UNIVERSITY OF CALIFORNIA,
IRVINE

Usage of Kernel Smoothing in Generalized Additive Models for Disease Mapping with Individual-level Point-referenced Data: Stratified Smoothers and Generalized Additive Mixed Models

DISSERTATION

submitted in partial satisfaction of the requirements
for the degree of

DOCTOR OF PHILOSOPHY

in Statistics

by

Yannan Tang

Dissertation Committee:
Professor Daniel L. Gillen, Chair
Professor Michele Guindani
Professor Veronica M. Vieira
Professor Scott M. Bartell

2020

TABLE OF CONTENTS

LIST OF FIGURES

Page

LIST OF TABLES

Page

LIST OF ALGORITHMS

Page

ACKNOWLEDGMENTS

I would like to thank ...

CURRICULUM VITAE

Yannan Tang

EDUCATION

University of California, Irvine — **Irvine, CA, USA**
Ph.D. in Statistics — *2015 - 2020*
Dissertation Advisor: Professor Daniel L. Gillen

The George Washington University — **Washington, DC, USA**
M.S. in Statistics — *2013 - 2015*

Tsinghua University — **Beijing, China**
B.E. in Civil Engineering — *2008 - 2012*

PUBLICATIONS

- **Tang Y.**, Vieira V., Bartell S. and Gillen D., "A Stratified Generalized Additive Model and Permutation Test for Temporal Heterogeneity of Smoothed Bivariate Spatial Effects" (revised and submitted to *Statistics in Medicine*)
- **Tang Y.**, Vieira V., Bartell S. and Gillen D., "Additive Mixed Models with Kernel Smoothers for Disease Mapping Using Individual-level Data" (Submitted)
- **Tang Y.**, Vieira V., Bartell S. and Gillen D., "Disease Mapping using Generalized Additive Mixed Models with Kernel Smoothers" (To submit)

COLLABORATIVE RESEARCH

- Spatio-temporal analysis of birth defects and infant morbidity in relation to air pollution using generalized additive models (GAM) in a geographic framework
  Grant Funding Number: P42ES007381, NIEHS Grant, NIH
  PI: Veronica Vieira, D.Sc., Professor of Public Health, UC Irvine
  Role: Research Assistant
- Leveraging external data for regulatory decision making using propensity scores, with application in label expansion for multiple medical devices
  Sponsor: Allergan plc
  Supervisor: Jingyuan Yang, Ph.D., Director, Biostatistics, Allergan plc

TEACHING EXPERIENCE

University of California, Irvine **Irvine, CA, USA**
Tutor for Ph.D. qualification exams *2017 - 2019*

University of California, Irvine **Irvine, CA, USA**
Teaching Assistant, Reader *2015 - 2019*

CONTRIBUTED PRESENTATIONS AT ACADEMIC MEETINGS

- 13th International Conference on Health Policy Statistics, "An Additive Linear Mixed-effects Model (ALMM) with Kernel Smoothers and a Permutation Test on Temporal Heterogeneity of Geospatial Risk Patterns" (San Diego, CA, USA; JAN 2020)
- Joint Statistical Meetings of the ASA, "Time-Stratified LOESS Smoothers for Estimating and Testing Temporal Heterogeneity in Spatial Risk Patterns" (Vancouver, Canada; JUL 2018)

PROFESSIONAL MEMBERSHIPS

- American Statistical Association (2017-present)
- International Chinese Statistical Association (2018-present)

DEPARTMENT SERVICE

Department of Statistics **UC Irvine**
Graduate Student Representative (elected) *2017 - 2018*

AWARDS

- Early Advancement Award, Department of Statistics, UC Irvine (2017)
- University Scholarship, The George Washington University (2014, 2015)

ABSTRACT OF THE DISSERTATION

Usage of Kernel Smoothing in Generalized Additive Models for Disease Mapping with Individual-level Point-referenced Data: Stratified Smoothers and Generalized Additive Mixed Models

By

Yannan Tang

Doctor of Philosophy in Statistics

University of California, Irvine, 2020

Professor Daniel L. Gillen, Chair

Epidemiologists frequently aim to quantify geospatial heterogeneity in disease occurrence to identify relevant hidden health disparities. With the growing prevalence of individual-level point-referenced data, generalized additive models (GAMs) are becoming increasingly popular to map geospatial disease risk patterns while adjusting for confounding effects when the study is a cross-sectional one with an exponential family response. In the meanwhile, local regression smoothers are frequently adopted for spatial effects estimation in GAM framework by researchers partially due to their intuitive ideas and adaptation to changing population density.

However, studies with records over a (potentially long) period of time, including those with repeated measurements on subjects, commonly come into play nowadays. For these studies, traditional GAMs could be problematic. Firstly, since data could be recorded over a period of time while spatial risk patterns should not be assumed to be invariant in many cases, statistical tools to access time-varying spatial effects are required. On the other hand, if the study is longitudinally designed, traditional GAMs could lead to incorrect inference due to their incapability of accomodating within-individual correlation.

This dissertation work sought to develop statistical methodologies to address these problems under the GAM framework with kernel smoothers, using local regression smoothers in particular. In Chapter 3, we proposed GAMs with stratified kernel smoothers that could be applied for time-specific spatial effects modeling. Based on the new class of GAMs, we further designed a hypothesis testing procedure to formally detect temporal heterogeneity of spatial effects. In Chapter 4 and 5, we incorporated random effects, as well as kernel smoothers, into GAM, resulting in a class of generalized additive mixed models (GAMMs) with kernel smoothers. We further elaborated the novel fitting and inference procedures for the proposed models.

Relevant empirical results showed the utility and advantages in model fitting under some fairly designed scenarios, with comparison to classic models. We further applied our proposed methods in a study on birth defects in Massachusetts in Chapter 3 and a study on residents' serum PFOA concentration in Lubeck, WV, and Little Hocking, OH region.

Chapter 1

Introduction

1.1 Disease mapping with individual-level point-referenced data in epidemiology studies

In epidemiology studies, geospatial disparities of certain disease risks are of common interest since the potential unequal risks over an specific geographic area could potentially be a result of location-related risk factors. These factors could be environmental, demographic, socioeconomic among others. In plain language, when investigating a specific disease, epidemiologists frequently aim to identify areas where residents are more likely to develop the disease. Based on the identified areas with high risk, it would be more probable to investigate the underlying risk factors that are associated with occurrence rate of the disease by exploring the difference in potentially relevant factors between high and low risk areas. Once one or more factors are identified, corresponding actions, such as environmental treatment or policy modification, would be possible. For instance, Bristow et al. (2015) conducted a spatial analysis on advanced-stage ovarian cancer mortality in California and found significant geospatial disparity in mortality rate, based on which they managed to identify whether a

patient received NCCN guideline adherent care and treatment at an HVH as the hidden risk factor. The study would therefore help reducing the advanced-stage ovarian cancer mortality by proposing corresponding guidelines or suggestions to certain medical centers.

Partially due to the lack of high resolution data collection and insufficient computing power, traditional disease mapping commonly concentrate on areal data where a specific area, such as a country, a state, or a county as one unit hence the inference are made on the whole areas rather than specific individuals or certain spots within the areas. This class of studies made meaningful inference but the resolution would not be sufficiently satisfying when data are collected on each individual or measurements include accurate geographic information (e.g. longitude and latitude in many cases).

Datasets are called individual if observations within the datasets contain information on specific individuals. In spatial analysis, point-referenced data, or point-level data indicate datasets where items are observed at precise spots on a map. Along with the development of data storage and measurement techniques, these data are becoming increasingly prevalent. My dissertation work therefore focused on methodology in spatial analysis for individual-level point-referenced data. Since these data provide information on unique individuals and locations, inference with higher resolution would be possible, compared to the classic methods that are designed for areal data. In particular, with individual-level point-referenced data, spatial epidemiologist, along with statisticians, naturally expect efficient usage of data where inference could be drawn at virtually all locations on the map of interest rather than one marginal inference over one whole area.

As such, statistical tools for individual-level point-referenced data would be in high demand. In general, nonparametric smoothing techniques, including both freqentist and Bayesian approaches, are used to estimate the underlying spatial risk pattern in order to render inference on a certain type of disease risk at virtually every single location. Further, generalized additive models (GAMs) (Hastie and Tibshirani, 1990) with bivariate smoothers become

increasingly popular when both geospatial and confounding effects exist and the response is assumed to follow an exponential distribution. Examples of these spatial epidemiology studies could be found in Vieira et al. (2009) and Bristow et al. (2015).

1.2 Motivating examples

1.2.1 Birth defects study in Massachusetts

A fairly recent study of birth defects in the state of Massachusetts was conducted by Girguis et al. (2016). In the study, all recorded births in the Massachusetts Birth Defects Registry (MBDR) having cardiac, orofacial and neural tube defects from 2001 to 2009 were identified as cases and 1000 live births per year without defects were sampled as common controls. Among the recorded defects, one of the most common was patent ductus arteriosus (PDA). PDA is a cardiovascular birth defect in which abnormal blood flow occurs between two of the major arteries connected to the heart and is associated with high morbidity and mortality. Residential longitude and latitude were recorded for all observations as well as potential confounding variables including maternal age, adequacy of prenatal care, maternal race, maternal education level and number of siblings.

A primary goal of the MBDR study is to quantify geospatial risks for PDA with adjustment for known risk factors, thus allowing epidemiologists to further explore the underlying space-related risk factors. Moreover, since data are collected over 9 years and the spatial risk pattern could possibly change over the years, statistical tools to estimate time-specific spatial risk pattern are in need, as well as a class of hypothesis tests that formally decide if the spatial risk patterns at each time significantly differ from each other.

1.2.2 Serum PFOA concentration study

Another recent spatial epidemiology study was conducted by Bartell et al. (2010) to investigate serum perfluorooctanoic acid (PFOA) concentration among residents in Lubeck, West Virginia and Little Hocking, Ohio. In this study, researchers aimed to understand the declining behavior of PFOA concentration after granular activated carbon filtration on the public water systems in 2007. By design, 200 residents were included and 6 blood samples were to collect from each resident from May 2007 to August 2008 so that a trend of PFOA concentration could be observed. Besides PFOA concentration, residents' information such as gender, age and recent water consumption type (public or bottled water) was recorded as well as precise residential location (recorded as longitude and latitude).

One of the objectives is to understand the geospatial distribution of residents' serum PFOA concentration in order to help identify potential latent space-confounded risk factors. However, the since this study is a longitudinal one where individuals get repeated measurements, the estimation of spatial effects should be achieved with adjustment of confounding variables as well as the within individual correlation.

1.3 Overview of this dissertation

In this Chapter, we introduce the background and motivating examples for my dissertation work. In Chapter 2, we present statistical methodology background based on which our approaches are developed. The covered statistical background include frequentist and Bayesian smoothing techniques, generalized additive models (GAMs) and generalized linear mixed models. In Chapter 3, we propose stratified smoothers and incorporate these smoothers into GAMs and further developed a class of permuted mean squared difference (PMSD) tests to detect temporal heterogeneity of geospatial effects, with application on

birth defects study in Massachusetts. In Chapter 4, we generalize kernel smoothers with variance-covariance adjustment, describe additive mixed models (AMMs) framework with kernel smoothers and further propose a novel backfitting algorithm to fit AMMs making use of our generalized kernel smoothers. Chapter 5 could be considered as an extension of Chapter 4, accommodating exponential family response by combining penalized quasi-likelihood (PQL) procedure and the fitting procedure in Chapter 4. Either of Chapter 4 and 5 includes an application of the proposed methods on the serum PFOA study and manages to identify high and low risk areas in Lubeck, WV and Little Hocking, OH area. Chapter 6 covers relevant discussion and some insights on probable future avenue of research.

Bibliography

Hirotogu Akaike. Information theory and an extension of the maximum likelihood principle. In *Selected papers of hirotugu akaike*, pages 199–213. Springer, 1998.

Lu Bai, Scott Bartell, Robin Bliss, and Veronica Vieira. Mapgam-package: Mapping smoothed effect estimates from individual-level... 2019.

Sudipto Banerjee, Bradley P Carlin, and Alan E Gelfand. *Hierarchical modeling and analysis for spatial data*. CRC press, 2014.

Scott M Bartell, Antonia M Calafat, Christopher Lyu, Kayoko Kato, P Barry Ryan, and Kyle Steenland. Rate of decline in serum pfoa concentrations after granular activated carbon filtration at two public water systems in ohio and west virginia. *Environmental health perspectives*, 118(2):222–228, 2010.

Leo Breiman and Jerome H Friedman. Estimating optimal transformations for multiple regression and correlation. *Journal of the American statistical Association*, 80(391):580–598, 1985.

Norman E Breslow and David G Clayton. Approximate inference in generalized linear mixed models. *Journal of the American statistical Association*, 88(421):9–25, 1993.

Robert E Bristow, Jenny Chang, Argyrios Ziogas, Daniel L Gillen, Lu Bai, and Veronica M Vieira. Spatial analysis of advanced-stage ovarian cancer mortality in california. *American journal of obstetrics and gynecology*, 213(1):43–e1, 2015.

Chris Brunsdon, A Stewart Fotheringham, and Martin E Charlton. Geographically weighted regression: a method for exploring spatial nonstationarity. *Geographical analysis*, 28(4): 281–298, 1996.

William S Cleveland. Robust locally weighted regression and smoothing scatterplots. *Journal of the American statistical association*, 74(368):829–836, 1979.

Noel Cressie. The origins of kriging. *Mathematical geology*, 22(3):239–252, 1990.

Noel Cressie. Statistics for spatial data. *Terra Nova*, 4(5):613–617, 1992.

Peter J Diggle and Paulo J Ribeiro Jr. Bayesian inference in gaussian model-based geostatistics. *Geographical and Environmental Modelling*, 6(2):129–146, 2002.

Peter J Diggle, Paulo J Ribeiro, and Ole F Christensen. An introduction to model-based geostatistics. In *Spatial statistics and computational methods*, pages 43–86. Springer, 2003.

Jean Duchon. Splines minimizing rotation-invariant semi-norms in sobolev spaces. *Constructive theory of functions of several variables*, pages 85–100, 1977.

Andrew O. Finley, Sudipto Banerjee, and Bradley P. Carlin. spBayes: An R package for univariate and multivariate hierarchical point-referenced spatial models. *Journal of Statistical Software*, 19(4):1–24, 2007. URL `http://www.jstatsoft.org/v19/i04/`.

Andrew O. Finley, Sudipto Banerjee, and Alan E.Gelfand. spBayes for large univariate and multivariate point-referenced spatio-temporal data models. *Journal of Statistical Software*, 63(13):1–28, 2015. URL `http://www.jstatsoft.org/v63/i13/`.

Alan E Gelfand and Sudipto Banerjee. Bayesian modeling and analysis of geostatistical data. *Annual Review of Statistics and Its Application*, 4:245–266, 2017.

Mariam S Girguis, Matthew J Strickland, Xuefei Hu, Yang Liu, Scott M Bartell, and Verónica M Vieira. Maternal exposure to traffic-related air pollution and birth defects in massachusetts. *Environmental research*, 146:1–9, 2016.

Gene H Golub, Michael Heath, and Grace Wahba. Generalized cross-validation as a method for choosing a good ridge parameter. *Technometrics*, 21(2):215–223, 1979.

Mark S Handcock and Michael L Stein. A bayesian analysis of kriging. *Technometrics*, 35 (4):403–410, 1993.

Trevor Hastie and Robert Tibshirani. *Generalized additive models*. Wiley Online Library, 1990.

Robert E Kass and Adrian E Raftery. Bayes factors. *Journal of the american statistical association*, 90(430):773–795, 1995.

Nan M Laird and James H Ware. Random-effects models for longitudinal data. *Biometrics*, pages 963–974, 1982.

Kung-Yee Liang and Scott L Zeger. Longitudinal data analysis using generalized linear models. *Biometrika*, 73(1):13–22, 1986.

Xihong Lin and Daowen Zhang. Inference in generalized additive mixed modelsby using smoothing splines. *Journal of the royal statistical society: Series b (statistical methodology)*, 61(2):381–400, 1999.

Peter McCullagh. *Generalized linear models*. Routledge, 2018.

Henning Omre. Bayesian kriging—merging observations and qualified guesses in kriging. *Mathematical Geology*, 19(1):25–39, 1987.

Henning Omre and Kjetil B Halvorsen. The bayesian bridge between simple and universal kriging. *Mathematical Geology*, 21(7):767–786, 1989.

Michael L Stein. *Interpolation of spatial data: some theory for kriging*. Springer Science & Business Media, 2012.

Yannan Tang, Verónica M Vieira, Scott M Bartell, and Daniel L Gillen. Additive mixed models with kernel smoothers for disease mapping using individual-level data. 2020.

Florin Vaida and Suzette Blanchard. Conditional akaike information for mixed-effects models. *Biometrika*, 92(2):351–370, 2005.

Verónica Vieira, Thomas Webster, Janice Weinberg, and Ann Aschengrau. Spatial analysis of bladder, kidney, and pancreatic cancer on upper cape cod: an application of generalized additive models to case-control data. *Environmental Health*, 8(1):3, 2009.

Grace Wahba. Spline bases, regularization, and generalized cross-validation for solving approximation problems with large quantities of noisy data. *Approximation theory III*, 2, 1980.

Thomas Webster, Verónica Vieira, Janice Weinberg, and Ann Aschengrau. Method for mapping population-based case-control studies: an application using generalized additive models. *International Journal of Health Geographics*, 5(1):26, 2006.

Russ Wolfinger and Michael O'connell. Generalized linear mixed models a pseudo-likelihood approach. *Journal of statistical Computation and Simulation*, 48(3-4):233–243, 1993.

Simon N Wood. Thin plate regression splines. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 65(1):95–114, 2003.

Simon N Wood. Stable and efficient multiple smoothing parameter estimation for generalized additive models. *Journal of the American Statistical Association*, 99(467):673–686, 2004.

Simon N Wood. Fast stable restricted maximum likelihood and marginal likelihood estimation of semiparametric generalized linear models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 73(1):3–36, 2011.

Simon N Wood. *Generalized additive models: an introduction with R*. CRC press, 2017.

Simon N Wood, Mark V Bravington, and Sharon L Hedley. Soap film smoothing. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 70(5):931–955, 2008.

Simon N Wood, Natalya Pya, and Benjamin Säfken. Smoothing parameter and model selection for general smooth models. *Journal of the American Statistical Association*, 111 (516):1548–1563, 2016.

Robin L Young, Janice Weinberg, Verónica Vieira, Al Ozonoff, and Thomas F Webster. Generalized additive models and inflated type i error rates of smoother significance tests. *Computational statistics & data analysis*, 55(1):366–374, 2011.