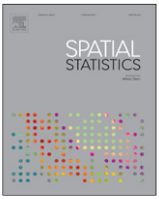
Spatial Statistics 2 (2012) 15–32

Contents lists available at SciVerse ScienceDirect

Spatial Statistics

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



Kernel averaged predictors for spatio-temporal regression models

Matthew J. Heaton [*], Alan E. Gelfand

Department of Statistical Science, Duke University, Box 90251, Durham, NC 27708-0251, United States

ARTICLE INFO

Article history:
Received 12 December 2011
Accepted 22 May 2012
Available online 1 June 2012

Keywords:
Distributed lag
Stochastic integral
Gaussian process
Ozone

ABSTRACT

In regression settings where covariates and responses are observed across space and time, a common goal is to quantify the effect of change in the covariates on the response while adequately accounting for the joint spatio-temporal structure in both. Customary modeling describes the relationship between a covariate and a response variable at a single spatio-temporal location. However, often it is anticipated that the relationship between the response and predictors may extend across space and time. In other words, the response at a given location and time may be affected by levels of predictors in spatio-temporal proximity. Here, a flexible modeling framework is proposed to capture such spatial and temporal lagged effects between a predictor and a response. Specifically, kernel functions are used to weight a spatio-temporal covariate surface in a regression model for the response. The kernels are assumed to be parametric and non-stationary with the data informing the parameter values of the kernel. The methodology is illustrated on simulated data as well as a physical data set of ozone concentrations to be explained by temperature.

1. Introduction

Consider quantifying the effect of a covariate on a response where both the covariate and response are point and time-referenced over a spatio-temporal domain. More concretely, let $Y(\boldsymbol{s}, t)$ denote a response variable at location $\boldsymbol{s}$ and time t and $X(\boldsymbol{s}, t)$ denote a spatio-temporal covariate that is, potentially, associated with the response (unless explicitly stated, $Y(\boldsymbol{s}, t)$ and $X(\boldsymbol{s}, t)$ will be

[*] Correspondence to: National Center for Atmospheric Research, PO Box 3000, Boulder, CO 80307, United States. Tel.: +1 303 497 2884.

E-mail addresses: heaton@ucar.edu, mattheaton@gmail.com (M.J. Heaton), alan@stat.duke.edu (A.E. Gelfand).

2211-6753/$ – see front matter © 2012 Elsevier B.V. All rights reserved.
doi:10.1016/j.spasta.2012.05.001

univariate). To simplify matters, throughout this article assume $(\boldsymbol{s}, t) \in \mathbb{R}^d \times \mathbb{R}$ for some $d \in \mathbb{N}$ but note that if $(\boldsymbol{s}, t) \in \mathcal{D} \times \mathcal{T}$ for bounded domains $\mathcal{D} \subset \mathbb{R}^d$ and $\mathcal{T} \subset \mathbb{R}$ then the methods below still apply with minimal alteration (see Section 2 for more details). Methods for capturing the spatio-temporal correlations within $Y(\boldsymbol{s}, t)$ and $X(\boldsymbol{s}, t)$ are now common with reviews provided by Stein (2005) and Gneiting et al. (2007). When relating $X(\boldsymbol{s}, t)$ to $Y(\boldsymbol{s}, t)$, the common method is to do so linearly through the mean according to

$$\ell(\mathbb{E}(Y(\boldsymbol{s}, t) \mid X(\boldsymbol{s}, t), \beta_0, \beta_1)) = \beta_0 + \beta_1 X(\boldsymbol{s}, t), \tag{1}$$

where $\ell(\cdot)$ is an appropriate link function (e.g. identity, log, etc.). Common extensions of (1) include spatially varying coefficient models (Gelfand et al., 2003) and dynamic spatial process models (Stroud et al., 2001; Huerta et al., 2004; Gelfand et al., 2005). However, a fundamental assumption of (1) is that only $X(\boldsymbol{s}, t)$ affects $\mathbb{E}(Y(\boldsymbol{s}, t))$; neighboring covariate levels $X(\boldsymbol{s}', t')$ for $(\boldsymbol{s}', t')$ close to $(\boldsymbol{s}, t)$ do not. In essence, by Eq. (1) the relationship between Y and X is confined to a single spatial location and time period. However, if $Y(\boldsymbol{s}, t)$ and/or $X(\boldsymbol{s}, t)$ exhibit spatio-temporal correlation then the relationship between them may be more complex. Here, we offer flexible spatio-temporal models to enable this, with a summary of our contribution provided below.

Ground-level ozone is the primary constituent of smog and has been linked to various negative health outcomes associated with the lungs such as chest pain, asthma, and bronchitis (www.epa.gov/ozone). For these reasons, the Environmental Protection Agency (EPA) monitors the levels of ozone near urban areas of the United States. Ozone formation is the result of a chemical reaction between volatile organic compounds (VOC) and nitric oxide (NO_x) in the presence of sunlight (i.e., solar radiation). In the absence of solar radiation data, temperature is often used as a surrogate predictor of the concentration of ozone (Abdul-Wahab et al., 2005; Reich et al., 2011). Specifying a suitable statistical model for the implicit relationship between ozone concentrations and temperature may require more than simply regressing ozone concentrations on temperature at a given location and time period. For example, if temperatures have been high for several days, ozone concentration may also be higher because such conditions, potentially, allow a greater number of reactions between VOC and NO_x to take place. Similarly, in the presence of wind, temperatures at one location in recent days may affect ozone concentrations at a different location on the current day. Finally, as temperature is serving as a surrogate for solar radiation, the relationship between temperature and ozone concentration may be more spatially and temporally complex than had solar radiation been used directly. Each of these possibilities suggests that the effect of temperature on ozone concentration may be spatially and/or temporally *lagged*.

Other examples where spatio-temporal lagged effects occur include the effect of pollution on public health (Schwartz, 2000; Welty and Zeger, 2005; Welty et al., 2009), economic indicators on consumption (Ravines et al., 2006), and disease incidence on disease propagation (Knorr-Held and Richardson, 2003). In all of these examples, the relationship between the response and covariate is not confined to a single spatio-temporal location and lagged effects need to be incorporated into the statistical model.

Models with temporally lagged effects are not new with the most common being the distributed lag model of Almon (1965) and its variations (see Ravines et al., 2006, and the references therein). Distributed lag models in time extend (1) to

$$\ell(\mathbb{E}(Y(\boldsymbol{s}, t) \mid X(\boldsymbol{s}, t), \dots, X(\boldsymbol{s}, t-L), \beta_0, \alpha_0, \dots, \alpha_L)) = \beta_0 + \sum_{l=0}^{L} \alpha_l X(\boldsymbol{s}, t-l), \tag{2}$$

for some known maximum lag L. Alternatively, L could be infinite yielding the Koyck distributed lag model (Koyck, 1954; Frances and van Oest, 2004). Distributed lag models, however, suffer from several limitations. First, if $X(\boldsymbol{s}, t)$ exhibits strong temporal correlation then the set of covariates $\{X(\boldsymbol{s}, t-l) : l = 0, \dots, L\}$ are highly collinear resulting in unstable estimates of the coefficients α_l, $l = 1, \dots, L$. To stabilize the estimates, various constraints are imposed on $\{\alpha_l\}$. For example, $\{\alpha_l\}$ may be assumed to follow some function such as a polynomial (Schwartz, 2000) or spline (Zanobetti et al., 2000). Welty et al. (2009) build constraints into a prior distribution and estimate the coefficients from a Bayesian perspective. Second, (2) only accounts for temporal lags while ignoring spatially