimate y_i . The function used in the regression uses weights for data points that are a function of distance between \mathbf{x}_i and the neighboring data points and the span. In general, the weight w_{ij} between two points i and j is calculated as

$$w_{ij} = W\left[\frac{d_{ij}}{b}\right], \quad (3)$$

where $W(\cdot)$ is a weighting function and d_{ij} is the Euclidean distance between the points.^{28,29,32} The weighting function implemented in loess in the R computing environment that we used is the tricube, which is defined as

$$W(d, d_b) = \begin{cases} \left(1 - \left(\frac{d}{d_b}\right)^3\right)^3 & \text{for } d < d_b \\ 0 & \text{for } d \geq d_b \end{cases} \quad (4)$$

where d_b corresponds to the maximum distance associated with the span. Increasing the span increases the number of data points that receive nonzero weight and generally increases the smoothness of the response surface.^{2,29}

Several methods are available for choosing an optimal span or degree of smoothing. An optimal smoothing parameter can be selected using weighted least squares cross-validation (CV), where CV minimizes an average-squared prediction error.^{29,30,33} The smoothing parameter can also be chosen by the generalized cross-validation (GCV) method for fitting penalized generalized ridge regressions with unknown scale parameters.^{29,30,34} CV may be computationally intensive, while GCV can lead to overfitting in models.³⁴ Similar to the GCV method, minimizing unbiased risk error (UBRE) derived from an expected mean square error can be used when the scale parameter is known.³⁰ In GAMs with loess smoothers, the minimization of the Akaike Information Criterion (AIC), a measure of goodness of fit used to control the bias-variance tradeoff, has been used to select the optimal span.^{28,35} The practice of minimizing the AIC to select

the span is recommended in a previous comparative study of methods.³⁴ Because a local optimal span can possibly be detected instead of a global optimal value, it is good practice to evaluate a sequence of candidate span values in GAM models.^{29,30,34} GAM coefficients can be estimated by a local scoring algorithm, which is an iteratively weighted backfitting algorithm. The backfitting algorithm is a Gauss–Seidel algorithm for fitting the additive models by iteratively smoothing partial residuals defined as $R_i = Y - f_0 - \sum f_i(x_i)$. The estimation consists of two loops. The local scoring algorithm is an outer loop and the weighted backfitting algorithm is an inner loop. The local scoring algorithm is performed by replacing the weighted linear regression for the adjusted dependent variable by the weighted backfitting algorithm. The algorithm starts with initial estimates of f_0, \dots, f_n . For each step, an adjusted dependent variable and a set weight are computed, and the smoothing components are fit to their partial residuals using the weighted backfitting algorithm. The algorithm stops when the deviance of estimates stops decreasing.^{29,36,37} The local scoring algorithm is used in the R package *gam*.³⁷

In addition to the bivariate loess smoothers, some types of smoothing spline functions have been applied in GAM models, such as a thin plate regression spline (TPRS) for modeling spatial-temporal risk of NHL in the NCI-SEER case–control study¹⁸ and for estimating spatial variation in forest biomass and biomass change in lodgepole pine stands in Alberta, Canada.³⁸ A two-dimensional penalized spline was used for modeling the spatial distribution of occupational accident risk in a labor market in Piracicaba, Southeast Brazil.³⁹ In general, a spline is a piece-wise polynomial determined by a sequence of knots, $k(s)$. A smoothing spline approach is fitting a spline with knots on each data point $k \ll n$ observations. Because fitting regression spline models typically depends on choosing knot locations, penalized regression splines include a “wiggleness” penalty term of predictors into the least square objective function to avoid the issue of knot placements. A thin plate spline (TPS) is a type of penalized regression spline and provides knot-free locations. It is available for any number of predictors and is flexible for derivative order selections and has relatively low execution cost.³⁰ This smoothing spline has a wiggleness penalty term and smoothing parameter and can serve as a bivariate smoother of coordinate data in spatial analysis.³⁰

A TPRS, an extension of the TPS, truncates the space of wiggly components of the TPS and estimates parameters by minimizing the sum of square error and the penalty term. The TPRS therefore fits the model better than the TPS.³⁰ The objective function for TPRS fitting is

$$\min \|y - U_k D_k \boldsymbol{\delta}_k - T \boldsymbol{\alpha}\|^2 + \lambda \boldsymbol{\delta}_k^T D_k \boldsymbol{\delta}_k, \text{ subject to } T^T U_k \boldsymbol{\delta}_k = 0 \quad (5)$$

with respect to vectors of coefficients $\boldsymbol{\delta}_k$ and $\boldsymbol{\alpha}$, where U_k contains the first k columns of eigenvectors of the observed predictors, D_k is a $k \times k$ submatrix of diagonal matrix of eigenvalues, T contains linearly independent polynomials, and λ is a

smoothing parameter.³⁰ Additionally, because the smoothing terms cannot be dropped from the model or completely zeroed out when they are not significantly associated with the fit, we can apply shrinkage to a TPRS (TPRS-S) by adding an extra parameter with associated smoothing parameter for more possibility of smoothness than wiggleness. The smoothing parameters that do not contribute to the model can be dropped in such a case.³⁰ The smoothing parameter in a TPRS is estimated by UBRE methods. The model coefficients for GAMs with TPRS smoothers can be estimated by maximizing the penalized likelihood function using a penalized iteratively reweighted least squares (P-IRLS) algorithm.³⁰ More details can be found in Wood.³⁰

Evaluation of smoothing functions in GAMs. Various hypothesis testing procedures have been proposed for evaluating significant spatial variation in risk and detection of significant areas of elevated or lowered risk. In a GAM model, an association between a smoothed predictor and a binary outcome can be determined using a likelihood ratio statistic and approximate Chi-square distribution by testing a difference in deviance statistics for the models with and without the smoothed term.^{28–30} As the approximate Chi-square test can lead to an inflated type I error, a permutation test based on an empirical distribution has been proposed.^{28,30,40,41} Monte Carlo and bootstrap sampling methods can also be used.^{40,42} The performance of permutation testing for detecting significant spatial variation in risk with GAMs has been investigated in simulation studies.^{21,31,41,43} Generally, the permutation procedure is based on Monte Carlo randomization of case labels and associated covariates, conditioning on the number and location of observed points.⁴⁴ Randomization of labels is consistent with the null hypothesis of constant risk throughout the study area. The permutation procedure randomly assigns subjects (outcome status and covariates together) to the observed geographical locations. Difference in model deviance is calculated for models with and without the smoothing term for the observed data and each permuted data set. The differences from the observed and permuted data are then ranked in ascending order and P -values for overall significant spatial variation in risk are calculated by comparing the rank of observed and permuted data and dividing by the total number of data sets. Significant risk areas are identified by the spatial points that have spatial log-odds that are outside the 2.5% and 97.5% ranked values of the point-wise permutation distribution.^{28,41} Typically, a spatial grid is used for predicting and evaluating the spatial log-odds, but one could also consider using the observed data points.

In a previous study of span selection for the proportion of data points to use to estimate y_i in a bivariate loess in a GAM, a fixed-span permutation test (FSPT) assuming one span or a multiple span permutation test (FMSPT) assuming multiple possible span sizes were used and compared to the span defined by observed data prior to conducting permutations.⁴³ A conditional permutation test (CPT) identifies an optimal

span from the observed data and holds the span constant for all permutations, while an unconditional permutation test (UPT) uses the optimal span size identified from the observed data and each of the permuted data sets. Hence, the bandwidth estimate can change for each permuted data set with UPT. A previous study suggested that the FMSPT had lower power than the FSPT, whereas the FSPT power depended on the spans previously determined.⁴³ Also, the CPT showed an inflated type I error rate, and an adjustment to the nominal *P*-value was necessary. The nominal *P*-value was reduced to 0.025 to have an effective significance level of 0.05 with the CPT. The UPT had the correct type I error rate, but required a considerable computational effort.^{31,43}

When detecting local areas of elevated risk, a circular area of elevated risk is the most common shape identified in spatial statistics, although actual true shapes may not always be circular. True areas of elevated risk may align with unusually shaped geographical regions such as streets, rivers, mountain ranges, or cape regions. The performance of spatial statistical tests may differ according to the true shape of an elevated risk area in a particular location. A prior simulation study evaluated the performance of GAMs with a loess, a local spatial scan statistic, and a Bayesian disease mapping model (BYM) in detecting local areas with different risk levels in real geographical surfaces.²⁴ The GAM method provided the highest sensitivity, but had low specificity. All the methods had difficulty in detecting areas with low risk levels and irregular shapes.²⁴ Another study found that the elliptical local spatial scan statistic was robust in performance to the shape of the true risk area (elliptical or circular), but that the circular local spatial scan statistic had greater power to detect a circular area of risk.³ Recently, a simulation study comparing statistical power and sensitivity between GAM permutation methods and Kulldorff's local spatial scan statistic⁴³ found that the GAM methods outperformed the scan statistic in sensitivity in all simulation study scenarios: a single point source located






in a circular cluster, a line source centered on a horizontal axis of a square region, and a single circular cluster centered on a circular region. The GAM permutation methods also had greater power in the first two scenarios and had similar power for the last scenario.⁴³

Simulation Study

Study design. We created five scenarios with different true risk areas in a square study region to evaluate the performance of GAMs with different types of smoothing functions (Table 1). We created three scenarios with one area of elevated risk that was circular, linear, or triangular and two scenarios with two circular areas of elevated risk in the first and third quadrants. The single circular and single linear risk areas were centered in the study region, and the triangular risk area was located in a corner of the study region. The elevated risk areas covered 15% of the study region for the single elevated area, 5% and 10% for the two elevated risk areas with different sizes, and 10% for the two with the same size. We uniformly simulated data where the true parameters were the odds ratios (ORs) of the specified risk areas (OR: 0.5, 1.0, 1.5, 2.0, 2.5, 3.0, and 3.5) and the probability of disease outside the risk areas (*P*: 0.05 and 0.20). We randomly generated the geographical coordinates based on uniform distribution in the study areas. We generated 1,000 data sets each containing 1,000 observations for each set of parameters for each scenario (code available upon request). In total, 70,000 data sets were generated based on the given parameters in the five scenarios for all model implementations.

Evaluation of methods. We evaluated the performance of four smoothing functions in GAMs: loess with UPT, loess with CPT, TPRS, and TPRS-S to detect overall spatial variation and areas of elevated risk. The model parameters of the GAM models with loess smoothers were fitted using the local score algorithm, whereas the P-IRLS estimation method was used for GAMs with TPRS and TPRS-S smoothers.

Table 1. Scenarios and parameters for the simulation study.

SCENARIO	FIGURE	RISK AREA SHAPE	RISK AREA SIZE	PROBABILITY OF DISEASE	ODDS RATIO (OR)
1		Circular	15%	0.05, 0.2	OR ₁
2		Linear	15%	0.05, 0.2	OR ₁
3		Triangle	15%	0.05, 0.2	OR ₁
4		Circular	5%, 10%	0.05, 0.2	OR ₁ = OR ₂
5		Circular	10%, 10%	0.05, 0.2	OR ₁ vs OR ₂ *

Notes: Blue squares represent study regions. Red shapes represent areas of true elevated or decreased risk. OR₁: Odds ratio in the first risk area, OR₂: Odds ratio in the second risk area (in the third quadrant). Scenarios 1, 2, and 3 consist of OR₁: 0.5, 1.0, 1.5, 2.0, 2.5, 3.0, and 3.5. Scenario 4 consists of OR₁ = OR₂: 0.5, 1.0, 1.5, 2.0, 2.5, 3.0, and 3.5. *Scenario 5: OR₁ vs OR₂ were defined as [3.5 vs 1.75, 3.0 vs 1.5, 2.5 vs 1.25, 2.0 vs 1.0, 1.5 vs 1.0, 1.0 vs 1.0, and 0.5 vs 0.75].

For the loess smoothers, the optimal span was selected for the observed and permuted data for UPT, and for CPT, the observed data alone was selected. Smoothing parameters were estimated based on minimizing AIC with span sizes ranging from 0.05 to 1.00 for the two loess smoothers and using the GCV method for the two TPRS smoothers. The four methods were applied to the same simulated data sets. The two loess and two TPRS smoothing functions in the GAM models were analyzed using *gam*³⁷ and *mgcv* packages^{30,45–49} in the R programming environment, respectively.

We performed hypothesis testing under the null hypothesis of constant risk with the alternative hypotheses specified in the five scenarios. We applied the likelihood ratio statistic approximate Chi-square and Monte Carlo permutation tests. One thousand Monte Carlo randomizations were performed for all the statistical models on the data sets to test the null hypothesis of constant risk and identify any areas of significantly lowered or elevated risk. We calculated type I error rate of the two tests and power based on the Monte Carlo permutation test as suggested in previous studies.^{31,41,43} The type I error rate was defined as the probability of false positives under the null hypothesis.⁴¹ We also calculated approximate Chi-square *P*-values and Monte Carlo *P*-values for the overall null hypotheses. *P*-values less than 0.05 were defined as statistically significant for the loess UPT, TPRS, and TPRS shrinkage models. Based on the previously described evidence of an inflated type I error rate in the CPT method, *P*-values less than 0.025 were defined as statistically significant when using CPT. In addition to the nominal type I error, we calculated average *P*-values of the approximate Chi-square and Monte Carlo permutation tests on different true risk areas based on our simulated data.

We measured performance of the four smoothing functions in the GAM models in detecting local significant risk areas of elevated and decreased risk. We determined whether the methods could detect centroids of the true risk areas based on grid cells (called grid detection) and based on observed point locations (called point detection). We defined observed spatial log-odds greater and lower than 97.5% and 2.5% of permuted log-odds ranked as significantly elevated and decreased risk areas, respectively. The detection of two centroids simultaneously was defined as detection in the scenarios with two circular elevated risk areas.

Typically, one would consider the local significance of areas only if the overall null hypothesis of constant risk has been rejected.^{4,33} To explore this approach, we evaluated three different ways of measuring detection with a GAM. The first approach calculated power for rejecting the null hypothesis of overall constant risk using Monte Carlo permutation tests. A centroid-grid detection approach defined detection as identifying the centroid of the true risk area based on a grid spatial reference. The third approach defined detection as rejecting the null hypothesis of constant risk and identifying the centroid of the true risk area using the centroid-grid detection. We calculated the detection rate using these approaches.

The ability of the smoothing functions was also quantified using four performance measures: sensitivity, false-positive rate, accuracy rate, and precision rate. The sensitivity was defined as the proportion of observations inside the true risk area that were identified as being in a significant risk area, whereas the false-positive rate was defined as the proportion of observations outside the true risk area that were identified as being inside a significant risk area.⁵⁰ The accuracy rate was defined as the proportion of observations that were correctly identified as being inside or outside a true risk area,⁵¹ and the precision rate was defined as the proportion of observations correctly identified as being in a risk area among those correctly identified as being inside or outside a risk area.⁵¹ All statistical analyses were implemented in the R computing environment.

Results

Overall spatial variation of risk. The overall *P*-value is used to reject the null hypothesis of constant risk. When the true OR = 1 in the designated risk area, the null hypothesis should only be rejected on average 5 out of 100 times. The approximate Chi-square test applied at the defined nominal significance levels had inflated type I error rates 2.3, 3.4, and 3.6 times higher than the nominal levels in the CPT, TPRS, and UPT methods for all scenarios and had a rate 12.4 times higher in the TPRS-S smoothers (Table 2). On the contrary, the type I error rates using the Monte Carlo permutation test at nominal significance levels were more reasonable and appropriate. The CPT smoother with the redefined significance rejection level had the type I error rate closest to the value of 0.05 in the study area with probability of disease of 0.2. Likewise, the TPRS smoother had the most correct type I error rate in the study area when the probability of disease was 0.05.

The average Chi-square and Monte Carlo *P*-values are intended to indicate how well a method can detect significant variation in risk in the data. With the approximate Chi-square test, each method had an average *P*-value below 0.05 when the elevated areas had OR ≥ 2.5 in scenarios 1 and 3 and OR ≥ 3.0 in scenarios 2, 4, and 5 (Supplementary Table 1). The TPRS-S smoother was able to detect significant elevated risk with OR ≥ 2.0 and probability of disease was 0.2, but this is at least partially attributable to the inflated type I error rate of the approximate Chi-square test. With the Monte Carlo permutation test, each smoothing function had an average *P*-value below 0.05 when OR ≥ 3.0 in scenarios 1 and 3 and probability of disease was 0.2. Only UPT had an average *P*-value > 0.05 for OR ≥ 3.0 and probability of disease = 0.2 in scenario 4. Overall, all methods with the two types of tests were able on average to detect overall spatial variation when OR ≥ 3.5 with the 0.2 probability of disease for all scenarios, while they did not on average detect significant variation in risk with a probability of disease of 0.05 outside the true risk area.

Table 2. Type I error rate for GAM methods in the simulation study.

PROBABILITY OF DISEASE	APPROXIMATE Chi-square <i>P</i> -VALUES				MONTE CARLO <i>P</i> -VALUES			
	UPT	CPT	TPRS	TPRS-S	UPT	CPT	TPRS	TPRS-S
0.05	0.178	0.115	0.168	0.620	0.055	0.059	0.047	0.050
0.20	0.173	0.111	0.153	0.588	0.061	0.047	0.057	0.058

Note: Four smoothing functions in GAMs: loess with UPT, loess with CPT, TPRS, and TPRS-S.

As anticipated, the power estimates were increased with increased disease risk. With probability of disease of 0.2, all the methods had remarkable power to detect the overall spatial variation in scenarios 1, 3, and 4, and satisfactory power in scenarios 2 and 5 when risk was highest (Fig. 1). The two TPRS methods had similar power in all scenarios. CPT with a significance level of 0.025 had the highest power to detect the overall spatial variation in scenarios 1 and 3. The two TPRS methods also had higher power than the UPT in these scenarios. Power estimates were decreased when the disease probability outside the true risk area was 0.05 for all risk levels in all scenarios (Supplementary Fig. 1). Nevertheless, the trends in power among methods were similar across the two levels of disease probability.

Detection of risk areas. We describe here the results of the detection of true risk areas based on a grid spatial reference when the probability of disease was 0.2. Similar to power

estimates, the detection rate of true risk areas was increased with increased risk for all methods and all scenarios. All methods performed best in scenario 1 and scenario 3, where the detection rate was over 90% in the elevated risk areas with $OR \geq 2.5$. The two TPRS methods and UPT were able to identify the true risk areas better than the CPT in scenarios 4 and 5 (two circular risk areas). In addition, the two TPRS methods and UPT performed better when the elevated risk areas had the same risk instead of different risk. However, the CPT outperformed the other methods in detecting the linear elevated risk area (scenario 2), which all methods had difficulty in detecting (Fig. 1 and Supplementary Fig. 1).

When considering the different approaches to detection, the CPT with a significance level of 0.025 had better performance than the other methods in the individual risk areas (scenarios 1, 2, and 3), whereas the two TPRS smoothers had detection rates better than the two loess smoothers in the

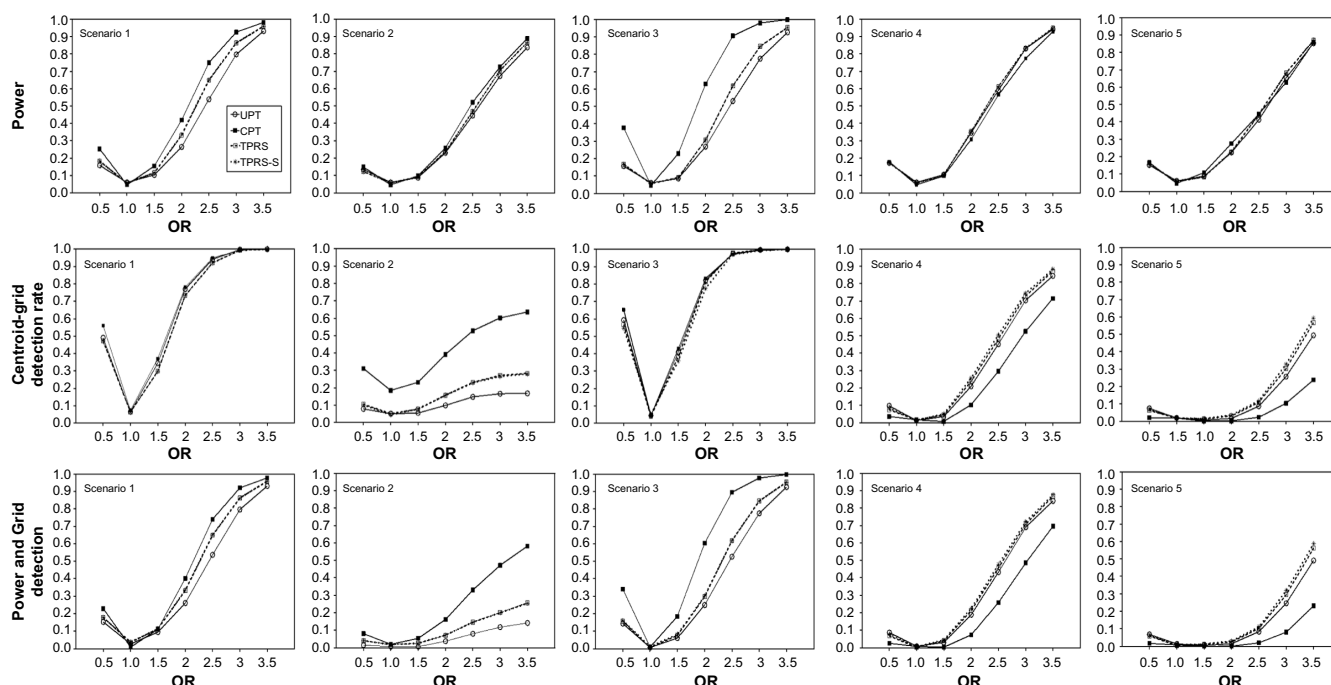


Figure 1. Power and centroid-grid detection of four smoothing functions in GAMs at different true ORs in risk areas. The power for rejecting the null hypothesis of overall constant risk was calculated based on Monte Carlo permutation tests. Centroid-grid detection for true risk areas was defined as detecting the centroid of the true risk area based on a grid spatial reference. The proportion of the detection by the power and centroid-grid detection together are shown in the third row. The detection of two centroids simultaneously was defined for the detection in the two circular elevated risk areas. Four smoothing functions in GAMs: loess with UPT, loess CPT, and TPRS and TPRS-S. Probability of disease unexposed risk was 0.2. Average values over 1,000 data sets are presented. Results of true risk areas in the first quadrant were evaluated for scenarios 4 and 5. The grid detection was not meaningful under the null hypothesis ($OR = 1.0$).

double true risk areas (scenarios 4 and 5). The detection rates in the scenarios 2, 4, and 5 were slightly lower when insisting on rejecting the null hypothesis (Fig. 1 and Supplementary Fig. 1).

As expected, all GAMs had a higher sensitivity in detecting elevated risk areas when the true risk was higher. Focusing on the results when the probability of disease was 0.2, all methods had higher sensitivity in scenarios 1 and 3, lower sensitivity in scenario 5, and similar sensitivity in scenarios 2 and 4. The two TPRS methods were able to detect the true risk area with more than 90% sensitivity when the $OR \geq 2.5$ in scenario 3 and $OR \geq 3$ in scenario 1. The two TPRS smoothers also had higher sensitivity than the two loess smoothers for $OR \geq 2$ in scenarios 1 and 5 and $OR \geq 1.5$ in scenarios 2 and 4 (Fig. 2). The two TPRS methods had a higher sensitivity than the CPT method in the areas with the highest risk. The UPT and two TPRS methods had very similar sensitivity in scenario 3 (Fig. 3).

The false-positive rate of detection for the methods was generally low (Fig. 2). It was typically below 10% for $OR \leq 2.0$ and below 20% for $OR > 2.0$. The false-positive rate tended to increase with increasing true risk, which was also when the sensitivity was higher. Hence, the two TPRS models had higher false-positive rates than the two loess models across the scenarios. The one exception was scenario 3, where

the methods had nearly identical false-positive rates, likely a consequence of them having nearly equal sensitivity.

As with sensitivity, the four types of GAMs had the highest accuracy when the true risk area had $OR \geq 2$ in scenarios 1 and 3 and $OR \geq 3$ in scenario 4 (Fig. 2). All methods in scenario 3 had similarly high accuracy with $OR = 3.5$ (Fig. 3) and also had a sufficiently high accuracy rate with $OR \geq 1.5$ or 2 in the other scenarios. All the smoothers had a consistently moderate-to-high accuracy rate when the true risk was higher ($OR \geq 1.5$). Similar results were found across the true risk areas with the probability of disease being either 0.05 or 0.2 (Supplementary Fig. 2).

Regarding precision, most methods had a higher rate of precision when true risk was higher (Fig. 2). However, in scenario 3, the methods had a decrease in precision when the OR of true risk area was ≥ 2 . This also occurred for the UPT and two TPRS methods in scenario 1. Comparing methods, the CPT method generally had a higher precision rate than the other methods in all scenarios. This method was also able to attain an acceptably high precision rate when the true risk area had $OR \geq 3$ for all scenarios, except for scenario 2, where it was difficult to detect the elevated risk area with a sufficient sensitivity (Figs. 2 and 3). The accuracy and precision rate overall were similar across the probabilities of disease. In other words, the accuracy and precision were not influenced by the disease

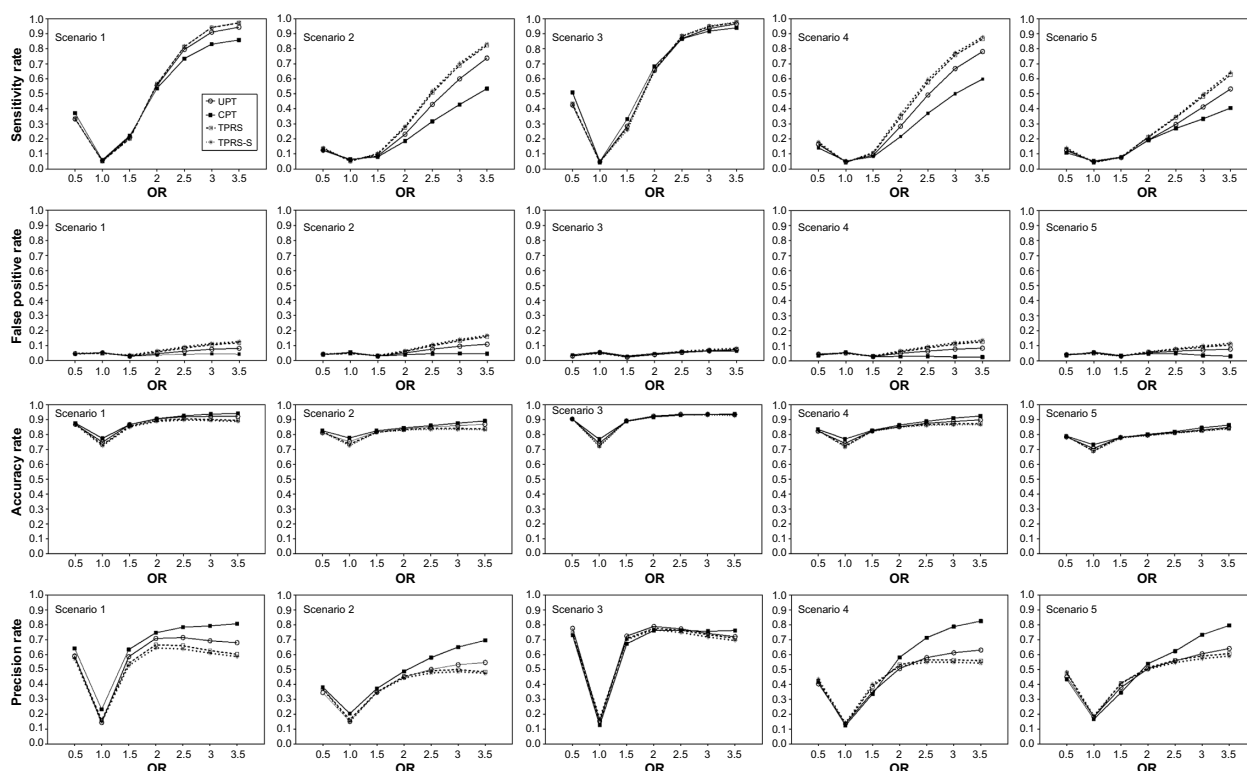


Figure 2. Sensitivity rate, false-positive rate, accuracy rate, and precision rate for detection of true risk areas with varying ORs based on a grid spatial reference. Four smoothing functions in GAMs: loess with UPT, loess CPT, and TPRS and TPRS-S. Probability of disease outside the true risk was 0.2. Averages over 1,000 data sets are shown in the figures. Results of true risk areas in the first quadrant were evaluated for scenarios 4 and 5. Performance rates were not meaningful under the null hypothesis ($OR = 1.0$).

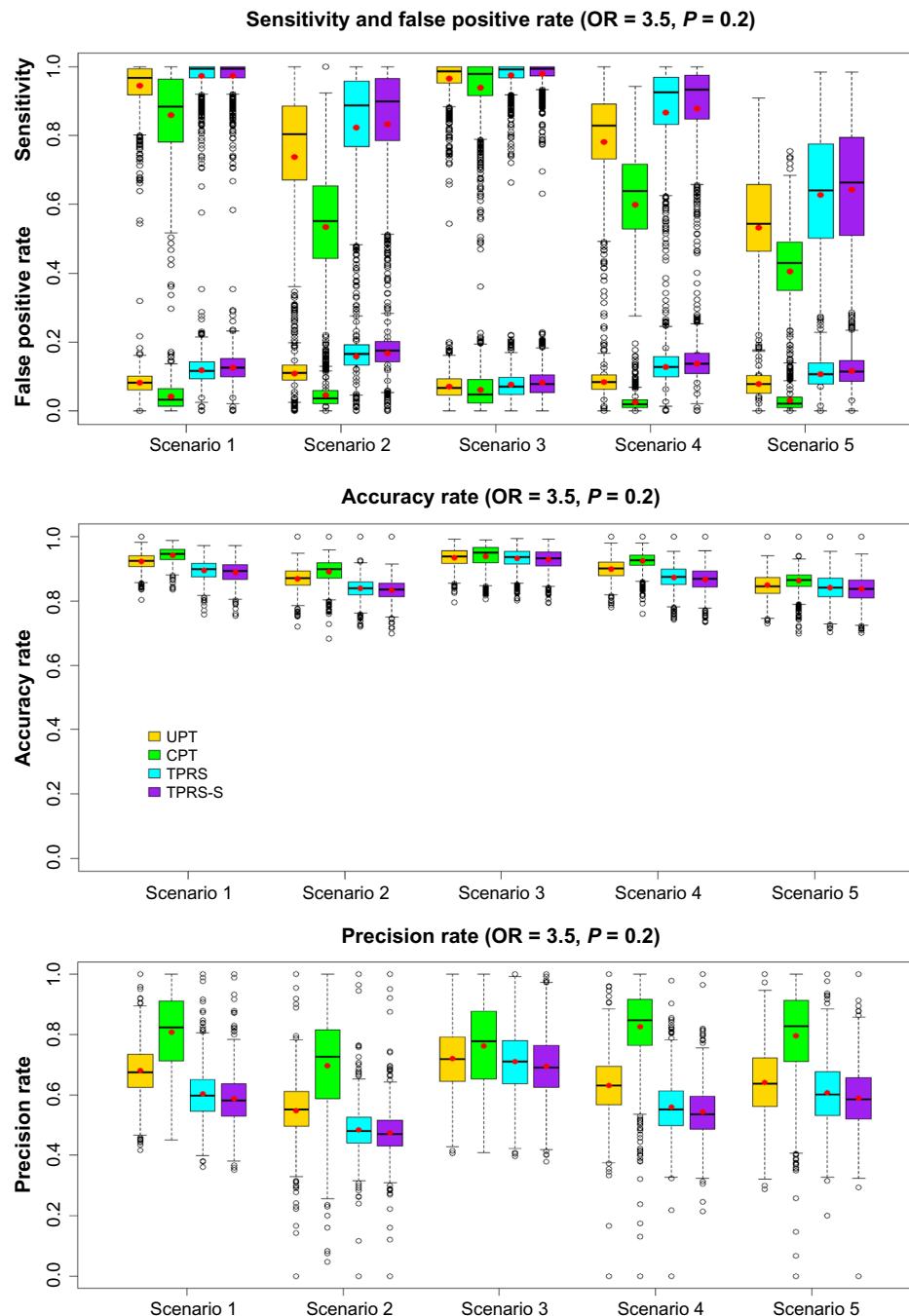


Figure 3. Boxplots of sensitivity, false-positive rate, accuracy rate, and precision rate for detection of elevated risk areas with an OR of 3.5 and probability of diseases outside the true risk (P) of 0.2 based on a grid spatial reference and 1,000 simulated data sets. Four smoothing functions in GAMs: loess with UPT, loess CPT, and TPRS and TPRS-S. Results of true risk areas in the first quadrant were evaluated for Scenario 4 and 5. Red dots represent mean values.

prevalence according to our results. In contrast, the sensitivity and false-positive rate were generally lower when probability of disease outside the true risk was lower. More details can be seen in Supplementary Figure 2.

In addition to the previous grid-based detection results, we calculated results using the observed points as the spatial reference. Focusing on the results when the probability of disease was 0.2, the detection rate based on point references was lower than the rate based on grid references for all the methods in all

scenarios. In particular, in scenarios 3, 4, and 5, the detection rate of elevated risk areas was much lower using point references compared with using grid references. Differences between the results from the grid and point-based approaches were less noticeable in scenarios 1 and 2 (Supplementary Fig. 3).

Discussion and Conclusion

Our large simulation study revealed several findings about the absolute and relative performance of four different types

of smoothing methods used in GAMs for evaluating overall spatial variation in risk and detection of true risk areas. Overall, all smoothing methods applied in the GAM models had appropriate type I error rates in all scenarios across the probabilities of disease after adjusting the nominal significance level of the CPT method. All methods performed well in detecting overall spatial variation and elevated risk in a single circular area centered in the middle of the study area (scenario 1) and in a triangular area in a corner of the study area (scenario 3). The methods displayed higher power, detection rates, and sensitivity in these scenarios compared with the other scenarios. Scenario 1 could represent an environmental exposure area with a point source located at the center, a, while scenario 3 could represent a large landfill or industrial waste site located on the periphery of a community. On the other hand, a linear area of elevated risk (scenario 2), which could represent elevated risk along a river or roadway, was more challenging to detect for all methods. Also, two elevated risk areas with equal size but different risks (scenario 4) were more difficult to detect than two circular areas with equal risk but different sizes (scenario 5). These scenarios could represent environmental exposures from multiple sources at a particular time or two different time periods.

Some of our findings agreed with those reported in previous studies. The ability of the two loess and TPRS smoothing functions in the GAMs were different depending on true risk area shapes, sizes, and locations. In our study, all methods easily detected a triangular area positioned in a corner of the study region. This indicates that in this scenario there were no significant edge effects impacting the performance of the GAM, which is similar to previous findings reported elsewhere.²⁸ To compare our results with previously published results, we designed scenario 1 to be similar to a scenario in a previous study.⁴³ Given the same risk and probability of disease outside the true risk area, we verified that our results regarding type I error, power, and sensitivity agreed with the previous findings.^{21,43}

Our study also supports that the CPT method with a loess smoother had an inflated type I error rate as suggested in previous studies.^{21,31,43} We found the inflated rates at all risks in all scenarios, except scenario 4 at $OR \geq 2$ and scenario 5 at $OR \geq 3$. The type I error rates in these cases were similar to those from the other methods. We retained the significance level of 0.025 for CPT in these cases because this threshold had no effect on power estimates and other evaluations. The 0.025 significance threshold was chosen in our study also to be comparable to the previous results.

In addition to significance thresholds, a recent study pointed out that the inflated type I error with the CPT method tended to occur when a small span was selected for the observed data. In this case, the difference in the model deviances with and without the spatial smooth was more inflated for the observed data compared with the distribution of the difference in model deviances based on permuted data that

reflect the null hypothesis of constant risk.³¹ In our study, as anticipated, small span sizes tended to be selected when the true risk was higher (results not shown).

Furthermore, the UPT was recognized as the most unbiased and appropriate smoothing method in a GAM.³¹ Our findings support that the UPT was an unbiased approach and had no inflated type I error. However, the UPT generally had lower power than the CPT method and required very large computation efforts. Using a Beowulf computer cluster, the computations for the UPT models for all our scenarios took almost a year to complete, whereas the two TPRS methods required lower execution time. The CPT method had the lowest execution time. Our results suggest that CPT with a significance level of 0.025 can be used as an alternative to UPT with a much lower computation cost. Our findings show that two TPRS methods performed similarly to the loess UPT. It was possible that the two TPRS methods and the loess UPT could give a similar result because they both controlled the bias–variance tradeoff using the smoothing parameters derived from the observed and permuted data, while the CPT controlled the tradeoff using the optimal parameter derived from the only observed data.

A strength of our study is that it is the largest and most comprehensive study comparing the performance of the loess and TPRS smoothing functions in GAMs in the context of spatial analysis. We considered a realistic set of scenarios for elevated risk with many levels of risk in a case–control study. An obvious limitation of our study is that we could not consider all possible exposure scenarios. The density of point locations can have a strong impact on performance of smoothing terms in GAMs.¹⁵ As we simulated point locations based on a uniform distribution, our results may not reflect those from other distributions. Our study was also limited to assessing the performance of smoothing functions in GAMs for spatial data at one time point. An assessment of GAMs for modeling risk in spatial-temporal data can be found in another study.⁵⁰

While not exhaustive in scope, our study confirms that the performance of smoothing functions in the GAM permutation methods in detecting overall spatial variation and elevated risk areas varies by different geographical areas of elevated risk and disease risk levels. Though some true risk area shapes are more difficult than others to detect, the performance of the GAMs overall was encouraging and we recommend their use in future studies.

Acknowledgments

The authors thank Robin Bliss for sharing initial computer code to simulate data in scenario 1 that they modified for this study.

Author Contributions

Conceived of the study: DCW. Designed the experiments: DCW, US. Conducted the experiments and analyzed the data:



US. Wrote the first draft of the manuscript: US. Contributed to the writing of the manuscript US, DCW. Made critical revisions: US, DCW. Agree with manuscript results and conclusions: US, DCW. Both authors reviewed and approved of the final manuscript.

Supplementary Data

Supplementary Figure 1.

Supplementary Figure 2.

Supplementary Figure 3.

Supplementary Table 1.

REFERENCES

- Lyseen AK, Nohr C, Sorensen EM, et al. A review and framework for categorizing current research and development in health related geographical information systems (GIS) studies. *Yearb Med Inform.* 2014;9(1):110–24.
- Ozonoff A, Webster T, Vieira V, Weinberg J, Ozonoff D, Aschengrau A. Cluster detection methods applied to the Upper Cape Cod cancer data. *Environ Health.* 2005;4:19.
- Kulldorff M, Huang L, Pickle L, Duczmal L. An elliptic spatial scan statistic. *Stat Med.* 2006;25(22):3929–43.
- Vieira V, Webster T, Weinberg J, Aschengrau A, Ozonoff D. Spatial analysis of lung, colorectal, and breast cancer on Cape Cod: an application of generalized additive models to case-control data. *Environ Health.* 2005;4:11.
- Christian W, Huang B, Rinehart J, Hopenhayn C. Exploring geographic variation in lung cancer incidence in Kentucky using a spatial scan statistic: elevated risk in the application coal-mining region. *Public Health Rep.* 2011;126(6):789–96.
- Chen J, Roth R, Naito A, Lengerich E, MacEachren A. Geovisual analytics to enhance spatial scan statistic interpretation: an analysis of U.S. cervical cancer mortality. *Int J Health Geogr.* 2008;7:57.
- Henry K, Niu X, Boscoe F. Geographic disparities in colorectal cancer survival. *Int J Health Geogr.* 2009;8:48.
- Jarup L, Best N, Toledano MB, Wakefield J, Elliott P. Geographical epidemiology of prostate cancer in Great Britain. *Int J Cancer.* 2002;97(5):695–99.
- Wheeler DC. A comparison of spatial clustering and cluster detection techniques for childhood leukemia incidence in Ohio, 1996–2003. *Int J Health Geogr.* 2007;6:13.
- Faure C, Mollie A, Bellec S, Guyot-Goubin A, Clavel J, Hemon D. Geographical variations in the incidence of childhood acute leukaemia in France over the period 1990–2004. *Eur J Cancer Prev.* 2009;18(4):267–79.
- Nyari TA, Ottoff G, Bartzyk K, et al. Spatial clustering of childhood acute lymphoblastic leukaemia in Hungary. *Pathol Oncol Res.* 2013;19(2):297–302.
- Bulka C, Nastoupil LJ, McClellan W, et al. Residence proximity to benzene release sites is associated with increased incidence of non-Hodgkin lymphoma. *Cancer.* 2013;119(18):3309–17.
- Salerno C, Berchialla P, Palin LA, Barasolo E, Vanhaecht K, Panella M. Geographical and epidemiological analysis of oncological mortality in a municipality of north-western Italy Vercelli years 2000–2009. *Ann Ig.* 2014;26(2):157–66.
- Ugarte MD, Goicoa T, Etxeberria J, Militino AF, Pollan M. Age-specific spatio-temporal patterns of female breast cancer mortality in Spain (1975–2005). *Ann Epidemiol.* 2010;20(12):906–16.
- Vieira VM, Webster TF, Weinberg JM, Aschengrau A. Spatial-temporal analysis of breast cancer in Upper Cape Cod, Massachusetts. *Int J Health Geogr.* 2008;7:46.
- Gallagher LG, Webster TF, Aschengrau A, Vieira VM. Using residential history and groundwater modeling to examine drinking water exposure and breast cancer. *Environ Health Perspect.* 2010;118(6):749–55.
- Wheeler DC, De Roos AJ, Cerhan JR, et al. Spatial-temporal analysis of non-Hodgkin lymphoma in the NCI-SEER NHL case-control study. *Environ Health.* 2011;10:63.
- Wheeler DC, Waller LA, Cozen W, Ward MH. Spatial-temporal analysis of non-Hodgkin lymphoma risk using multiple residential locations. *Spat Spatiotemporal Epidemiol.* 2012;3(2):163–71.
- Baastrop Nordsborg R, Meliker JR, Kjaer Ersboll A, Jacquez GM, Raaschou-Nielsen O. Space-time clustering of non-Hodgkin lymphoma using residential histories in a Danish case-control study. *PLoS One.* 2013;8(4):e60800.
- Ozonoff A, Bonetti M, Forsberg L, Pagano M. Power comparisons for an improved disease clustering test. *Comput Stat Data Anal.* 2005;48:679–84.
- Bliss RY, Weinberg J, Vieira V, Ozonoff A, Webster TF. Power of permutation tests using generalized additive models with bivariate smoothers. *J Biom Biostat.* 2010;1(104):1000104.
- Bonetti M, Pagano M. The interpoint distance distribution as a descriptor of point patterns, with an application to spatial disease clustering. *Stat Med.* 2005;24(5):753–73.
- Besag J, York J, Mollié A. Bayesian image restoration, with two applications in spatial statistics. *Ann Inst Stat Math.* 1991;43:1–20.
- Aamodt G, Samuelsen SO, Skrandal A. A simulation study of three methods for detecting disease clusters. *Int J Health Geogr.* 2006;5:15.
- Ramis R, Diggle P, Boldo E, Garcia-Perez J, Fernandez-Navarro P, Lopez-Abente G. Analysis of matched geographical areas to study potential links between environmental exposure to oil refineries and non-Hodgkin lymphoma mortality in Spain. *Int J Health Geogr.* 2012;11:4.
- Jackson MC, Huang L, Luo J, Hachey M, Feuer E. Comparison of tests for spatial heterogeneity on data with global clustering patterns and outliers. *Int J Health Geogr.* 2009;8:55.
- Song C, Kulldorff M. Power evaluation of disease clustering tests. *Int J Health Geogr.* 2003;2(1):9.
- Webster T, Vieira V, Weinberg J, Aschengrau A. Method for mapping population-based case-control studies: an application using generalized additive models. *Int J Health Geogr.* 2006;5:26.
- Hastie T, Tibshirani R. *Generalized Additive Models*. UK: CRC Press; 1990.
- Wood SN. *Generalized Additive Models: An Introduction with R*. UK: Chapman and Hall CRC; 2006.
- Young RL, Weinberg J, Vieira V, Ozonoff A, Webster TF. Generalized additive models and inflated type I error rates of smoother significance tests. *Comput Stat Data Anal.* 2011;55(1):366–74.
- Cleveland WS, Devlin SJ. Locally-weighted regression: an approach to regression analysis by local fitting. *J Am Stat Assoc.* 1988;83(403):596–610.
- Hastie T. *Gam: generalized additive models. R package version 1.09.1*. 2013.
- Kelsall JE, Diggle PJ. Spatial variation in risk of disease: a nonparametric binary regression approach. *J R Stat Soc Series C.* 1998;47(4):559–73.
- Krause R, Tutz G. Genetic algorithms for the selection of smoothing parameters in additive models. *Comput Stat.* 2006;21(1):9–31.
- Hurvich CM, Simonoff JS, Tsai C. Smoothing parameter selection in nonparametric regression using an improved Akaike information criterion. *J R Stat Soc.* 1998;60(2):271–93.
- Hastie T. Generalized additive models. In: Chambers SJM, Hastie TJ, eds. *Statistical Models*. (Chap. 7). Pacific Grove: Wadsworth; 1991:377–419.
- Monserud RA, Huang S, Yang Y. Biomass and biomass change in lodgepole pine stands in Alberta. *Tree Physiol.* 2006;26(6):819–31.
- Bailey TC, Cordeiro R, Lourenco RW. Semiparametric modeling of the spatial distribution of occupational accident risk in the casual labor market, Piracicaba, Southeast Brazil. *Risk Anal.* 2007;27(2):421–31.
- Tusell F. A permutation test for randomness with power against smooth variation. *Stat Comput.* 2001;11:147–54.
- Bliss RL, Weinberg J, Vieira VM, Webster TF. Adjusted significance cutoffs for hypothesis tests applied with generalized additive models with bivariate smoothers. *Spat Spatiotemporal Epidemiol.* 2011;2(4):291–300.
- Härdle W, Huet S, Mammen E, Sperlich S. Bootstrap inference in semiparametric generalized additive models. *Econometric Theory.* 2004;20:265–300.
- Young RL, Weinberg J, Vieira V, Ozonoff A, Webster TF. A power comparison of generalized additive models and the spatial scan statistic in a case-control setting. *Int J Health Geogr.* 2010;9:37.
- Waller L, Gotway C. *Applied Spatial Statistics for Public Health Data*. Hoboken, NJ: John Wiley & Sons, Inc; 2004.
- Wood S. *mgcv: mixed GAM computation vehicle with GCV/AIC/REML smoothness estimation (1.7–28)*. Available at: <http://Cran.r-project.org/web/packages/mgcv/index.html>. 2014.
- Wood SN. Thin-plate regression splines. *J R Stat Soc Series B.* 2003;65(1):95–114.
- Wood SN. Fast stable restricted maximum likelihood and marginal likelihood estimation of semiparametric generalized linear models. *J R Stat Soc Series B.* 2011;73(1):3–36.
- Wood SN. Stable and efficient multiple smoothing parameter estimation for generalized additive models. *J Am Stat Assoc.* 2004;99:673–86.
- Wood SN. Modelling and smoothing parameter estimation with multiple quadratic penalties. *J R Stat Soc Series B.* 2000;62(2):413–28.
- Wheeler DC, Siangphoe U. Modeling spatial variation in disease risk in epidemiologic studies. In: Kanaroglou P, Delmelle E, Páez A, eds. *Spatial Analysis in Health Geography*. first ed. UK: Ashgate; 2015.
- Fawcett T. An introduction to ROC analysis. *Pattern Recognit Lett.* 2006;27(8):14.