Submitted to the Annals of Applied Statistics
arXiv: math.PR/0000000

BETA REGRESSION FOR TIME SERIES ANALYSIS OF BOUNDED DATA, WITH APPLICATION TO CANADA GOOGLE® FLU TRENDS

By Annamaria Guolo and Cristiano Varin

Università di Verona and Università Ca' Foscari Venezia

Bounded time series consisting of rates or proportions are often encountered in applications. This manuscript proposes a practical approach to analyze bounded time series, through a beta regression model. The method allows the direct interpretation of the regression parameters on the original response scale, while properly accounting for the heteroskedasticity typical of bounded variables. The serial dependence is modeled by a Gaussian copula, with a correlation matrix corresponding to a stationary autoregressive and moving average process. It is shown that inference, prediction, and control can be carried out straightforwardly, with minor modifications to standard analysis of autoregressive and moving average models. The methodology is motivated by an application to the influenza-like-illness incidence estimated by the Google® Flu Trends project.

1. Introduction. Continuous bounded response variables, such as proportions and rates, are frequently encountered in many areas of statistical practice. This kind of data is usually examined through linear regression after a logistic transformation. Despite its feasibility, such a modeling strategy can suffer from some shortcomings, the most relevant being that regression parameters are not directly interpretable on the original response scale, as a consequence of the Jensen's inequality. See Kieschnick and McCullough (2003) and Cribari-Neto and Zeileis (2010) for detailed discussions.

An alternative to linear modeling after logistic transformation consists in a direct analysis of the bounded responses on their original scale. To this purpose, the beta regression model has attracted increasing interest in recent years, as a consequence of the flexibility of the beta distribution in accommodating a variety of distributional shapes over the unit interval. Beta regression modeling of independent observations has been illustrated in Paolino (2001), Ferrari and Cribari-Neto (2004), and Smithson and Verkuilen (2006). Recent applications of beta regression in life sciences have been encountered in clinical medicine (Zou, Carlsson, and Quinn, 2010; Wang et al., 2011),

Keywords and phrases: Beta regression, Bounded time series, Gaussian copula, Google® Flu Trends, Surveillance

neuroscience (Wang, 2012), pharmacometrics (Rogers et al., 2012), and virology (Love et al., 2010).

Recent developments of beta regression analysis of bounded time series have been addressed to observation-driven models (Rocha and Cribari-Neto, 2009; Casarin, Dalla Valle, and Leisen, 2012) and to parameter-driven models (Da-Silva and Migon, 2012). Straightforward likelihood inference makes the observation-driven model appealing. A possible drawback arises in case of regression analysis, since the interpretation of the coefficients depends on past transformed observations in the mean. Parameter-driven models are attractive given their hierarchical construction. Nevertheless, inference and prediction are complicated by the presence of correlated latent variables.

As an alternative to the conditional observation- and parameter-driven models, we suggest a marginal regression approach, through the specification of a convenient class of beta regression models with autoregressive and moving average errors. The serial dependence is modeled by a Gaussian copula. Likelihood inference, prediction, and control are carried out in a straightforward manner, with a computational complexity similar to that of an ordinary ARMA model. In addition, the approach allows an attractive interpretation of model components.

This article is motivated by surveillance of influenza through analysis of influenza-like-illness percentage estimated from aggregated web search queries by the Google® Flu Trends project. Analysis of influenza time series is a key step in disease surveillance for monitoring the progress of epidemics, early identification of pandemics, and ascertainment of factors associated to unexpected changes in flu levels.

The plan of the article is as follows. Section 2 describes the motivating Google® Flu Trends data. Section 3 summarizes beta regression modeling and some extensions for time series analysis. The proposed methodology is detailed in Section 4 and its finite sample performance is investigated through simulation in Section 5. Section 6 describes on-line monitoring of influenza outbreaks through control charts applied to beta regression predictive quantile residuals. The application to the real data set of interest is given in Section 7. Final remarks in Section 8 conclude.

Methods described in the paper are implemented within the more general `R` (R Core Team, 2012) package `gcmr` "Gaussian copula marginal regression", version 0.6.1. The package is freely available at the CRAN repository, URL `http://cran.r-project.org/web/packages/gcmr`.

2. Motivating example. The Google® Flu Trends project aims at early detection of influenza-like-illness (ILI) activity around the world. The

ILI activity is measured in terms of cases per 100 000 persons. The number of cases is reconstructed starting from aggregated Google® search queries related to the disease, such as, for example, *influenza complication*, *flu remedy*, *influenza symptoms*, and *antiviral medication*. See Ginsberg et al. (2009) for details about ILI counts estimation. The Google® estimated ILI time series are publicly available at URL `www.google.org/flutrends`. Data start on the last week of 2002 for Brazil and Peru. Information has been successively extended to other 26 countries all around the world. Researchers at the U.S. Centers for Disease Control and Prevention consider Google® Flu Trends as an early warning of an outbreak, although not a substitute for traditional epidemiological surveillance networks. In fact, recent data from U.S. indicate that peak influenza levels in Winter 2012-2013 have been overestimated, as a consequence of an increased number of search queries related to influenza strains which caused more serious illness and deaths than usual (Butler, 2013).

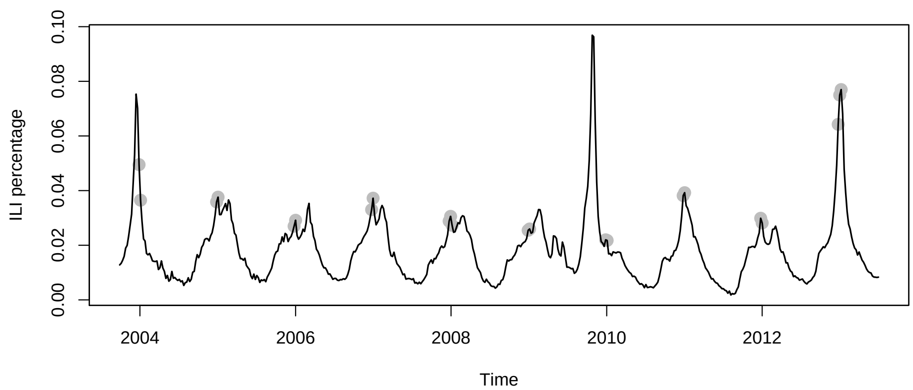

FIG 1. *Google® Flu Trends estimated ILI percentage for Canada. Circles denote Christmas/New Year holidays. Data source:* `www.google.org/flutrends`.

Figure 1 displays the time series of Google® estimated ILI percentage, obtained as estimated ILI counts divided by 100 000 persons, for Canada. The time series covers 510 consecutive weeks in the period October 2003 - June 2013. Canada has been chosen since Google® estimated ILI percentage highlights three epidemic peaks in December 2003, October - November 2009, and December 2012 - January 2013. In these periods, ILI peaked at about 7.5%, 9.7%, and 7.7% of Canadians, respectively, against normal seasonal influenza peaks of about 3.5%.

3. Beta regression. Let Y_t be a response variable bounded on the unit interval $(0,1)$, $t = 1,\ldots,n$, and let $\boldsymbol{x}_t$ be a vector of p concomitant covariates. According to Paolino (2001) and Ferrari and Cribari-Neto (2004), beta regression assumes that Y_t given $\boldsymbol{x}_t$ follows a beta distribution $\mathrm{Beta}(\mu_t, \kappa_t)$ parametrized in terms of the mean parameter $0 < \mu_t < 1$ and the precision parameter $\kappa_t > 0$. It follows that $\mathrm{var}(Y_t) = \mu_t(1-\mu_t)/(1+\kappa_t)$ and the density function of Y_t is

$$p_t(y_t;\boldsymbol{\beta}) = \frac{\Gamma(\kappa_t)}{\Gamma(\mu_t\kappa_t)\Gamma\left\{(1-\mu_t)\kappa_t\right\}} y_t^{\mu_t\kappa_t - 1}(1-y_t)^{(1-\mu_t)\kappa_t - 1}, \tag{3.1}$$

where $\Gamma(\cdot)$ denotes the Gamma function and subscript t in $p_t(\cdot)$ emphasizes the time dependence of the beta density through μ_t and κ_t.

Dependence of the response Y_t on the covariates $\boldsymbol{x}_t$ is obtained by assuming a logit-linear model for the mean parameter, $\mathrm{logit}(\mu_t) = \boldsymbol{x}_t^\top \boldsymbol{\beta}_{\boldsymbol{x}}$, where $\boldsymbol{\beta}_{\boldsymbol{x}}$ is a p-dimensional vector of coefficients. Alternative link functions $g : (0,1) \to \mathbb{R}$ are allowed, provided that they are monotonic and differentiable, such as, for example, probit and log-log. Since the distribution of bounded variables is characterized by heterogeneity, it is reasonable to model the precision parameter with a log-linear model $\log(\kappa_t) = \boldsymbol{z}_t^\top \boldsymbol{\beta}_{\boldsymbol{z}}$, where $\boldsymbol{z}$ is a set of q covariates with associated vector of coefficients $\boldsymbol{\beta}_{\boldsymbol{z}}$. Implementations of beta regression analysis for independent observations are available through `R` packages `betareg` (Cribari-Neto and Zeileis, 2010; Grün, Kosmidis, and Zeileis, 2012) and `gamlss` (Stasinopoulos and Rigby, 2007).

Within the time series framework, serial correlation in nonlinear regression analysis can be accounted for through conditional or marginal models. Following Cox (1981), conditional models are further classified as observation- and parameter-driven models. Rocha and Cribari-Neto (2009) consider observation-driven beta regression models where the response Y_t is modeled as a function of past information,

$$Y_t | \{y_{t-1}, \ldots, y_1\} \sim \mathrm{Beta}(\mu_t, \kappa_t),$$

with μ_t depending on both covariates $\boldsymbol{x}_t$ and logit-past transformed observations through the $\mathrm{ARMA}(p,q)$ model

$$\mathrm{logit}(\mu_t) = \boldsymbol{x}_t^\top \boldsymbol{\beta}_{\boldsymbol{x}} + \sum_{i=1}^{p} \psi_t \left\{ \mathrm{logit}(y_{t-i}) - \boldsymbol{x}_{t-i}^\top \boldsymbol{\beta}_{\boldsymbol{x}} \right\} + \sum_{j=1}^{q} \lambda_j \epsilon_{t-j}.$$

In the expression above, ϵ_t is a random error and $\boldsymbol{\psi} = (\psi_1, \ldots, \psi_p)^\top$ and $\boldsymbol{\lambda} = (\lambda_1, \ldots, \lambda_q)^\top$ are the autoregressive and moving average parameter vectors, respectively. Straightforward likelihood inference makes the observation-driven model appealing, although the interpretation of the regression coefficients is complicated by the presence of past transformed observations in the mean. Casarin, Dalla Valle, and Leisen (2012) develop Bayesian inference for purely autoregressive beta regression observation-driven models and discuss selection of the optimal order.

Da-Silva and Migon (2012) investigate parameter-driven beta regression models, extending Da-Silva, Migon, and Correia (2011). Da-Silva and Migon (2012) suppose responses distributed as independent beta random variables conditionally on latent variables. Serial correlation is accounted for by assuming that the latent variables evolve in time according to a state-space model. Although the hierarchical model construction is attractive, likelihood computation is complicated by the presence of n correlated latent variables. Likelihood approximation can be based on sequential simulation methods, such as, for example, the Markov chain Monte Carlo approach discussed by Da-Silva and Migon (2012).

4. Marginal beta regression time series modeling. In this paper, we develop a marginal extension of the beta regression model for time series analysis which avoids the difficulties of interpretation of the observation-driven models and the computational complications of the parameter-driven models. Thereafter, the cumulative distribution function of a non-standard normal variable with mean m and variance s^2 will be denoted by $\Phi(\cdot; m, s)$. A similar notation will be used for the density function $\phi(\cdot; m, s)$. The common simplified notation $\Phi(\cdot) = \Phi(\cdot; 0, 1)$ and $\phi(\cdot) = \phi(\cdot; 0, 1)$ is adopted for a standard normal variable.

The proposed marginal beta regression model exploits the probability integral transformation to relate response Y_t to covariates $\boldsymbol{x}_t$ and $\boldsymbol{z}_t$ and to a standard normal error ϵ_t,

$$Y_t = F_t^{-1} \left\{ \Phi(\epsilon_t); \boldsymbol{\beta} \right\}, \tag{4.1}$$

where $F_t(\cdot; \boldsymbol{\beta})$ is the cumulative distribution function associated to density (3.1), $\boldsymbol{\beta} = (\boldsymbol{\beta}_{\boldsymbol{x}}^\top, \boldsymbol{\beta}_{\boldsymbol{z}}^\top)^\top$. The probability integral transformation implies that Y_t is *marginally* beta distributed, $Y_t \sim \text{Beta}(\mu_t, \kappa_t)$. Remaining serial correlation not accounted for by covariates $\boldsymbol{x}_t$ and $\boldsymbol{z}_t$ is modeled by assuming that errors ϵ_t follow a stationary ARMA(p, q) process,

$$\epsilon_t = \sum_{i=1}^{p} \psi_i \epsilon_{t-i} + \sum_{j=1}^{q} \lambda_j \eta_{t-j} + \eta_t, \tag{4.2}$$

where η_t are independent zero-mean normal variables. In order to assure ϵ_t having unit variance, the variance of η_t is an appropriate function of the autoregressive parameter vector $\boldsymbol{\psi}$ and moving average parameter vector $\boldsymbol{\lambda}$. For example, if errors follow the AR(1) process $\epsilon_t = \psi\epsilon_{t-1} + \eta_t$, then $\text{var}(\eta_t) = 1 - \psi^2$.

The proposed beta regression model expressed by equations (4.1)-(4.2) has the advantage of separating the time series component ϵ_t from the regression part. This allows a straightforward interpretation of the regression coefficients as if observations were independent. Model (4.1)-(4.2) is an instance of Gaussian copula marginal regression (Song, 2007, Chapter 6; Masarotto and Varin, 2012).

Let $\boldsymbol{\theta}$ denote the whole parameter vector formed by the regression parameter vector $\boldsymbol{\beta}$ and the ARMA parameter vectors $\boldsymbol{\psi}$ and $\boldsymbol{\lambda}$. Inference on $\boldsymbol{\theta}$, diagnostics of departures from model assumptions, and prediction of future outcomes require the specification of the k-lags ahead predictive density $p_{t+k}(y_{t+k}|y_t, \dots, y_1; \boldsymbol{\theta})$. Such a density can be obtained by standard transformation rules as the product of the k-lags ahead predictive density of the errors and the Jacobian of the transformation $\epsilon_{t+k} = \Phi^{-1}\{F_{t+k}(y_{t+k}; \boldsymbol{\beta})\}$,

$$\begin{aligned} p_{t+k}(y_{t+k}|y_t, \dots, y_1; \boldsymbol{\theta}) &= p(\epsilon_{t+k}|\epsilon_t, \dots, \epsilon_1; \boldsymbol{\theta}) \left| \frac{d\epsilon_{t+k}}{dy_{t+k}} \right| \\ &= p_{t+k}(y_{t+k}; \boldsymbol{\beta}) \frac{p(\epsilon_{t+k}|\epsilon_t, \dots, \epsilon_1; \boldsymbol{\theta})}{p(\epsilon_{t+k}; \boldsymbol{\beta})} \\ &= p_{t+k}(y_{t+k}; \boldsymbol{\beta}) \frac{\phi(\epsilon_{t+k};\, m_{t+k|t}, s_{t+k|t})}{\phi(\epsilon_{t+k})}, \end{aligned} \tag{4.3}$$

where $m_{t+k|t} = \text{E}(\epsilon_{t+k}|\epsilon_t, \dots, \epsilon_1; \boldsymbol{\theta})$ and $s^2_{t+k|t} = \text{var}(\epsilon_{t+k}|\epsilon_t, \dots, \epsilon_1; \boldsymbol{\theta})$. Both conditional expectations can be efficiently evaluated in a linear number of operations via Kalman filter recursions.

Expression (4.3) is particularly attractive in terms of interpretability, since it separates the marginal density associated to the future observation, $p_{t+k}(y_{t+k}; \boldsymbol{\beta})$, from a measure of the serial correlation within the errors. Figure 2 provides an illustration of the beta regression model with ARMA$(2, 1)$ errors used for the simulation study in Section 5. The marginal density $p_{t+k}(y_{t+k}; \boldsymbol{\beta})$ and the predictive density $p_{t+k}(y_{t+k}|y_t, \dots, y_1; \boldsymbol{\theta})$ substantially differ for short time prediction, with the predictive density being more peaked since it accounts for the information in the past observations. As the prediction lag increases, past data become less informative, thus making the predictive density closer to the marginal density, as expected.

Basic properties of the ARMA(p, q) process are inherited by the proposed model. In fact, it is immediate from (4.3) that if errors ϵ_t follow a MA(q)

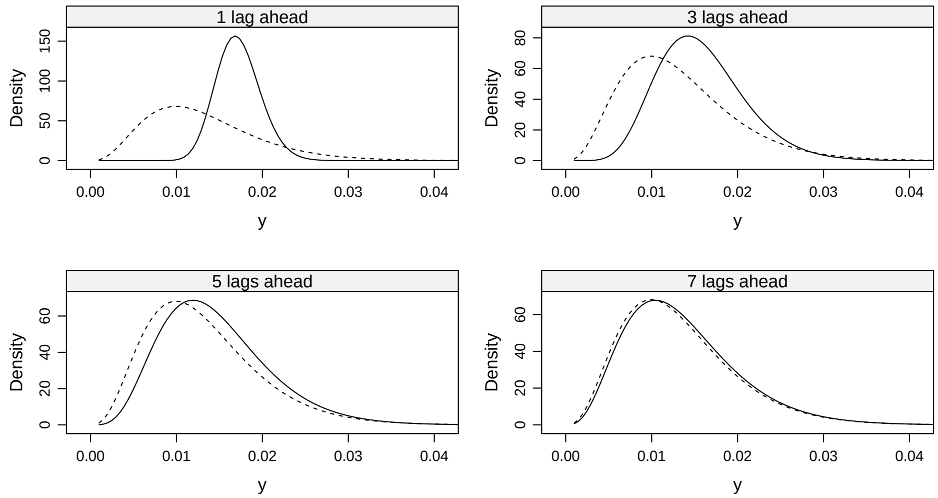

FIG 2. *Predictive density (solid line) and marginal density (dashed line) at different lags ahead for the marginal beta regression model with ARMA(2,1) errors described in the simulation study, Section 5.*

process, then observations far apart more than q units are independent. Moreover, if errors ϵ_t follow an AR(p) process, then observations follow a Markovian process of order p.

By model construction, the predictive cumulative distribution function of Y_{t+k} given $\{y_t, \dots, y_1\}$ coincides with the predictive cumulative distribution function of ϵ_{t+k} given $\{\epsilon_t, \dots, \epsilon_1\}$,

$$\begin{aligned} F_{t+k}(y_{t+k}|y_t, \dots, y_1; \boldsymbol{\theta}) &= \int_0^{y_{t+k}} p_{t+k}(u|y_t, \dots, y_1; \boldsymbol{\theta}) du \\ &= \int_{-\infty}^{\Phi^{-1}\{F_{t+k}(y_{t+k};\boldsymbol{\beta})\}} p(\epsilon_{t+k}|\epsilon_t, \dots, \epsilon_1; \boldsymbol{\theta}) d\epsilon_{t+k} \\ &= \Phi\left(\epsilon_{t+k};\, m_{t+k|t}, s_{t+k|t}\right). \end{aligned} \tag{4.4}$$

Accordingly, the α-quantile of the predictive distribution is

$$y_{t+k|t;\alpha} = F_{t+k}^{-1}\left[\Phi\left\{m_{t+k|t} + \Phi^{-1}(\alpha)\, s_{t+k|t}\right\}; \boldsymbol{\beta}\right].$$

4.1. *Likelihood inference.* We suggest to perform inference by relying on maximum likelihood estimation. Let $L_{\text{ind}}(\boldsymbol{\beta}; \boldsymbol{y}) = \prod_{t=1}^n p_t(y_t; \boldsymbol{\beta})$ denote the likelihood constructed under the assumption of independence. Then, given the result in (4.3), the likelihood function for $\boldsymbol{\theta}$ is

$$L(\boldsymbol{\theta};\boldsymbol{y}) = p_1(y_1;\boldsymbol{\beta})\prod_{t=2}^{n} p_t(y_t|y_{t-1},\dots,y_1;\boldsymbol{\theta})$$
$$= L_{\text{ind}}(\boldsymbol{\beta};\boldsymbol{y})\prod_{t=2}^{n}\frac{p(\epsilon_t|\epsilon_{t-1},\dots,\epsilon_1;\boldsymbol{\theta})}{p(\epsilon_t;\boldsymbol{\beta})}.$$

The likelihood function is the product of the independence likelihood L_{ind} and a calibration term accounting for the presence of dependence of ϵ_t on past values. A calibration term significantly different from one is indicative of dependence.

From a practical point of view, the closed-form of the likelihood implies an effortless computation. As already noted for the predictive density, Kalman filter can be employed for efficient computation of the predictive densities of the ARMA(p, q) errors, $p(\epsilon_t|\epsilon_{t-1},\dots,\epsilon_1;\boldsymbol{\theta})$, thus making the computational complexity of likelihood evaluation of linear order.

4.2. *Predictive quantile residuals.* Following Dunn and Smyth (1996) and Masarotto and Varin (2012), model validation can be based on the analysis of the predictive quantile residuals

$$r_t = \Phi^{-1}\left\{F_t(y_t|y_{t-1},\dots,y_1;\hat{\boldsymbol{\theta}})\right\},$$

where $\hat{\boldsymbol{\theta}}$ denotes the maximum likelihood estimate of $\boldsymbol{\theta}$. Given (4.4), predictive quantile residuals r_t assume the familiar form

$$r_t = \frac{\hat{\epsilon}_t - \hat{m}_{t|t-1}}{\hat{s}_{t|t-1}},$$

where $\hat{\epsilon}_t$, $\hat{m}_{t|t-1}$ and $\hat{s}_{t|t-1}$ are evaluated at $\hat{\boldsymbol{\theta}}$. Residuals r_t are realizations of n independent standard normal variables if the model assumptions are met.

5. Simulation study. A simulation study has been performed in order to evaluate maximum likelihood estimation and prediction for the proposed marginal beta regression model. The simulation set-up consists of 1 000 weekly time series from the marginal beta regression model specified as follows. The length of the time series is set equal to 368, with the first $n = 52 \times 7 = 364$ observations used for model fitting and the remaining four observations used for prediction. Following common practice in surveillance literature (Unkel et al., 2012), mean μ_t and precision κ_t include linear trend and annual seasonal components representing temperature variations,

$$\begin{aligned} \text{logit}(\mu_t) &= \beta_{0\boldsymbol{x}} + \beta_{1\boldsymbol{x}}\tilde{t} + \beta_{2\boldsymbol{x}} \sin\left(\frac{2\pi t}{52}\right) + \beta_{3\boldsymbol{x}} \cos\left(\frac{2\pi t}{52}\right), \\ \log(\kappa_t) &= \beta_{0\boldsymbol{z}} + \beta_{1\boldsymbol{z}}\tilde{t} + \beta_{2\boldsymbol{z}} \sin\left(\frac{2\pi t}{52}\right) + \beta_{3\boldsymbol{z}} \cos\left(\frac{2\pi t}{52}\right), \end{aligned} \tag{5.1}$$

where $\tilde{t}$ indicates the time index t centered and scaled by factor 100 in such a way to avoid numerical instabilities. The residual serial correlation is modeled by assuming an ARMA(2,1) process for the errors. The values of the parameters are set equal to $\beta_{0\boldsymbol{x}} = -4.00$, $\beta_{1\boldsymbol{x}} = 0.15$, $\beta_{2\boldsymbol{x}} = -0.22$, $\beta_{3\boldsymbol{x}} = -0.67$, $\beta_{0\boldsymbol{z}} = 6.00$, $\beta_{1\boldsymbol{z}} = 0.10$, $\beta_{2\boldsymbol{z}} = -0.06$, $\beta_{3\boldsymbol{z}} = -0.19$, $\psi_1 = 1.50$, $\psi_2 = -0.60$, and $\lambda = -0.30$. The values of $\beta_{2\boldsymbol{x}}$, $\beta_{3\boldsymbol{x}}$, $\beta_{2\boldsymbol{z}}$, and $\beta_{3\boldsymbol{z}}$ are chosen in order to guarantee an amplitude equal to 0.7 and 0.2 for the mean and the precision, respectively, and a phase shift equal to 0.6π for both mean and precision. These values resemble a typical ILI weekly time series.

Table 1 displays average and standard deviation of the parameter estimates, and average of the standard errors computed from the inverse of the observed Fisher information. The results are satisfactory, as they show (i) a negligible bias in the estimation of all the parameters and (ii) averages of the standard errors close to standard deviations of the estimates.

TABLE 1
Average (`ave`), standard deviation (`s.d.`), and average of standard errors (`s.e.`) for 1 000 simulated estimates based on a beta regression model with ARMA(2,1) errors and with independent errors.

			ARMA(2,1)			independence		
		true	ave	s.d.	s.e.	ave	s.d.	s.e.
mean	intercept	-4.00	-4.01	0.06	0.05	-4.01	0.06	0.02
	trend	0.15	0.15	0.05	0.04	0.15	0.05	0.02
	cosine term	-0.22	-0.22	0.07	0.06	-0.22	0.07	0.02
	sine term	-0.67	-0.67	0.08	0.07	-0.67	0.08	0.03
precision	intercept	6.00	6.11	0.17	0.17	6.15	0.18	0.08
	trend	0.10	0.10	0.07	0.07	0.12	0.18	0.07
	cosine term	-0.06	-0.06	0.11	0.11	-0.06	0.24	0.10
	sine term	-0.19	-0.20	0.11	0.11	-0.22	0.25	0.11
errors	ar1	1.50	1.51	0.12	0.11	–	–	–
	ar2	-0.60	-0.62	0.11	0.09	–	–	–
	ma1	-0.30	-0.33	0.15	0.13	–	–	–

Table 2 reports the empirical coverage of prediction intervals at lags one to four, either for the fitted model with ARMA(2,1) errors and for the independence model. Prediction intervals from model with ARMA(2,1) errors are sensibly closer to the nominal level than those based on the independence model.

TABLE 2
Empirical coverage of prediction intervals at various lags ahead for 1 000 simulated time series based on a beta regression model with ARMA(2,1) errors and with independent errors.

		ARMA(2,1)				independence			
		lag 1	lag 2	lag 3	lag 4	lag 1	lag 2	lag 3	lag 4
	90%	0.895	0.886	0.870	0.885	0.880	0.868	0.857	0.851
levels	95%	0.948	0.933	0.930	0.930	0.932	0.932	0.913	0.900
	99%	0.985	0.985	0.978	0.973	0.971	0.970	0.956	0.948

6. Monitoring outbreaks of disease. Quality control charts are typically employed for on-line detection of outbreaks of infectious diseases, *e.g.*, Woodall (2006) and Unkel et al. (2012). To this aim, the first step is the identification of a model describing the pattern of ordinary influenza seasons. Then, departures from the model-expected influenza levels are interpreted as symptoms of anomalies. Cumulative sum (CUSUM) charts (Montgomery, 2009, Chapter 9) are appropriate for monitoring long-lasting illnesses such as ILI, given the capability of early detection of small variations in the mean disease level. In fact, CUSUM charts are employed by the Centers for Disease Control and Prevention for routinely syndromic surveillance (Hutwagner et al., 2003).

CUSUM charts are typically constructed under the assumption of independent observations from a normal distribution, at least approximately. Accordingly, below we suggest to monitor influenza disease through predictive quantile residuals r_t. Bilateral CUSUM chart is based on the positive C_t^+ and the negative C_t^- cumulative sums of r_t,

$$
\begin{aligned}
C_t^+ &= \max\{0, r_t - k + C_{t-1}^+\},\\
C_t^- &= \max\{0, -k - r_t + C_{t-1}^-\},
\end{aligned}
$$

for a *reference value* k and with $C_0 = 0$. The process is out-of-control if either C_t^+ or C_t^- exceeds the *decision limit* h. Parameters k and h are chosen in order to guarantee an acceptable capability to detect influenza levels anomalies and, in the meanwhile, a low number of false alarms. Following standard recommendations in quality control literature (Montgomery, 2009), the chart parameters can be set to values $k = 0.5$ and $h = 4$.

Standard application of CUSUM charts involves two phases. In Phase I, historical data are analyzed to calibrate the chart when the process is under control. Phase II is the on-line monitoring stage based on the chart calibrated at the previous phase. Details are given below.

1. Phase I
   (a) Fit the beta marginal regression model including trend, seasonality, and ARMA(p, q) errors, with p and q large enough to guarantee residual autocorrelation to be captured. As a rule of thumb, we suggest $p = q = 3$.
   (b) Remove the anomalous observations identified by a CUSUM chart of the predictive quantile residuals derived from the model fitted at step (a).
   (c) Re-estimate the beta marginal regression model on the time series without the anomalous observations. Choose the most appropriate ARMA(p, q) structure, $p \leq 3$ and $q \leq 3$, via information criteria or cross-validation. The chosen model is the best model representation of a regular seasonal influenza.
2. Phase II
   (d) On-line monitor influenza outbreaks by the unilateral positive CUSUM chart of the predictive quantile residuals derived from the model selected at Phase I, step (c).

7. Application to Canada Google® Flu Trends. In this section, we illustrate the application of the methodology previously described to the analysis of Canada Google® Flu Trends data.

In order to illustrate the surveillance procedure of Section 6, we used data until June 2010 for model calibration (Phase I), while the following three years of observations are used for on-line monitoring (Phase II). The initial CUSUM chart based on the ARMA$(3, 3)$ model in Phase I identifies 19 anomalous observations over 354 observations. The subsequent step is the estimation of all possible models with ARMA(p, q) errors, $p \leq 3$ and $q \leq 3$, to the data after removal of the 19 anomalous observations. Table 3 ranks the sixteen possible models in terms of Akaike Information Criterion. The preferred model is the one with ARMA$(2, 1)$ errors. However, results highlight that a precise identification of p and q is not crucial, since many models induce essentially the same autocorrelation structure, see Table 3.

The application of the CUSUM chart in Phase II requires the predictive quantile residuals being comparable to a set of independent normal variables. The graphical examination of the predictive quantile residuals reported in Figure 3 sustains such a requirement.

TABLE 3
Canada Google® Flu Trends data. Estimated beta marginal regression models with ARMA(p, q) errors ranked according to the Akaike Information Criterion (AIC) and corresponding autocorrelation of the errors at lags one to four.

	ARMA			autocorrelations			
rank	p	q	AIC	lag 1	lag 2	lag 3	lag 4
1	2	1	-3372.45	0.94	0.84	0.74	0.64
2	3	0	-3372.37	0.94	0.84	0.74	0.64
3	2	0	-3371.57	0.94	0.84	0.75	0.66
4	1	2	-3371.47	0.94	0.84	0.74	0.66
5	3	1	-3370.49	0.94	0.84	0.74	0.64
6	2	2	-3370.46	0.94	0.84	0.74	0.64
7	1	3	-3369.77	0.94	0.84	0.74	0.65
8	3	2	-3368.66	0.94	0.84	0.74	0.64
9	2	3	-3367.87	0.94	0.84	0.74	0.65
10	3	3	-3367.23	0.94	0.84	0.74	0.64
11	1	1	-3366.89	0.93	0.85	0.77	0.70
12	1	0	-3353.23	0.93	0.87	0.81	0.75
13	0	3	-3269.01	0.78	0.42	0.12	0.00
14	0	2	-3185.59	0.68	0.24	0.00	0.00
15	0	1	-3038.51	0.49	0.00	0.00	0.00
16	0	0	-2766.91	0.00	0.00	0.00	0.00

Phase II CUSUM chart for on-line monitoring is illustrated in Figure 4. The corresponding points above the decision limit $h = 4$ in the influenza time series are highlighted in the bottom panel of Figure 4. The process is under control until December 9, 2012, and then it remains out-of-control for eigth consecutive weeks before returning under control. The out-of-control weeks correspond to the epidemic peak occurred in December 2012 - January 2013.

7.1. *Holiday peaks.* As observed by a referee, Canada Google® Flu Trends data show a peak-valley-peak pattern within a couple of weeks at the beginning of most of the observed years, see Figure 1. Accordingly, we investigated the presence of a "holiday effect", related to the Christmas/New Year period. Table 4 reports estimates and standard errors for the parameters of the beta marginal regression model with trend, sine and cosine terms describing seasonal temperature variations, ARMA$(2, 1)$ errors, and the dummy variable for the holiday weeks. Results indicate no significant trend in the mean, which is instead significant for the precision. The annual seasonal component is highly significant in both mean and precision, as expected. The analysis confirms a very significant increase of ILI in correspondence with the holiday weeks, given an estimated holiday effect parameter in the

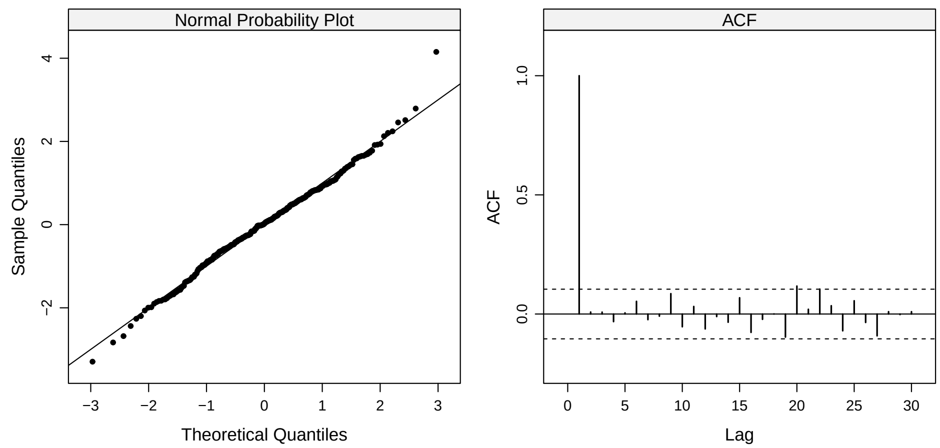

FIG 3. *Canada Google® Flu Trends data. Normal probability plot (left panel) and autocorrelation function (right panel) of the predictive quantile residuals for the fitted marginal beta regression model with ARMA(2,1) errors.*

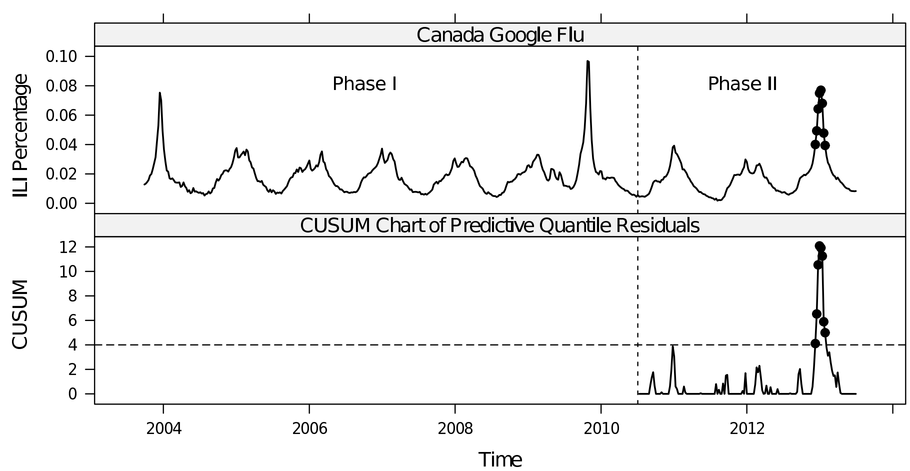

FIG 4. *Canada Google® Flu Trends data. Positive CUSUM chart for surveillance of influenza outbreaks. Circles indicate out-of-control weeks.*

mean equal to 0.11, with a standard error of 0.02. Conversely, there is no significant effect in terms of precision (estimate 0.12, standard error 0.09).

Further confirmations of the relevance of the holiday effect are provided by AIC, which increases from -5057.31 to -5028.74, and by the profile loglikelihood for the associated coefficient, displayed in Figure 5.

Table 4
Canada Google® Flu Trends data. Estimates and standard errors for the parameters of fitted marginal beta regression model without and with holiday effect. Akaike Information Criterion (AIC) statistic also reported.

	parameter	no holiday effect est.	no holiday effect s.e.	holiday effect est.	holiday effect s.e.
mean	intercept	-4.14	0.05	-4.14	0.05
	trend	-0.16	0.33	0.05	0.33
	sine term	0.66	0.06	0.65	0.06
	cosine term	-0.31	0.06	-0.31	0.06
	Christmas/New Year	–	–	0.11	0.02
precision	intercept	6.23	0.11	6.19	0.11
	trend	1.46	0.43	1.68	0.43
	sine term	-0.48	0.09	-0.37	0.10
	cosine term	-0.04	0.10	-0.08	0.09
	Christmas/New Year	–	–	0.12	0.09
ARMA	ar1	1.52	0.07	1.57	0.06
	ar2	-0.60	0.07	-0.64	0.06
	ma1	-0.25	0.09	-0.28	0.08
	AIC	-5028.74		-5057.31	

8. Conclusions. This paper suggested a practical approach for analysis of bounded time series defined on the unit interval. One of the advantages of the proposed marginal model is the reproducible interpretation of the regression parameters, whose meaning does not depend on the ARMA structure. The robust interpretation of the regression parameters is a property not shared by alternative conditionally specified models, such as observation- and parameter-driven beta regression models briefly described in Section 3. Another advantage of the proposed approach is that inferential and prediction tasks have convenient expressions, thus making modeling time series on the unit scale feasible as a practical alternative to the common logit-transformation approach.

Several extensions of the proposed modeling framework are possible. First, the approach has a trivial extension to time series defined on an arbitrary (a, b) interval. Second, spatial and spatio-temporal beta regression models can be constructed by assuming that the errors are realizations of a Gaussian random field. Finally, the model can be extended to allow for exact zeros and ones, by using the zero-or-one beta inflated regression model (Ospina and Ferrari, 2012) to define the univariate marginal distributions.

References.

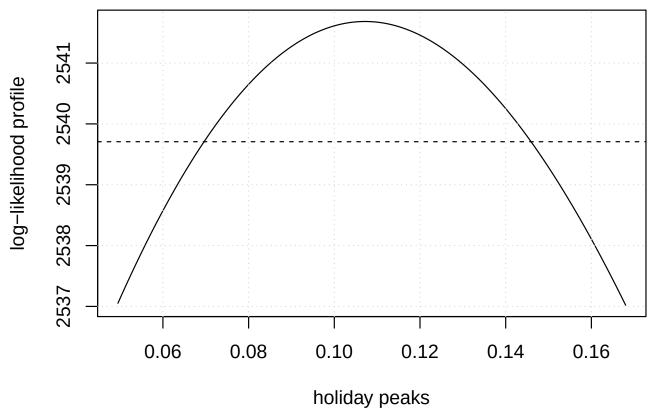

Fig 5. *Canada Google® Flu Trends data. Profile log-likelihood for holiday effect parameter. Horizontal dashed line corresponds to 95% asymptotic confidence interval.*

Butler, D. (2013). When Google got flu wrong. *Nature* **494**, 155–156.

Casarin, R., Dalla Valle, L., and Leisen, F. (2012). Bayesian model selection for beta autoregressive processes. *Bayesian Analysis* **7**, 385–410.

Cribari-Neto, F. and Zeileis, A. (2010). Beta regression in R. *Journal of Statistical Software* **34**, Issue 2, 1–24.

Cox, D.R. (1981). Statistical analysis of time series: Some recent developments. *Scandinavian Journal of Statistics* **8**, 93–115.

Da-Silva, C.Q., Migon, H.S., and Correia, L.T. (2011). Dynamic Bayesian beta models. *Computational Statistics and Data Analysis* **55**, 2074–2089.

Da-Silva, C.Q. and Migon, H.S. (2012). Hierarchical dynamic beta model. Technical report n. 253. Department of Statistics, Federal University of Rio de Janeiro.

Dunn, P.K. and Smyth, G.K. (1996). Randomized quantile residuals. *Journal of Computational and Graphical Statistics* **5**, 236–244.

Ferrari, S.L.P. and Cribari-Neto, F. (2004). Beta regression for modelling rates and proportions. *Journal of Applied Statistics* **31**, 799–815.

Ginsberg, J., Mohebbi, M.H., Patel, R.S., Brammer, L., Smolinski, M.S., and Brilliant, L. (2009). Detecting influenza epidemics using search engine query data. *Nature* **457**, 1012–1014.

Grün, B., Kosmidis, I., and Zeileis, A. (2012). Extended beta regression in R: Shaken, stirred, mixed, and partitioned. *Journal of Statistical Software* **48**, Issue 11, 1–25.

Hutwagner, L., Thompson, W.W., Seeman, G.M., and Treadwell, T. (2003). The bioterrorism preparedness and response early aberration reporting system (EARS). *Journal of Urban Health* **80**, 89–96.

Love, T.M.T., Thurson, S.W., Keefer, M.C., Dewhurst, S., and Lee, H.Y. (2010). Mathematical modeling of ultradeep sequencing data reveals that acute CD8+ T-lymphocyte responses exert strong selective pressure in simian immunodeficiency virus-infected macaques but still fail to clear founder epitope sequences. *Journal of Virology* **84**, 5802–5814.

Kieschnick, R. and McCullough, B.D. (2003). Regression analysis of variates observed on (0,1): Percentages, proportions and fractions. *Statistical Modelling* **3**, 193–213.

Masarotto, G. and Varin, C. (2012). Gaussian copula marginal regression. *Electronic Journal of Statistics* **6**, 1517–1549.

Montgomery, D.C. (2009). *Introduction to Statistical Quality Control*. New York: John Wiley & Sons, 6th Edition.

Ospina, R. and Ferrari, S.L.P. (2012). A general class of zero-or-one inflated beta regression models. *Computational Statistics & Data Analysis* **56**, 1609–1623.

Paolino, P. (2001). Maximum likelihood estimation of models with beta-distributed dependent variables. *Political Analysis* **9**, 325–346.

R Core Team (2012). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, `http://www.R-project.org/`.

Rocha, A.V. and Cribari-Neto, F. (2009). Beta autoregressive moving average models. *Test* **18**, 529–545.

Rogers, J.A., Polhamus, D., Gillespie, W.R., Ito, K., Romero, K., Qiu, R., Stephenson, D., Gastonguay, M.R., and Corrigan, B. (2012). Combining patient-level and summary-level data for Alzheimer's disease modeling and simulation: A beta regression meta-analysis. *Journal of Pharmacokinetics and Pharmacodynamics* **39**, 479–498.

Smithson, M. and Verkuilen, J. (2006). A better lemon squeezer? Maximum-likelihood regression with beta-distributed dependent variables. *Psychological Methods* **11**, 54–71.

Song, P.X.-K. (2007). *Correlated Data Analysis: Modeling, Analytics, and Applications*. New York: Springer.

Stasinopoulos, D.M. and Rigby, R.A. (2007). Generalized additive models for location scale and shape (gamlss) in `R`. *Journal of Statistical Software* **23**, Issue 7, 1–46.

Unkel, S., Farrington, P., Garthwaite, P., Robertson, C., and Andrews, N. (2012). Statistical methods for the prospective detection of infectious disease outbreaks: A review. *Journal of the Royal Statistical Society Series A* **175**, 49–82.

Wang, X.F. (2012). Joint generalized models for multi-dimensional outcomes: A case study of neuroscience data from multi-modalities. *Biometrical Journal* **54**, 264–280.

Wang, W., Scharfstein, D., Wang, C., Daniels, M., Needham, D., and Brower, R. (2011). Estimating the causal effect of low tidal volume ventilation on survival in patients with acute lung injury. *Applied Statistics* **60**, 475–496.

Woodall, W. (2006). The use of control chart in health-care and public-health surveillance. *Journal of Quality Technology* **38**, 89–104.

Zou, K.H., Carlsson, M.O., and Quinn, S.A. (2010). Beta-mapping and beta-regression for changes of ordinal-rating measurements on Likert scales: A comparison of the change scores among multiple treatment groups. *Statistics in Medicine* **29**, 2486–2500.

Annamaria Guolo
Department of Economics
Università di Verona
Via dell'Artigliere, 19
I-37129 Verona, Italy
E-mail: annamaria.guolo@univr.it

Cristiano Varin
Department of Environmental Sciences, Informatics and Statistics
Università Ca' Foscari Venezia
San Giobbe Cannaregio, 873
I-30121 Venice, Italy
E-mail: sammy@unive.it