Environmental Modelling & Software 82 (2016) 121–127

Contents lists available at ScienceDirect

Environmental Modelling & Software

journal homepage: www.elsevier.com/locate/envsoft

Validating negative binomial lyme disease regression model with bootstrap resampling

Phoebe Tran [a, *], Lam Tran [b]

[a] Department of Epidemiology, Harvard School of Public Health, Boston, MA 02115, United States
[b] Department of Biology, College of Arts and Sciences, University of Pennsylvania, Philadelphia, PA 19104, United States

ARTICLE INFO

Article history:
Received 1 December 2015
Received in revised form
17 April 2016
Accepted 20 April 2016

Keywords:
Lyme disease
Bootstrap resampling
Landscape fragmentation
Bootstrap confidence intervals

ABSTRACT

Various negative binomial regression models have been developed to study Lyme disease in connection to climate and/or landscape factors. However, no internal validation of any of those models has been reported in the literature. This study used bootstrap resampling to conduct an internal validation of a negative binomial regression model on Lyme disease incidence. The model used county-level Lyme disease incidence in thirteen states in the Northeastern United States during 2002–2006 and linked it with several previously identified key landscape and climatic variables used in an earlier study. Results showed that there were significant differences between the outcomes from the initial model and those from bootstrap resampling. Arguably bootstrap resampling, as illustrated in this study, can serve as a sound and valuable means to provide a second line of evidence on model outcomes and shed more insight on variables (e.g., climate and landscape factors) included in the models.

1. Introduction

Since being discovered in the 1970s, Lyme disease has been the most frequently reported vector borne illness in the United States (Bown et al., 2003; Feder et al., 2007). The highest reported incidences of Lyme disease have been observed in the Northeast, the North Central states, and the West Coast where there is a junction of Lyme disease causing ticks (*Ixodes scapularis*), reservoir hosts, sources of blood meal, and suitable climate conditions (Bown et al., 2003; Diuk-Wasser et al., 2006). Studies have shown that infected ticks mainly pass on the disease to humans during the nymph and adult stage of their life (CDC, 2014; Killilea et al., 2008). Furthermore, numerous studies have found the association to varying degrees between different natural factors, such as climatic conditions, forest fragmentation, abundance of acorns, and vector hosts' populations, with Lyme disease cases (e.g., Bouchard et al., 2013; Diuk-Wasser et al., 2012; Finch et al., 2014; Jackson et al., 2006; Killilea et al., 2008; and Schauber et al., 2005).

Due to the nature of Lyme disease incidence data, many researchers have chosen negative binomial models to explore the association between Lyme disease and different natural factors (Bouchard et al., 2013; Diuk-Wasser et al., 2012; Finch et al., 2014). However, while the same type of regression model was used in each study, the results of these negative binomial Lyme disease models (e.g., Bouchard et al., 2013; Diuk-Wasser et al., 2012; Finch et al., 2014; Tran and Waller, 2013, 2015) were very different from one study to another. There are various factors that might contribute to the diverse findings among different studies. Some probable factors include the difference in spatial and/or temporal scales of data and/or the explanatory variables used in those studies (i.e., model specification). Furthermore, from our knowledge no study on internal validation of Lyme disease negative binomial regression models has been reported in the literature, making the comparison of different negative binomial models of Lyme disease incidence more difficult. In that context, this paper reports the use of bootstrap resampling to internally validate a negative binomial regression model to determine the effects of landscape fragmentation and climate variables on Lyme disease incidence in the Northeastern United States. Note that, while bootstrapping has been applied quite extensively in environmental modeling (e.g., Mudelsee and Alkio, 2007; Selle and Hannah, 2010; Srivastav et al., 2014; Hirsch et al., 2015), we could not find any bootstrapping analysis on negative binomial regression in literature, much less those for negative binomial Lyme disease model. Our goal in this study was to explore the consistency between the outcomes of the model with initial dataset (*initial model* hereafter) and those from

* Corresponding author.
E-mail address: pmt585@mail.harvard.edu (P. Tran).

http://dx.doi.org/10.1016/j.envsoft.2016.04.019
1364-8152/

bootstrap resampling (e.g., confidence intervals). Our hypothesis is that, if a variable is statistically significant in the initial model and confirmed by bootstrapping confidence interval (e.g., high percentage of statistically significant runs in the Monte Carlo simulation), the variable is arguably a global factor on Lyme disease incidence in the study area. On the other hand, if a variable is statistically significant in the initial model but not in bootstrapping (e.g., low percentage of statistically significant Monte Carlo runs), such inconsistency can be caused by various reasons (e.g., the variable acts only at local scale (i.e., spatial heterogeneity and/or spatial dependence), and/or the impact of atypical/unusual observations (i.e., outliers) on the model outcomes) and need further analysis (not part of this paper).

2. Materials and methods

2.1. Study area

The study area includes thirteen states in the Northeastern United States: Connecticut, Delaware, Maine, Maryland, Massachusetts, New Hampshire, New Jersey, New York, North Carolina, Pennsylvania, Rhode Island, Vermont, and Virginia (Fig. 1). Data on Lyme disease incidence per county from the contiguous 48 states of the United States from 2002 to 2006 were retrieved from the Geo.Data.gov database (http://geo.data.gov/geoportal/catalog/main/home.page). Land cover data for the year of 2001 were from the National Land Cover Database (NLCD) while temperature and precipitation data were retrieved from the Oregon State Parameter-elevation Regressions on Independent Slope Model (PRISM) group website (http://www.prism.oregonstate.edu/) (Homer et al., 2007; MRLC, 2001).

2.2. Data preparation

FRAGSTAT 4.0 (McGarigal et al., 2012) was used on 2001 NLCD land cover data to derive various landscape indicators at three different levels: patch, class, and landscape. Based on findings from other studies (e.g., Bown et al., 2003; Jackson et al., 2006), eight particular land cover classes (e.g., *developed – open space*, *developed – low intensity*, *developed – medium intensity*, *developed – high intensity*, *deciduous forest*, *evergreen forest*, *mixed forest*, and *grassland & herbaceous*) were included in the study. Table 1 shows the 62 variables used in this study.

2.3. Generalized linear model

Past applications of GLMs for Lyme disease incidence mainly include Poisson regression and its various extensions, such as zero-inflated Poisson regression (Khatchikian et al., 2012), negative binomial regression (Bown et al., 2003; Jackson et al., 2006), and

Fig. 1. Map of the study area.