Bias correction of bounded location errors in binary data

by

Nelson B. Walker

B. S., Brigham Young University, 2014

A REPORT

submitted in partial fulfillment of the
requirements for the degree

MASTER OF SCIENCE

Department of Statistics
College of Arts and Sciences

KANSAS STATE UNIVERSITY
Manhattan, Kansas

2018

Approved by:

Major Professor
Dr. Trevor Hefley

Copyright

© Nelson B. Walker 2018.

Abstract

Binary regression models for spatial data are commonly used in disciplines such as epidemiology and ecology. Many spatially-referenced binary data sets suffer from location error, which occurs when the recorded location of an observation differs from its true location. When location error occurs, values of the covariates associated with the true spatial locations of the observations cannot be obtained. We show how a change of support (COS) can be applied to regression models for binary data to provide bias-corrected coefficient estimates when the true values of the covariates are unavailable, but the unknown location of the observations are contained within non-overlapping polygons of any geometry. The COS accommodates spatial and non-spatial covariates and preserves the convenient interpretation of methods such as logistic and probit regression. Using a simulation experiment, we compare binary regression models with a COS to naive approaches that ignore location error. We illustrate the flexibility of the COS by modeling individual-level disease risk in a population using a binary data set where the location of the observations are unknown, but contained within administrative units. Our simulation experiment and data illustration corroborate that conventional regression models for binary data which ignore location error are unreliable, but that the COS can be used to eliminate bias while preserving model choice.

Table of Contents

List of Figures

List of Tables

Acknowledgments

I would like to acknowledge the fine and diligent mentorship of my adviser, Dr. Trevor Hefley. Without his guidance and support, this work would not be possible. I likewise acknowledge the support and input of my masters committee, Dr. Perla Reyes-Cuellar and Dr. Weixing Song. Without the support of my wife, Kelsee Walker, this degree itself would not be possible, and I extend my appreciation and love to her. I acknowledge the collaboration and early input from Dr. Daniel Walsh (National Wildlife Health Center) and thank the Wisconsin Department of Natural Resources for their efforts in monitoring the spread of chronic wasting disease and providing the data that was used in this research. Lastly, I acknowledge and thank the Department of Statistics and Dr. Lolafaye Coyne for their financial support through the Lolafaye Coyne Graduate Research Scholarship, which made this research possible during the summer months.

Bias Correction of Bounded Location Errors in Binary Data

1 Introduction

Location error, at some scale, is ubiquitous in all spatial data sets. Whenever the recorded location of an observation is different from the true location, location error occurs. For example, to protect the identity of human subjects in epidemiological studies the location of an observation is often reported as the centroid of a geopolitical unit (Goldberg and Cockburn, 2012). Correcting for location error is important in regression models because the covariates influencing the expected value of the response at the true location may be different than the value of the covariates at the recorded location, thus causing covariate measurement error (see **Figure 1** for an example). Both location error and covariate measurement error cause bias in regression coefficient estimates, which can result in incorrect inference (Carroll et al., 2006). In some cases, the bias in regression coefficient estimates caused by covariate measurement error can reverse the sign of regression coefficient estimates, thereby reversing the interpretation of the coefficient estimates (Carroll et al., 2006; Hefley et al., 2017a). Correction of bias caused by location error is, therefore, required to draw reliable statistical inference from regression models.

Spatially-referenced binary data are commonly used in epidemiology, public health, ecology, and the environmental sciences, and at least some of these data sets will include observations with location error. Aside from privacy concerns, many public health data sets aggregate and report the number of cases for a disease within administrative units. Location error generated by aggregation is bounded because the exact locations are contained within discrete spatial boundaries, such as a county, district, province, or other administrative

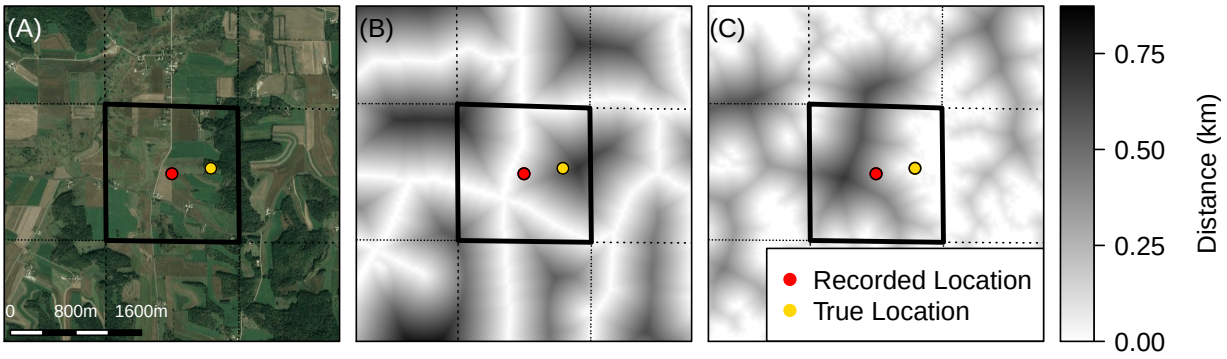


Figure 1: (A) The motivating dataset shows an observation with location error on a map of a 2.59 km^2 section of land in Wisconsin, USA. The recorded location of the observation is given as the centroid and the true location is shown near the right edge of the land section. The covariate distance to nearest development (B; roads, buildings, etc.), and distance to nearest forest (C) is overlaid on the same area. Binary regression using covariate values at the centroid, rather than the true location, would cause bias in the regression coefficient estimates.

unit. When modeling disease risk, researchers commonly incorporate observation-specific and location-specific covariates. For our purposes, we define observation-specific covariates as those which carry an attribute inherent to the observation or subject itself, such as sex, age, species, etc. We define location-specific covariates as those which are spatially-referenced or anchored to a spatial location, inherent to the location of the observation or subject, such as temperature, radon concentration, etc.

Research has focused primarily on reducing bias caused by location error in models for kriging (Cressie and Kornak, 2003; Fanshawe and Diggle, 2011; Orton et al., 2017; Wang et al., 2017) and point processes (Bradley et al., 2016; Chakraborty and Gelfand, 2010; Collins et al., 2017; Hefley et al., 2014, 2017a; Lund and Rudemo, 2000; Park and Davis, 2017; Sadahiro, 2003). Most established methods that correct for location error require some form of auxiliary data, truthing data, or a calibration data set, in the form of exact locations for a subset of observations. However, comparatively few methods exist for binary data.

The literature developing methods that obviate location error for binary data is sparse in describing how regression models could incorporate both observation-specific and location-

covariates. For instance, Zimmerman and Fang (2012) developed location error bias correction for a non-parametric kernel smoothing approach that uses binary data to model and map disease risk. However, no attempt was made to illustrate how location-specific covariates could be used, and it was not readily apparent how observation-specific covariates could be incorporated in the approach. Huque et al. (2016) developed a penalized least squares regression method by assuming smoothness in location-specific covariates that are measured with error. However, extensions for binary regression models, non-smooth spatial covariates, and location error in observations were not presented. Wang et al. (2017) presented a Bayesian hierarchical modeling approach for correcting bias due to bounded location error in binary data, by integrating a spatially-referenced incidence function over discrete sub-regions. Wang et al. (2017) however did not propose extensions that could incorporate observation-specific covariates, which is needed to preserve the interpretability of commonly used models like logistic or probit regression.

Many of the references to bias correction for location error are focused on developments for point process models, including the Poisson inhomogeneous point process (Poisson IPP). Point process models, like the Poisson IPP, describe the random locations of points in a study area based on location-specific covariates. A connection exists between the univariate Poisson point process and the Poisson distribution via a transformation of the Poisson IPP called a change of support (COS; Cressie and Wikle, 2011). If the study area is partitioned into disjoint polygons of arbitrary geometries, the number of points found within each polygon may be modeled as a Poisson random variable. Applying the COS requires that the locations of the points be contained within non-overlapping polygons of any geometry, which is equivalent to the assumption of bounded location error. Previous research shows that bias caused by location error can be corrected using the COS for a point pattern if the data arises from a point process. For example, Bradley et al. (2016) used a COS to make correct for bias due to aggregation over administrative units and Hefley et al. (2017a) used a COS to correct bias due to bounded location error.

Diggle and Rowlingson (1994) presented a model for spatial binary data by representing two Poisson point patterns as a bivariate Poisson IPP. The two point patterns can be linked

using a binary regression model, which accounts for the probability of each point taking a binary mark of zero or one. Linking the two point patterns with a binary regression model is useful because incorporating, for example, a logit or probit link preserves the interpretations of models commonly used by practitioners. Just as the COS may be used to correct bias caused by location error for the univariate Poisson IPP, we show that a COS may be used to correct bias with the bivariate Poisson IPP.

The remainder of this report proceeds as follows: In section 2, we review the Poisson IPP and the COS technique which can be used to correct bias caused by location error. In section 3, we explain the relationship between spatially-referenced binary data and the bivariate IPP. Next, in section 4, we present the COS for the bivariate IPP and, in section 5, we introduce a partial COS that can be used when only a portion of the observations are affected by location error. In section 6, we show how the partial COS can incorporate location-specific covariates and observation-specific covariates, thereby enjoying the same flexibility and convenient interpretation as binary regression with an arbitrary link. In section 7, we evaluate and compare the bivariate IPP with COS to naive logistic regression using a simulation study. Finally, we illustrate our technique by correcting location error in a disease risk factor analysis that demonstrates the partial COS in section 8, interpret the results of both the simulation study and the risk factor analysis in section 9, and discuss our results from the simulation study and risk factor analysis in section 10.

2 The Poisson IPP and Change of Support

2.1 Poisson IPP

The true locations of a random number of observations, n, in a study area $\mathcal{S} \subset \mathbb{R}^2$ can be modeled using a Poisson IPP. The Poisson IPP is defined by $\lambda(\mathbf{s})$, which is a spatially varying intensity function where $\mathbf{s} \equiv (s_1, s_2)$ and $\mathbf{s} \subseteq \mathcal{S}$. The intensity function is often modeled as

$$\log(\lambda(\mathbf{s})) = \beta_0 + \mathbf{x}(\mathbf{s})'\boldsymbol{\beta}\,, \tag{1}$$

where β_0 is the intercept, $\boldsymbol{\beta} \equiv (\beta_1, ..., \beta_p)'$ is a vector of regression coefficients, and $\mathbf{x}(\mathbf{s})$ is a $p \times 1$ vector that contains location-specific covariates at the point $\mathbf{s}$. The probability density function (PDF) of the Poisson IPP is constructed from a Poisson probability mass function (PMF) and a multivariate location density, as follows:

$$f(\mathbf{U}|\lambda) = \frac{e^{-\bar{\lambda}}\bar{\lambda}^n}{n!} \prod_{i=1}^{n} \frac{\lambda(\mathbf{u}_i)}{\int_{\mathcal{S}} \lambda(\mathbf{s}) d\mathbf{s}}\,, \tag{2}$$

where $\mathbf{U}$ is an $n \times 2$ matrix with rows that contain the coordinates of the n points, $\mathbf{u}_i$ is the i^{th} row of $\mathbf{U}$, and $\bar{\lambda} = \int_{\mathcal{S}} \lambda(\mathbf{s}) d\mathbf{s}$ is the integrated intensity function. The log likelihood is:

$$\ell(\lambda|\mathbf{U}) = \sum_{i=1}^{n} \log(\lambda(\mathbf{u}_i)) - \int_{\mathcal{S}} \lambda(\mathbf{s}) d\mathbf{s} - \log(n!)\,. \tag{3}$$

2.2 Change of Support

When a study area is divided into non-overlapping sub-regions, a property of the Poisson IPP is the number of points in each sub-region is a Poisson random variable. The relationship between location-specific covariates and the expected value of Poisson random variables may be estimated using Poisson regression, where the rate parameter is the intensity function integrated over each sub-region. This change, from modeling observations using an IPP with a continuous spatial support, to modeling observations as a Poisson random variable with discrete spatial support, is known as a change of support (COS; Cressie and Wikle, 2011).

To implement a COS, the study area $\mathcal{S}$ is partitioned into m non-overlapping sub-regions $(\mathcal{A}_1, \mathcal{A}_2, ..., \mathcal{A}_m)$, dictated by how the data were reported, such that $\mathcal{S} = \sum_{j=1}^{m} \cup \mathcal{A}_j$. We define the vector $\mathbf{a} \equiv (a_1, a_2, ..., a_m)'$, where a_j is the number of points contained within the j^{th} sub-region $\mathcal{A}_j$. The expected number of points in the j^{th} sub-region is given by the integrated intensity function for the j^{th} sub-region, $\bar{\lambda}_{\mathcal{A}_j} = \int_{\mathcal{A}_j} \lambda(\mathbf{s}) d\mathbf{s}$. The distribution of the random vector $\mathbf{a}$ is

$$\mathbf{a} \sim \text{Pois}(\bar{\boldsymbol{\lambda}}_{\mathcal{A}})\,, \tag{4}$$

where $\bar{\boldsymbol{\lambda}}_{\mathcal{A}} \equiv (\bar{\lambda}_{\mathcal{A}_1}, \bar{\lambda}_{\mathcal{A}_2}, ..., \bar{\lambda}_{\mathcal{A}_m})'$ is an $m \times 1$ vector of integrated intensity functions, dependent upon β_0 and $\boldsymbol{\beta}$ in (1). Because of the relationship in (4), β_0 and $\boldsymbol{\beta}$, from (1) can be estimated when the true locations of observations are unknown, but are contained within known sub-regions.

3 Binary Data and Bivariate Point Process

Regression models for binary data, such as logistic and probit regression, are often used to model spatial binary data because of their convenient interpretation (Cox and Snell, 1989; Cressie, 1993). Here, we use a connection between binary regression and the bivariate point process identified by Diggle and Rowlingson (1994). With this connection, the bivariate point process model maintains the interpretability of binary regression models with an arbitrary link, such as the logit or probit. However, in the presence of location error, using a bivariate point process for binary data can lead to unreliable inference. In section 4 we discuss the COS technique for bivariate point processes that corrects bias caused by bounded location error.

A bivariate Poisson IPP is a PDF which combines two Poisson IPPs, for n_1 observations with a mark of one, and for n_0 observations with a mark of zero. The PDF of the bivariate Poisson IPP may be thought of as simply multiplying (2) by a binomial coefficient that accounts for all possible orders of the marks, and a Bernoulli PMF for each point:

$$f(\mathbf{U}, \mathbf{y}|\lambda, p)) = \binom{n_1+n_0}{n_1} \frac{e^{-\bar{\lambda}} \bar{\lambda}^{n_1+n_0}}{(n_1+n_0)!} \prod_{i=1}^{n_1+n_0} \frac{\lambda(\mathbf{u}_i)}{\int_{\mathcal{S}} \lambda(\mathbf{s}) d\mathbf{s}} \; p(\mathbf{u}_i)^{y_i} (1 - p(\mathbf{u}_i))^{1-y_i} \;, \tag{5}$$

where $\mathbf{U}$ is now an ordered $(n_1+n_0) \times 2$ matrix, and $p(\mathbf{u}_i)$ is the probability of a point having a mark of one or zero at location $\mathbf{u}_i$. We define $\mathbf{y}$ as the $(n_1 + n_0)$ length vector containing the mark of one or zero for each point. For convenience, the first n_1 entries in both $\mathbf{U}$ and $\mathbf{y}$ correspond to points with a mark of one. Thus the PDF for the bivariate Poisson IPP can be expressed as

$$f(\mathbf{U},\mathbf{y}|\lambda,p) = \frac{e^{-\bar{\lambda}_1}\bar{\lambda}_1^{n_1}}{n_1!}\prod_{i=1}^{n_1}\frac{\lambda(\mathbf{u}_i)p(\mathbf{u}_i)}{\int_{\mathcal{S}}\lambda(\mathbf{s})p(\mathbf{s})d\mathbf{s}} \times \frac{e^{-\bar{\lambda}_0}\bar{\lambda}_0^{n_0}}{n_0!}\prod_{i=n_1+1}^{n_1+n_0}\frac{\lambda(\mathbf{u}_i)(1-p(\mathbf{u}_i))}{\int_{\mathcal{S}}\lambda(\mathbf{s})(1-p(\mathbf{s}))d\mathbf{s}} \ . \tag{6}$$

Here, $p(\mathbf{s})$ is a constrained (thinning) function such that $0 < p(\mathbf{s}) < 1$. Thus, $p(\mathbf{s})$ represents the probability of an observation having a mark of one, and $(1 - p(\mathbf{s}))$ represents the probability of an observation having a mark of zero at location $\mathbf{s}$. Additionally, $\bar{\lambda}_1 = \int_{\mathcal{S}} \lambda(\mathbf{s})p(\mathbf{s})\,d\mathbf{s}$ is the integrated intensity that gives the expected number of observations with a mark of one. Likewise, $\bar{\lambda}_0 = \int_{\mathcal{S}} \lambda(\mathbf{s})(1 - p(\mathbf{s}))d\mathbf{s}$ is the integrated intensity that gives the expected number of observations with a mark of zero. Just as with $p(\mathbf{u}_i)$ in (5), $p(\mathbf{s})$ can be linked to covariates using an arbitrary link function. As an example, we use a logit function:

$$\text{logit}(p(\mathbf{s})) = \alpha_0 + \mathbf{z}(\mathbf{s})'\boldsymbol{\alpha}\,, \tag{7}$$

where α_0 is an intercept, $\boldsymbol{\alpha} \equiv (\alpha_1, ..., \alpha_q)'$ is a $q \times 1$ vector of regression coefficients, and $\mathbf{z}(\mathbf{s})$ is a $q \times 1$ vector that contains covariates. The log-likelihood of the bivariate IPP is:

$$\ell(\lambda, p|\mathbf{U}) = \begin{array}{l} \sum_{i=1}^{n_1}\log(\lambda(\mathbf{u}_i)p(\mathbf{u}_i)) - \int_S \lambda(\mathbf{s})p(\mathbf{s})\,d\mathbf{s} - \log(n_1!) \ + \\ \sum_{i=n_1+1}^{n_1+n_0}\log(\lambda(\mathbf{u}_i)(1-p(\mathbf{u}_i)))) - \int_S \lambda(\mathbf{s})(1-p(\mathbf{s}))d\mathbf{s} - \log(n_0!). \end{array} \tag{8}$$

The thinning function $p(\mathbf{s})$ arises naturally from the joint distribution, composed of the bivariate Poisson IPP and the product of Bernoulli PMFs. Thus binary outcomes may be modeled jointly as two point patterns while preserving the interpretation of the binary regression model's arbitrary link. As inference obtained from modeling data as a bivariate Poisson IPP is subject to bias caused by location error, we next present an application of the COS to correct for this issue.

4 Bivariate Change of Support

The COS technique is straight-forward to apply to the bivariate Poisson IPP. As before, the study area $\mathcal{S}$ is partitioned into m non-overlapping sub-regions $\mathcal{A}_j$, such that $\mathcal{S} = \sum_{j=1}^{m} \cup \mathcal{A}_j$.

We define the vectors $\mathbf{a} \equiv (a_1, a_2, ..., a_m)'$ and $\mathbf{b} \equiv (b_1, b_2, ..., b_m)'$ where a_j and b_j are the number of points in the j^{th} sub-region $\mathcal{A}_j$, with a mark of one or zero respectively. We define $\bar{\lambda}_{1\mathcal{A}_j}$ as the expected count of observations with a mark of one in the j^{th} sub-region, and $\bar{\lambda}_{0\mathcal{A}_j}$ as the expected count of observations with a mark of zero in the same j^{th} sub-region. The distributions of $\mathbf{a}$ and $\mathbf{b}$ are

$$\begin{aligned} \mathbf{a} &\sim \text{Pois}(\bar{\boldsymbol{\lambda}}_{1\mathcal{A}}) \\ \mathbf{b} &\sim \text{Pois}(\bar{\boldsymbol{\lambda}}_{0\mathcal{A}}), \end{aligned} \tag{9}$$

where $\bar{\boldsymbol{\lambda}}_{1\mathcal{A}} \equiv (\bar{\lambda}_{1\mathcal{A}_1}, \bar{\lambda}_{1\mathcal{A}_2}, ..., \bar{\lambda}_{1\mathcal{A}_m})'$ is a $m \times 1$ vector with the j^{th} element corresponding to the integrated intensity function $\bar{\lambda}_{1\mathcal{A}_j} = \int_{\mathcal{A}_j} \lambda(\mathbf{s}) p(\mathbf{s})\, d\mathbf{s}$, for the j^{th} sub-region. Similarly, $\bar{\boldsymbol{\lambda}}_{0\mathcal{A}} \equiv (\bar{\lambda}_{0\mathcal{A}_1}, \bar{\lambda}_{0\mathcal{A}_2}, ..., \bar{\lambda}_{0\mathcal{A}_m})'$ is a $m \times 1$ vector, with the j^{th} element corresponding to the integrated intensity function $\bar{\lambda}_{0\mathcal{A}_j} = \int_{\mathcal{A}_j} \lambda(\mathbf{s})(1 - p(\mathbf{s}))d\mathbf{s}$.

5 Partial Change of Support

We have described how the COS is applied to the bivariate Poisson IPP. In this section, we introduce a partial COS technique, which may be used when the true locations are known for just one of the two point patterns in the bivariate Poisson IPP. Consider an example where, in a case-control study, the locations of cases are only known to be within a specific sub-region of the study area. For the individuals who were selected as controls, the privacy constraints are more relaxed, so we know their exact locations. Since location error is only an issue with one of the two point patterns, we can apply a partial COS when estimating regression coefficients associated with the thinning function (7). A partial COS can use both the exact location information from the control observations and bounded location error from case observations to estimate the effect due to the covariates.

To construct the component of the partial COS log-likelihood which corresponds to the cases (observations with a mark of one), we use the log of the Poisson PMF from the bivariate COS in (9). To construct the component of the log-likelihood corresponding to the controls (observations with a mark of zero), we use the portion of (8) associated with the n_0

observations and thinning function $(1 - p(\mathbf{s}))$. The complete log-likelihood is:

$$\ell(\lambda, p|\mathbf{a}, \mathbf{U}) = \begin{array}{l} \sum_{j=1}^{m}(-\bar{\lambda}_{1\mathcal{A}_j} + a_j \log(\bar{\lambda}_{1\mathcal{A}_j}) - \log(a_j!)) \; + \\ \sum_{i=n_1+1}^{n_1+n_0} \log(\lambda(\mathbf{u}_i)(1 - p(\mathbf{u}_i))) - \int_S \lambda(\mathbf{s})(1 - p(\mathbf{s}))d\mathbf{s} - \log(n_0!). \end{array} \tag{10}$$

By extension, this technique could be used when true locations are known for only some of the observations in one of the two point patterns. If we assume that some cases in the study area have exact locations, we assign n_g as the number of these cases. Further, if we assume the n_g exact locations are found in different sub-regions from the cases that have inexact locations, our log-likelihood is:

$$\ell(\lambda, p|\mathbf{U}) = \begin{array}{l} \sum_{j=1}^{m-l}(-\bar{\lambda}_{1\mathcal{A}_j} + a_j \log(\bar{\lambda}_{1\mathcal{A}_j}) - \log(a_j!)) \; + \\ \sum_{i=n_1-n_g}^{n_1} \log(\lambda(\mathbf{u}_i)p(\mathbf{u}_i)) \; - \int_S \lambda(\mathbf{s})p\,(\mathbf{s})d\mathbf{s} - \log((n_1 - n_g)!) + \\ \sum_{i=n_1+1}^{n_1+n_0} \log(\lambda(\mathbf{u}_i)(1 - p(\mathbf{u}_i))) - \int_S \lambda(\mathbf{s})(1 - p(\mathbf{s}))d\mathbf{s} - \log(n_0!). \end{array} \tag{11}$$

Here, the vector $\mathbf{a} \equiv (a_1, a_2, ..., a_{m-l})'$ is the counts of observations (cases) contained in $m-l$ sub-regions. We give l as the number of sub-regions that contain exact locations for all cases. This demonstrates that the COS technique is capable of accommodating binary data with varying levels of location accuracy.

6 A Spatial-only Change of Support with Observation-Specific Covariates

Analyses of binary data often require using location-specific and observation-specific covariates. We now present a modification, which allows estimation of regression coefficients for both location-specific covariates in the presence of location error and observation-specific covariates known without measurement error. The ability to include observation-specific

covariates in the model, such as sex and age, is an important development because it preserves the choice of model by allowing both types of covariates to be included. The bivariate IPP with COS therefore preserves the utility of other standard binary regression models. Additionally, the bivariate IPP model with COS could easily be modified to accommodate observation-specific covariates suffering from traditional measurement error (e.g. Buonaccorsi et al., 2018).

Consider the bivariate Poisson IPP. We have a PDF that describes the locations of a random number of points with a mark of one (n_1), and PDF that does the same for a random number of points with a mark of zero (n_0). The PDFs incorporate a thinning function, $p(\mathbf{s})$, to describe the probability of having a mark of one or zero, dependent upon location-specific covariates. If we include observation-specific covariates in the thinning functions, the joint PDF of the bivariate Poisson IPP depends on the intensity function $\lambda(\mathbf{s})$, and the thinning functions $p(\mathbf{s}, \mathbf{w})$ and $1 - p(\mathbf{s}, \mathbf{w})$. Here, $p(\mathbf{s}, \mathbf{w})$ depends on a $q \times 1$ vector of location-specific covariates $\mathbf{z}(\mathbf{s})$, an $r \times 1$ vector of observation-specific covariates ($\mathbf{w}$; e.g. sex or age), and an $r \times 1$ vector of regression coefficients $\boldsymbol{\gamma} \equiv (\gamma_1, \gamma_2, ..., \gamma_r)'$. Similar to (7), we use a logit link for $p(\mathbf{s}, \mathbf{w})$,

$$\text{logit}(p(\mathbf{s}, \mathbf{w})) = \alpha_0 + \mathbf{z}(\mathbf{s})'\boldsymbol{\alpha} + \mathbf{w}'\boldsymbol{\gamma}. \tag{12}$$

The PDF of this particular bivariate Poisson IPP is:

$$f(\mathbf{U}|\lambda, p) = \begin{array}{l} \frac{e^{-\bar{\lambda}_1}\bar{\lambda}_1^{n_1}}{n_1!} \prod_{i=1}^{n_1} \frac{\lambda(\mathbf{u}_i)p(\mathbf{u}_i,\mathbf{w}_i)}{\int_{\mathcal{S}}\int_{\mathcal{W}}\lambda(\mathbf{s})p(\mathbf{s},\mathbf{w})d\mathbf{w}d\mathbf{s}} \times \\ \frac{e^{-\bar{\lambda}_0}\bar{\lambda}_0^{n_0}}{n_0!} \prod_{i=n_1+1}^{n_1+n_0} \frac{\lambda(\mathbf{u}_i)(1-p(\mathbf{u}_i,\mathbf{w}_i))}{\int_{\mathcal{S}}\int_{\mathcal{W}}\lambda(\mathbf{s})(1-p(\mathbf{s},\mathbf{w})d\mathbf{w}d\mathbf{s}}, \end{array} \tag{13}$$

where $\bar{\lambda}_1 = \int_{\mathcal{S}}\int_{\mathcal{W}}\lambda(\mathbf{s})p(\mathbf{s}, \mathbf{w})d\mathbf{w}d\mathbf{s}$ is the integrated intensity for observations with mark of one, $\bar{\lambda}_0 = \int_{\mathcal{S}}\int_{\mathcal{W}}\lambda(\mathbf{s})(1 - p(\mathbf{s}, \mathbf{w}))d\mathbf{w}d\mathbf{s}$ is the integrated intensity for observations with a mark of zero, $\mathcal{S}$ is the two-dimensional study area, and $\mathcal{W}$ is the support of observation-specific covariates.

Additionally, $\mathbf{w}_i$ is the vector of observation-specific covariates for the i^{th} observation in $\mathbf{U}$. The location PDF of an IPP for any given observation with a mark of one, $\mathbf{u}_i$

(where $\mathbf{u}_i$ is in the first n_1 rows of $\mathbf{U}$), can be constructed by normalizing $\lambda(\mathbf{u}_i)p(\mathbf{u}_i, \mathbf{w}_i)$ with the integrated intensity function $\int_{\mathcal{S}} \int_{\mathcal{W}} \lambda(\mathbf{s})p(\mathbf{s}, \mathbf{w})d\mathbf{w}d\mathbf{s}$. This allows the location PDF to integrate to one over both the study area $\mathcal{S}$ and the sample space of the individual-level covariates $\mathcal{W}$. For our purposes, the support of $\mathcal{W}$ may be approximated empirically by Monte Carlo sampling of the observations. The log-likelihood for this bivariate IPP distribution is:

$$\ell(\lambda, p|\mathbf{U}) = \begin{array}{l} \sum_{i=1}^{n_1} \log(\lambda(\mathbf{u}_i)p(\mathbf{u}_i, \mathbf{w}_i)) \;\; - \;\; \int_{\mathcal{S}} \int_{\mathcal{W}} \lambda(\mathbf{s})p(\mathbf{s}, \mathbf{w})d\mathbf{w}d\mathbf{s} - \log(n_1!) \;\; + \\ \sum_{i=n_1+1}^{n_1+n_0} \log(\lambda(\mathbf{u}_i)p(\mathbf{u}_i, \mathbf{w}_i)) \;\; - \;\; \int_{\mathcal{S}} \int_{\mathcal{W}} \lambda(\mathbf{s})(1 - p(\mathbf{s}, \mathbf{w}))d\mathbf{w}d\mathbf{s} - \log(n_0!)\,. \end{array} \tag{14}$$

If we assume that, for the $n_1 + n_0$ observations, each location was only known to be contained within the respective sub-region $\mathcal{A}_j$, we can apply a spatial-only COS (or partial COS) to the bivariate IPP PDF. To do this, we integrate the intensity function for each observation over the spatial domain of the observation's assigned sub-region. The resulting log-likelihood is:

$$\ell(\lambda, p|\mathbf{U}) = \begin{array}{l} \sum_{i=1}^{n_1} \log(\int_{\mathcal{A}_i} \lambda(\mathbf{s})p(\mathbf{s}, \mathbf{w}_i)d\mathbf{s}) - \int_{\mathcal{S}} \int_{\mathcal{W}} \lambda(\mathbf{s})p(\mathbf{s}, \mathbf{w})d\mathbf{w}d\mathbf{s} \;\; + \\ \sum_{i=n_1+1}^{n_1+n_0} \log(\int_{\mathcal{A}_i} \lambda(\mathbf{s})(1 - p(\mathbf{s}, \mathbf{w}_i))d\mathbf{s}) - \int_{\mathcal{S}} \int_{\mathcal{W}} \lambda(\mathbf{s})(1 - p(\mathbf{s}, \mathbf{w}))d\mathbf{w}d\mathbf{s}. \end{array} \tag{15}$$

7 Simulation Experiment

We conduct a simulation experiment to compare the COS technique with binary regression in three scenarios that may be commonly encountered in practice:

1. Bivariate Poisson IPP with a logistic link function and COS, where the true location is not known;

2. Ordinary logistic regression, where the true location is not known and the model uses the covariate value at the center of each sub-region;

3. Ordinary logistic regression where the true location is known.

Our purpose is to compare estimates of a slope parameter, α_1, among the three scenarios. This parameter, (α_1), is associated with a location-specific covariate $z(\mathbf{s})$ when defining our ordinary logistic regression model as $\text{logit}(p(\mathbf{s})) = \alpha_0 + z(\mathbf{s})\alpha_1$. Similarly, we are interested in α_1 in the bivariate IPP with COS because it influences the probability of an observation having a mark of one. For this simulation experiment, we compare estimates of α_1 between models by evaluating empirical bias, coverage probabilities, and efficiency. Thus, we test the ability of the COS technique to match the interpretability of logistic regression coefficient estimates, while correcting bias due to location error.

For our simulation, we use a unit square study area, $\mathcal{S} = [0,1] \times [0,1]$, that is divided into twenty-five grid cells (sub-regions) such that $\mathcal{S} = \cup_{j=1}^{25} \mathcal{A}_j$ and $|\mathcal{A}_j| = \frac{1}{25}$. We define one location-specific spatially correlated covariate for the study area $\mathcal{S}$, $z(\mathbf{s})$, which is a mixture of a Gaussian process that results from a convolution of an exponential kernel with white noise and near-minimal values at the sub-region centers (Higdon, 2002). We then use $z(\mathbf{s})$ to generate simulated data from a bivariate Poisson IPP model, where $x(\mathbf{s}) \equiv z(\mathbf{s})$. We thus employ the log link function (1) to determine the intensity and location of observations and a Bernoulli PMF to assign marks of one or zero. Under this data model, β_0 and β_1 influence the location of points. Our choice of parameters is $\beta_0 = 5.5$ or 8.0 (small and large sample size), $\beta_1 = 0$, $\alpha_0 = 0$, and $\alpha_1 = 1$. Our choice of α_0 and α_1 yields a relatively equal number of points with each mark.

We simulate 1000 data sets from four different settings using a combination of two factors, each with two levels: sample size (small, large); partial presence of bounded location error (location error present in only the point pattern with marks of one, location error present in both point patterns). For each data set, we estimate the parameter of interest, α_1, among the three scenarios using maximum likelihood estimation. Particularly, under scenario one we use the Nelder-Mead algorithm (Nelder and Mead, 1965) with the likelihood from 10 and 14 in the program R. We use the glm function, also in the program R, for estimating α_1 in scenarios two and three (R Core Team, 2016). For each setting, we calculate and

compare 95% coverage probabilities for α_1 for Wald-type confidence intervals and construct plots comparing the empirical distribution of $\hat{\alpha}_1$ among the methods obtained from the 1000 data sets. We also calculate the mean sample size of points with a mark of one and zero for each setting, and the efficiency of the estimates of α_1 obtained using the COS technique. We assess efficiency by dividing the standard error of the estimates for α_1 obtained using the COS (i.e. scenario one), by the standard error of the estimates for α_1 obtained using ordinary logistic regression, using data without location error (i.e. scenario three). We report the mean of these ratios as our measure of efficiency for each setting.

Under scenario one, we expect the COS technique to yield unbiased estimates for α_1 when data is generated according to a bivariate Poisson IPP model. We expect the 95% confidence intervals obtained from the COS to have an approximately 95% coverage probability. Under scenario two, when location error is present and ignored, we expect estimates of α_1 will be biased when using ordinary logistic regression. We expect the coverage probability for estimates obtained using ordinary logistic regression to be less than 0.95 in the presence of location error. When location error occurs, we expect the 95% confidence intervals obtained using the COS technique (scenario one) to be wider than those produced using ordinary logistic regression (scenario two), due to a bias-variance trade-off.

8 Disease Risk Factor Analysis

The partial, spatial-only COS technique outlined in section 6 is useful for disease risk factor analyses when both location-specific and observation-specific covariates are considered. We give an example from wildlife ecology to demonstrate the impact of location error on disease risk factor analysis for Chronic Wasting Disease (CWD) in Wisconsin, USA. Chronic wasting disease is an invariably fatal transmissible spongiform encephalopathy of cervids. CWD is known to affect wild populations of white-tailed deer (*Odocoileus virginianus*), mule deer (*Odocoileus hemionus*), elk (*Cervus canadensis*), moose (*Alces alces*), and reindeer (*Rangifer tarandus*; Williams and Young, 1980). First discovered in a captive facility in Colorado in 1967, the geographic extent of CWD has expanded such that it has now been documented

in 24 states, as well as Canada, South Korea, and Norway (Carlson et al., 2018).

The CWD endemic region used in our analysis is located in southwestern Wisconsin, and covers approximately 5,970 km^2. The area is comprised largely of agricultural areas interspersed with woodlands, and white-tailed deer are abundant throughout the region. Chronic wasting disease was first discovered in three adult male white-tailed deer in our study area in 2001, and since that time prevalence in harvested deer has been increasing throughout the region based on estimates from surveillance of hunter-harvested deer conducted by the Wisconsin Department of Natural Resources (Hefley et al., 2017b; Heisey et al., 2010). Currently, for some locales within our study area, CWD prevalence rates are reported as exceeding 50% in adult males and 35% in adult females (WIDNAR, 2017). We focus our analysis on data collected in 2012 from hunter-harvested white-tailed deer in the study region. For demonstration purposes we limit our analysis to observations where the recorded location for each harvested deer was reported as the center of a section of land (an approximately 2.59 km^2 square area defined by the public land survey system), thus exhibiting bounded location error (shown in Figure 1). The result is a sample size of 2,497 deer tested for CWD, with 284 positive cases.

We estimate parameters of the bivariate IPP with and without a partial COS using the 2012 CWD data set with the location-specific covariate distance to nearest development included in the intensity function. We used a logit link as the thinning function with the location-specific covariate distance to nearest forest. We made this latter determination because nearness to forest may be an ecologically relevant predictor in the spread of CWD. Additionally, age and sex of deer are known to be important observation-specific covariates when determining the probability of CWD infection, as older male deer have the highest prevalence among demographic groups of harvested deer (Hefley et al., 2017b; Heisey et al., 2010; Walsh and Miller, 2010). As such, we used both age and sex of deer in the thinning function of the bivariate IPP model with a partial COS. We obtain coefficient estimates using maximum likelihood estimation applied to the likelihood from 15, with the Nelder-Mead algorithm (Nelder and Mead, 1965) in the program R. We compare coefficient estimates and Wald-type confidence intervals from the bivariate Poisson IPP with partial COS to coefficient

estimates and Wald-type confidence intervals obtained using a naive logistic regression model, all in the program R (R Core Team, 2016). This logistic regression model incorporates the covariates distance to nearest forest, sex, and age, and is considered naive because it uses values of the location-specific variable at the recorded locations.

9 Results

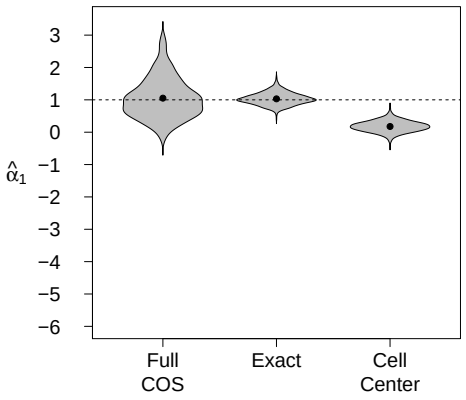
9.1 Simulation Experiment

For our simulation experiment, we crossed two factors (sample size and partial presence of location error), with two levels each. With our choice of β_0 as 5.5 or 8.0, we obtained a mean total sample size of 245.1 for our small sample settings, and 2979.2 for our large sample settings. Sample sizes for points with a mark of one or zero under each setting may be found in **Table 1**.

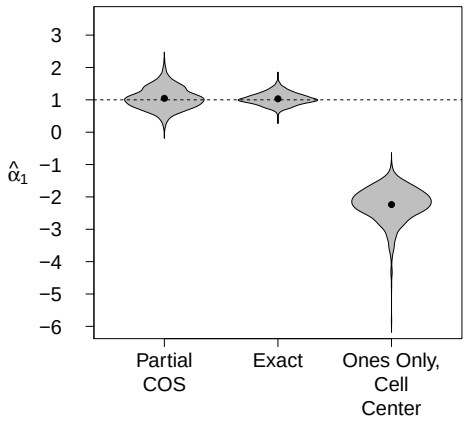
Under scenario one, biases in MLEs for α_1 were corrected successfully by the bivariate COS technique under both settings of location error. The 95% coverage probabilities for $\hat{\alpha}_1$, using the COS technique, were between 94% and 96% in all settings. Additionally, the efficiency of $\hat{\alpha}_1$, obtained from the COS, ranged from 1.76 (large sample with partial change of support) to 3.9 (small sample with full change of support).

Under scenario two, the MLEs for α_1 were biased when values for covariates were used from the center of each sub-region, regardless of which level of location error was present. For settings involving location error only in the point pattern with marks of one, the sign of the estimates for α_1 was reversed for all data sets. Additionally, the empirical distributions of $\hat{\alpha}_1$ obtained through ordinary logistic regression for these settings were dramatically wider than the settings where bounded location error was present in both point patterns. Coverage probabilities for estimates of α_1 using covariates from cell centers were less than 1% for all settings covered by the experiment.

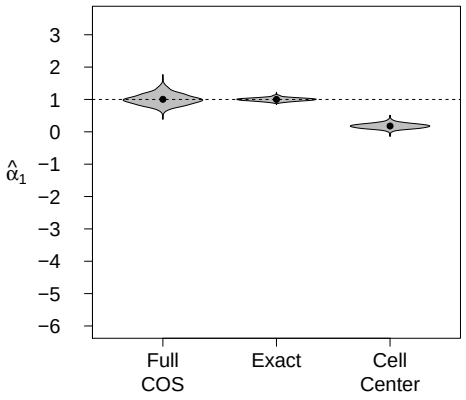
Under scenario three, the MLEs for α_1 performed as expected without location error. Coverage probabilities for estimates of α_1 using covariates from the true locations were



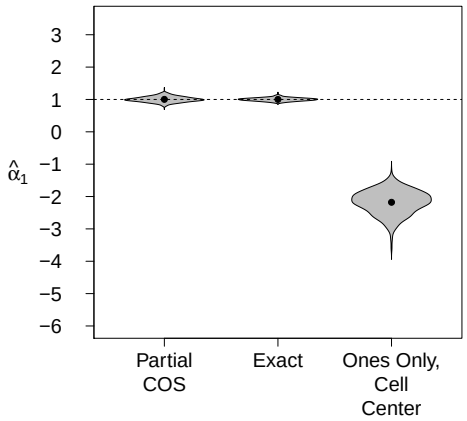
(A) Small Sample, Location Error in Both



(B) Small Sample, Location Error in Points with Mark of One Only



(C) Large Sample, Location Error in Both



(D) Large Sample, Location Error in Points with Mark of One Only

Figure 2: Violin plots showing the empirical distribution of MLEs for the regression coefficient, α_1, from three different scenarios using 1,000 simulated data sets. Each panel shows the true value of $\alpha_1 = 1$ (dotted line). For panels (A) and (C), we use a full COS, and use a partial COS in (B) and (D) because location error is present only in the point pattern with marks of one. We show estimates obtained using logistic regression with the exact point locations (Exact) and using logistic regression with the locations reported as (Cell Center) or (Ones Only, Cell Center) depending on the level of location error. We excluded estimates from (A) when $\mid \hat{\alpha}_1 \mid > 3$ or $\hat{\alpha}_1$ had infinite variance with the COS due to complete separation, which eliminated only 5% of the parameter estimates.

between 94.6% and 94.9% for all settings.

Table 1: Results from the simulation experiment obtained using two sample sizes and two levels of bounded location error (error in both point patterns regardless of mark, error in point pattern with marks of one only). We report the average sample size of points with a mark of one ($\bar{n}_1$) and points with a mark of zero ($\bar{n}_0$) from 1,000 simulated data sets for each setting. We give the relative efficiency of the COS technique in estimating α_1 and the estimated 95% CI coverage probability (CP) for the the COS correction. We also report the estimated 95% CI coverage probabilities for logistic regression using the exact locations of points (exact) and locations reported as the center (cell center).

Presence of Location Error	$\bar{n}_1$	$\bar{n}_0$	Efficiency	CP Bivariate COS	CP Logit (cell center)	CP Logit (exact)
Both	123.5	121.6	3.882	0.948	0.002	0.941
Positives only	123.5	121.6	1.800	0.952	0.000	0.941
Both	1502.2	1477.2	3.430	0.946	0.000	0.944
Positives only	1502.1	1476.9	1.758	0.956	0.000	0.944

9.2 Disease Risk Factor Analysis

We are most interested in the effects of the three models (i.e. the bivariate IPP, ordinary logistic regression, and the bivariate IPP with partial COS) on the ability to detect a difference in the parameters for the distance to nearest forest covariate. A table summarizing the results is found below:

Table 2: Regression coefficient estimates and 95% CIs for effect of distance to nearest forest, sex, and age on the probability of chronic wasting disease infection for deer in the study area. For each covariate and model, we give the coefficient estimate followed by the 95% CI, shown in parentheses.

Method	$\hat{\alpha}_{forest}$	$\hat{\alpha}_{age}$	$\hat{\alpha}_{sex}$
Bivariate IPP	-0.97 (-2.20, 0.26)	0.44 (0.31, 0.57)	0.65 (0.38, 0.94)
Logistic Regression	-0.97 (-2.20, 0.26)	0.44 (0.31, 0.57)	0.65 (0.37, 0.93)
Bivariate Partial COS	-3.66 (-6.21, -1.12)	0.44 (0.31, 0.57)	0.65 (0.37, 0.93)

Our results show that the bivariate Poisson IPP and logistic regression models obtain identical point estimates, to the second decimal place, for the effects of distance to nearest forest, age, and sex on the probability of CWD infection. These two methods give nearly identical 95% confidence intervals for the same three covariates. We see from the 95% confidence intervals for these two methods that distance to nearest forest (forest) is not a significant predictor for the probability of a deer being infected with CWD, while age and sex are significant predictors. The partial COS technique obtains different inference, giving a confidence interval for distance to nearest forest that does not contain zero. This suggests that location error in the data caused attenuation in the estimate of α_{forest}, and an attempt to draw conclusions based on the bivariate IPP or logistic regression results would have been in erroneous.

10 Discussion

Our results suggest the COS technique successfully corrected bias due to location error in the simulation experiment and made an appreciable impact on the inference obtained when modeling probability of CWD infection in individual deer. Our results for the disease risk factor analysis highlight differences in conclusions that would be drawn from the same data using the bivariate Poisson IPP, ordinary logistic regression, and the COS correction technique.

When location-specific covariates are of interest for modeling spatially-referenced binary data, location error in the data amounts to measurement error in the covariates. In turn, measurement error can cause unpredictable changes to conclusions when drawing inference (Carroll et al., 2006). Our simulation experiment highlighted two possible hazards of location error (attenuation and sign reversal), demonstrated the susceptibility of a common method (i.e. logistic regression) to these hazards, and demonstrated the ability of the COS technique to protect against these hazards. We recommend the COS as the technique of choice when the end goal is inference from spatial binary regression in the presence of bounded location error. The COS may be widely applicable because the correction requires no assumptions about

how the location error was generated, apart from being bounded, whereas other methods require truthing data or error distribution assumptions (Chakraborty and Gelfand, 2010; Hefley et al., 2014; Lund and Rudemo, 2000; Sadahiro, 2003). Some situations may render this bounded location error assumption unreasonable. Therefore, future work will address an extension of the COS which allows for unbounded location error correction.

Our simulation study and data example make specific use of a logit link for the binary regression model. The COS technique is easily modified to incorporate other link functions (probit, for example). A spatial random effect may also easily be added to the thinning function to model residual spatial structure in the data (Diggle et al., 1998; Gotway and Stroup, 1997). Additionally, one need not assume a linear relationship in the intensity function as shown in (1); a smooth function such as a semi-parametric or kernel density estimator may be used for the intensity function in the bivariate IPP COS (Diggle, 2013). We refer to these options as modifications rather than extensions because these changes may be made without extending the basic framework given in this paper.

Bibliography

J. R. Bradley, C. K. Wikle, and S. H. Holan. Bayesian spatial change of support for count-valued survey data with application to the american community survey. *Journal of the American Statistical Association*, 111:472–487, 2016.

J. P. Buonaccorsi, G. Romeo, and M. Thoresen. Model based bootstrapping when correcting for measurement error with application to logistic regression. *Biometrics*, 74:35–144, 2018.

Christina M. Carlson, M. Camille Hopkins, Natalie T. Nguyen, Bryan J. Richards, Daniel P. Walsh, and W. David Walter. Chronic wasting disease–status, science, and management support by the U.S. Geological Survey. Technical report, Reston, VA, 2018. URL http://pubs.er.usgs.gov/publication/ofr20171138.

R. J. Carroll, D. Ruppert, L. A. Stefanski, and C. M. Crainiceanu. *Measurement Error in Nonlinear Models: A Modern Perspective*. Chapman & Hall /CRC, 2006.

A. Chakraborty and A. E. Gelfand. Analyzing spatial point patterns subject to measurement error. *Bayesian Analysis*, 5:97–122, 2010.

S. D Collins, J. C. Abbott, and N. E. McIntyre. Quantifying the degree of bias from using county-scale data in species distribution modeling: Can increasing sample size or using county-averaged environmental data reduce distributional overprediction? *Ecology and Evolution*, 7:6012–6022, 2017.

D. R. Cox and E. J. Snell. *Analysis of Binary Data*. Chapman and Hall/CRC, second edition edition, 1989.

N. Cressie. *Statistics for Spatial Data*. John Wiley and Sons, 1993.

N. Cressie and J. Kornak. Spatial statistics in the presence of location error with an application to remote sensing of the environment. *Statistical Science*, 18:436–456, 2003.

N. Cressie and C. Wikle. *Statistics for Spatio-Temporal Data*. Wiley, 2011.

P. J. Diggle. *Statistical Analysis of Spatial and Spatio-Temporal Point Patterns*. CRC Press, 2013.

P. J. Diggle and B. S. Rowlingson. A conditional approach to point process modeling of elevated risk. *Journal of the Royal Statistical Society: Series A*, 153:349–362, 1994.

P.J. Diggle, J. A. Tawn, and R. A. Moyeed. Model based geostatistics. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 47:299–350, 1998.

T. R. Fanshawe and P. J. Diggle. Spatial prediction in the presence of positional error. *Environmetrics*, 22:109–122, 2011.

D. W. Goldberg and M. G. Cockburn. The effect of administrative boundaries and geocoding error on cancer rates in california. *Spatial and Spatio-temporal Epidemiology*, 3:39–54, 2012.

C. A. Gotway and W. W. Stroup. A generalized linear model approach to spatial data analysis and prediction. *Journal of Agricultural, Biological, and Environmental Statistic*, 2:157–178, 1997.

T. J. Hefley, D. M. Baasch, A. J. Tyre, and E. E. Blankenship. Correction of location errors for species distribution models. *Methods in Ecology and Evolution*, 5:207–214, 2014.

T. J. Hefley, B. M. Brost, and M. B. Hooten. Bias correction of bounded location errors in presence-only data. *Methods in Ecology and Evolution*, 8:1566–1573, 2017a.

T. J. Hefley, M. B. Hooten, E. M. Hanks, R. E. Russell, and D. P. Walsh. The bayesian group lasso for confounded spatial data. *Journal of Agricultural, Biological, and Environmental Statistics*, 22:42–59, 2017b.

D. M. Heisey, E. E. Osnas, P. C. Cross, D. O. Joly, J. A. Langenberg, and M. W. Miller. Linking process to pattern: Estimating spatiotemporal dynamics of a wildlife epidemic from cross-sectional data. *Ecological Monographs*, 80:221–240, 2010.

D. Higdon. Space and space-time modeling using process convolutions. In Clive W. Anderson, Vic Barnett, Philip C. Chatwin, and Abdel H. El-Shaarawi, editors, *Quantitative Methods for Current Environmental Issues*, pages 37–56, London, 2002. Springer London.

M. H. Huque, H. D. Bondell, R. J. Carroll, and L. M. Ryan. Spatial regression with covariate measurement error: A semiparametric approach. *Biometrics*, 72:678–686, 2016.

J. Lund and M. Rudemo. Models for point processes observed with noise. *Biometrika*, 87: 235–249, 2000.

J. A. Nelder and R. Mead. A simplex algorithm for function minimization. *Computer Journal*, 7:308–313, 1965.

T. G. Orton, M. R. Dobarco, and N. P. A. Saby. Kriging based on areal summary statistics data: Effects of within-unit variability on predictions and uncertainties. *Spatial Statistics*, 19:42–67, 2017.

D. S. Park and C. C. Davis. Implications and alternatives of assigning climate data to geographical centroids. *Journal of Biogeography*, 44:2188–2198, 2017.

R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2016. URL https://www.R-project.org/.

Y. Sadahiro. Cluster detection in uncertain point distributions: a comparison of four methods. *Computers, Environment and Urban Systems*, 27:33–52, 2003.

D. P. Walsh and M. W. Miller. Designing a weighted surveillance system for chronic wasting disease in colorado. *Journal of Wildlife Diseases*, 46:118–135, 2010.

F. Wang, J. Wang, A. Gelfand, and F. Li. Accommodating the ecological fallacy in disease mapping in the absence of individual exposures. *Statistics in Medicine*, 36:4930–4942, 2017.

E. Williams and S. Young. Chronic wasting disease of captive mule deer: a spongiform encephalopathy. *Journal of Wildlife Diseases*, 16:18–98, 1980.

Wisconsin Department of Natural Resources (WIDNAR). *Prevalence & surveillance*, 2017. URL https://dnr.wi.gov/topic/wildlifehabitat/prevalence.html.

D. L. Zimmerman and X. Fang. Estimating spatial variation in disease risk from locations coarsened by incomplete geocoding. *Statistical Methodology*, 9:239–250, 2012.