A Novel Binary Time Series Model of the Gastric Dilation-Volvulus

by

M. Levine
Purdue University

G.E. Moore

Technical Report #08-03

Department of Statistics
Purdue University

August 2008

A novel binary time series model of the gastric dilatation-volvulus

Michael Levine [*], George E. Moore [†‡]

Abstract

Gastric dilatation-volvulus (GDV) is a life-threatening condition in many mammals, including humans, that has a special importance in large breed dogs. The study of its etiological factors is notoriously difficult due to the variety of possible living conditions. This study introduces the binary time series approach to the investigation of the possible GDV risk factors. Possible outcomes are coded as 0 for a day with no GDV case and 1 for a day when GDV occurs. The data collected in the population of high-risk working dogs in Texas was used.

The binary time series method had not been applied to the GDV occurrence investigation before. The method is based on interpreting the GDV occurrence data as binary time series with probabilities of an outcome depending on extraneous risk factors as well as past values of the series itself. Since the census data for subjects are not available for the full observation period, some of them are imputed using a simple linear time series model. A number of temperature- and atmospheric pressure-related factors turns out to be important in determining the likelihood of GDV occurrence on a given day. All of the temperature related factors are negatively associated with the likelihood of GDV while all of the atmospheric pressure related factors are positively associated with it. All of the odds of a day being GDV day are interpreted conditionally on the past GDV occurrences. Possible ways of handling the problem when the number of working dogs is large and the possible number of GDV cases may be more than one a day are also suggested.

[*]Corresponding author. Address: 250 N. University Street, Purdue University, West Lafayette, IN 47907. E-mail: mlevins@stat.purdue.edu. Phone: 765-496-7571. Fax: 765-494-0558

[†]725 Harrison Street, School of Veterinary Medicine, Purdue University, West Lafayette, IN 47907. Email: gemoore@purdue.edu

[‡]Keywords and phrases: time series, categorical, gastric dilatation-volvulus

1 Introduction

Gastric dilatation-volvulus (GDV) is a life threatening condition in which the stomach dilates and rotates on itself. It is very frequent in dogs [21] but can also affect other species. It is noteworthy that it had been diagnosed in humans as well where it can be a cause of upper gastrointestinal obstruction (see [20, 13, 8, 19]). The physical mechanisms of this condition are well understood but the etiology is not [4]. A number of possible risk factors in dogs have been suggested; among them, temperament of the dog (excitability), large or giant breed, increased thoracic depth to width ratio, rapid food consumption etc. [4, 18, 16]. However, many clinical findings of the onset of disease remain unexplained by these hypotheses.

Thus, the study of the GDV risk factors is an important and not very well researched area. Statistical methods applied in the past include the principal component analysis [10] to select possibly important climatologic factors. Univariate associations between seasonal risk factors and GDV were tested using χ^2 analysis without developing any models. Due to the nature of the principal component analysis, this obscured the impact of a single weather-related variable. [6] reported a statistically significant association between temperature and GDV occurrence; in particular, the mean temperature on days with GDV cases was significantly higher than on the days without GDV cases. However, the absolute size of the effect was too small for it to have any clinical significance. [15] used a simple logistic regression analysis to find the odds of a number of meteorologic variables influencing the incidence of GDV. In particular, significant negative association between maximum daily temperature and the risk of GDV occurrence was detected. The applicability of the simple logistic regression approach is, probably, limited because often it cannot be safely assumed that GDV incidences recorded over time are uncorrelated. The total number of subjects being observed over time may be changing and not always known which also impedes the applicability of the logistic regression model.

As far as we know, no investigators so far paid any attention to the possible temporal autocorrelation of GDV occurrence data. Many commonly presumed etiological factors, such as barometric pressure, air temperature, are time series that are correlated over time. It has been hypothesized that there is association between GDV and the weather conditions shortly before its occurrence [10, 6]. Since there are many possible climatologic factors describing weather conditions, it is likely that some of them may not be considered by a researcher when constructing the model, for example due to a lack of time. If the GDV occurrence is influenced by past values of such a factor (which is commonly a time

series), its omission will induce the temporal autocorrelation in the GDV occurrence data. The response is usually recorded as a number of GDV cases over a certain period time. Therefore, it is possible to end up with a series of categorical data recorded over time that exhibits temporal autocorrelation. Such data are commonly called *categorical time series.*

The categorical time series approach has almost no history in medical and/or veterinary applications of statistic. So far, the only attempt to apply categorical time series approach in the veterinary literature that we are aware of is [3]. In medical literature, while studying the influence of the climate on the occurrence of human plague, [17] mentions that, in addition to its chosen approach of Poisson regression with time-dependent covariates, alternative models "... are available that mix methods from time series analysis and generalized linear models theory" but do not pursue this line of inquiry any further. The current research aims to fill this gap.

A number of ways to model categorical time series is possible. Historically, the first one was using a Markov chain of some order $p > 0$ to model the conditional probability of the future outcome Y_t given p past outcomes $Y_{t-1}, \ldots, Y_{t-p}$. An excellent monograph on the practical aspects of Markov chain modeling is [9]. However, this approach has serious limitations in the context of GDV data. First, the number of parameters to be estimated grows exponentially as the order of the Markov chain p grows. Second, this approach requires prior identification of the joint dynamics of response and covariates which is usually quite difficult to do in practice. Finally, it is usually necessary to use other covariates besides just the past values of the response in order to build a reasonable model. If this is the case, such covariates (often called extraneous covariates) are very difficult to incorporate using the Markov chain approach.

To avoid the just described problems, a different approach to modeling categorical time series has been adopted that has been suggested by [12]. It puts the categorical time series into the generalized linear model (GLM) framework. The most important benefit of this approach is that neither Markov nor stationarity assumptions are required. It is also important to notice here that it utilizes heavily the classical estimation and inference framework for GLM that is very well researched and has been a subject of the classical monograph by [14]. The only substantial difference is that instead of the standard maximum likelihood estimation commonly used for classical GLM's a partial likelihood inference has to be used. It effectively excludes information about the parameter(s) of interest that is contained in the extraneous covariates; it has been shown that this is a reasonable approach under some simple regularity conditions. For details, see [12].

As mentioned before, the covariates commonly considered as possible etiological factors of GDV research, such as barometric pressure, humidity, air temperature and others, are the time series themselves. Since there is always some likelihood that some of them may be omitted, the categorical time series models is likely to be a very suitable tool for the research on the possible etiological factors of GDV. In general, the choice of the risk factors for GDV is difficult due to a variety of the living conditions of the dogs, e.g. diet, exercise, stress, housing environments etc., for affected dogs. (see [4]). To overcome this, [10] tried to control the environment by studying GDV in military dogs in training at the Military Working Dog (MWD) Training Center at Lackland Air Force Base (LAFB) where a large population of large-breed dogs is trained. These dogs are trained 5 days a week to acquire a number of skills, such as contraband detection, using standard training protocols. They are fed standard diet, housed in outdoor runs and are under observation 24 hours a day. All of the above provides a relatively controlled environment in which to study the GDV occurrence.

Earlier, [15] studied possible associations between maximum (minimum) daily temperature and/or atmospheric pressure on the occurrence of GDV in the abovementioned dog population at LAFB. It has been hypothesized that these factors may influence not only current, but also future rates of GDV occurrence [10, 6]; because of that, values of the maximum daily atmospheric pressure and temperature recorded on the day before the GDV event have also been used as possible external covariates. The conclusion in [15] was that the maximum daily temperature on the day of GDV event and the day before the GDV event were the most important factors determining the rate of GDV occurrence. Other factors that were found to have played a significant role were the minimum atmospheric pressure on the day of GDV event and the day before it, the maximum hourly rise in temperature on the day of GDV and the maximum hourly drop in temperature on the day of GDV event. The current research aims to consider the factors first enumerated in [15] as possible causes of GDV in the much more general framework of categorical time series modeling.

Finally, it is necessary to describe the data used in this study in detail.In addition to the GDV occurrence data, our study also uses the climate dataset. The GDV occurrence data set consists of all recorded cases of GDV among the military working dogs (MWD) at the Lackland Airforce Base (LAFB) from Jan 1993 through Jan 1998. In each case, the breed of affected dog, its sex, date of birth, age at the onset of the disease and weight were recorded. The first recorded case of GDV occurred on Jan 6, 1993 and the last one on Dec 25th, 1998. The total number of recorded cases (i.e. the days on which GDV case

was registered) is 60. Out of 60, only two days involved more than one case of GDV; in both of them, there were 2 affected dogs.

The number of dogs under observation at LAFB was not constant but rather changing from month to month. The monthly dog census data were available Oct 1993 through Dec 1997 only. It started with 357 dogs in October 1993 and ended with 281 dog in December 1997. It is not clear why the number of dogs under observation was constantly changing. In particular, it is not known whether the new dogs were being brought in or removed from the population during the period of study or if the only reason for the change was the natural birth/death process. It is also not known how long the dogs who were already present at the beginning of the observation period have been at the LAFB before then.

Finally, in order to investigate the influence of meteorologic factors on the GDV occurrence, a large database of weather data has been assembled from the National Climatic Data Center at the Kelly Air Force Base, located immediately adjacent to LAFB. It contains hourly data on the wind direction, speed and wind gust; hourly temperature in Fahrenheit degrees, both adjusted and unadjusted for humidity; atmospheric pressure in inches of mercury, adjusted to the sea-level and the unadjusted one in millibars; also ceiling (the height of the lowest clouds) and the sky condition, based on the subjective evaluation by observer (codes used are from NOAA glossary and, finally, visibility (the units were not explained in the original database).)

2 The categorical time series for modeling the future incidents of GDV

2.1 Categorical time series models

It is quite reasonable to assume that the incidences of GDV should be viewed as binary time series, that is, the time series Y_t that can only take values 0 or 1. Such time series is a special case of categorical time series that can take on up to m different values with m being a positive integer. The categorical time series is not a well-known research area in the general time series field of study; of the few literary references, one can mention a monograph on the subject by [12] as well as their excellent review paper [7]. One of the reasons why such an assumption is plausible is that the past values of possible etiological factors of GDV, such as atmospheric pressure and air humidity, influence the current GDV occurrence. In practice, since etiological factors of GDV are not well known (see [4]) and the number of possible factors is large, it is quite likely that a number of

possible explanatory covariates are unintentionally omitted from the model. It has been mentioned already in the previous section that such a situation is likely to lead to temporal autocorrelation of response values which suggests categorical time series as means of study.

Historically, the most common approach to modeling categorical time series has been based on viewing them as Markov chain of some order $p > 0$. A Markov chain of order p is defined as the process Y_t, $t = 1, \ldots, N$ such that its current value Y_t depends on at most p past values $Y_{t-1}, \ldots, Y_{t-p}$. The process Y_t is *discrete-valued* in the sense that at any time t, Y_t can take on any of m values with m being some integer; m is called the number of states of the Markov chain. More formally, $P(Y_t = k|Y_{t-1}, Y_{t-2}, \ldots) = P(Y_t = k|Y_{t-1}, \ldots, Y_{t-p})$, $k = 1, \ldots, m$. This approach is somewhat problematic since it implies the exponential growth of the number of parameters that needs to be estimated as the sample size grows. As a matter of fact, it is easy to realize that for a Markov chain of order p the number of parameters to be estimated is equal to $m^p(m-1)$. Since we do not know in advance how many lags of the covariates p should be included in the model, this number can become unacceptably large in practice. Thus, even for a small sample size the number of parameters to be estimated may become prohibitively large.

We intend to use a more general approach advocated by [12] that is based on modeling categorical time series as generalized linear models. It may also be called a regression model based approach. Unlike Markov chain approach, it does not cause the number of parameters to be estimated to grow exponentially with the sample size; it is also broad enough to encompass most of the practically important categorical time series models. Note that neither Markov property nor stationarity have to be assumed which is important since both of these properties may be difficult to verify in practice. These models can be estimated using the same method (iterative reweighted least squares, IWLS for short) as regular generalized linear models; the only difference is that the results have to be interpreted conditionally on the past.

Some guidance on the possible covariate selection can be gleaned from the past work on the subject in the medical literature. As pointed out before (see, for example, [10, 6]), there may be some correlation between the occurrence of GDV and the weather conditions. [10] noted seasonal variation in the occurrence of GDV, but did not find a statistically significant association between any synoptic climatologic category and the incidence of GDV. Their analysis used the principal component analysis (PCA) to group weather-related covariates together and, therefore, may have obscured the influence of individual variables. [15] used a logistic regression type approach to analyze if there is significant association between changes in hourly-measured temperature and/or atmospheric pressure

and the occurrence of GDV in a dog population. The odds of a day being a GDV day given certain temperature and atmospheric pressure conditions for that day or the day before was estimated using logistic regression models. Disease risk was negatively associated with daily maximum temperature; an increased risk of GDV was weakly associated with the occurrence of large hourly drops in temperature that day and with the higher minimum barometric pressure that day and the day before GDV occurrence.

To view possible models in the categorical time series framework, some notation is needed. As a reminder, the GDV occurrence is viewed as binary: 0 signifies no incidence of the disease on a given day while 1 means that a GDV case was observed on that day. The response Y_t is viewed as a binary time series. Let us denote the mean of Y_t $\mu_t = E\,Y_t$. Then, for a given monotone link function g, the systematic component of the model is defined as

$$g(\mu_t) \equiv \eta_t = \sum_{j=1}^{p} \beta_j Z_{t-1,j} \tag{1}$$

where the covariate vector $\mathbf{Z}_{t-1} = (Z_{t-1,1}, \ldots, Z_{t-1,p})^T$ may include among its p components both past values of possible external covariate(s) $X_{t-1}, X_{t-2}, \ldots$ and past values of the time series $Y_{t-1}, Y_{t-2}, \ldots$. Let $\mathcal{F}_{t-1}$ be the σ-algebra generated by $\mathbf{Z}_{t-1}, \mathbf{Z}_{t-2}, \ldots$. Then, the data is assumed to have been generated from the conditional distribution $f(y_t; \theta_t | \mathcal{F}_{t-1})$ that belongs to some exponential family; it is assumed that θ_t is a finite-dimensional parameter. For the Bernoulli distributed Y_t, assuming that π_t is the probability of success given $\mathcal{F}_{t-1}$, the conditional distribution of the data is

$$f(y_t; \theta_t | \mathcal{F}_{t-1}) = \exp\left\{ y_t \log \frac{\pi_t}{1-\pi_t} + \log(1-\pi_t) \right\}$$

where $\theta_t = \log \frac{\pi_t}{1-\pi_t}$ is the natural parameter of the exponential family. If the natural parameter is used as a canonical link -

$$\theta_t(\pi_t) = \log \frac{\pi_t}{1-\pi_t} \equiv \eta_t = \mathbf{Z}_{t-1}^{'} \beta,$$

-the result is the logistic model for the binary time series Y_t. It is possible to use other link functions with binary data, such as logit, probit and complementary log-log. However, the logit, being the canonical link function, seems to be the most natural choice. To stress the dependence of possible GDV occurrence on the explanatory covariates, it can be also defined as $\pi(\mathbf{z}_t)$- a function of the vector $\mathbf{z}_t$ which is a realization of $\mathbf{Z}_t$. Then, under the assumption of the logit link function, the systematic component of the model becomes, as in ordinary logistic regression,

$$\log \frac{\pi(\mathbf{z}_{t-1})}{1-\pi(\mathbf{z}_{t-1})} = \mathbf{z}_{t-1}^{'} \beta. \tag{2}$$

2.2 Modeling of the dog census data

The logit link assumption means that the data being considered are, effectively, proportions of GDV occurrences with respect to the total MWD population on that day. As mentioned before, the dog census data is only available for the last three months of 1993 and then 1994 through 1997. In the spirit of time series approach, the missing data is imputed by fitting an auxiliary time series model to the GDV occurrence data first that is then used to compute probabilities needed for logistic modeling.

To that end, it is assumed that the dog census data can be modeled using a classical structural approach. The census data N_t is viewed as consisting of the trend and an additive random error:

$$N_t = T_t + \varepsilon_t \tag{3}$$

where T_t is a deterministic trend and ε_t is the zero mean error process. It is assumed that the trend is a simple deterministic linear trend $T_t = a + bt$ that is estimated using the regular least squares and then a linear model of the ARIMA class is fit to the residuals.

Remark 1 *According to the traditional approach, the general structural model also contains the additive periodic component S_t. It is not included here due to the limited amount of data available. To illustrate this point, the Figure* (1) *shows the dog census data over time. While it may suggest possible seasonality with the period of about* 20 *months, the total length of the series available is only* 51 *months and this is hardly enough to estimate the possible seasonal component.*

Remark 2 *In theory it is possible to fit a more complicated trend T_t then a linear one suggested above; it is, however, better to start with the more parsimonious model first. If this proves inadequate, it is always possible to try a more complicated functional shape of the trend later*

2.3 Inference and model selection

As a first step, the daily atmospheric pressure characteristics (min, max or mean) and daily air temperature characteristics (again, min, max or mean) are considered as possible covariates. The temperature is not adjusted for humidity. Not just daily, but also hourly weather data is available; however, since the GDV occurrence data is available on the daily basis only, there is no way the hourly weather data can be utilized directly. As mentioned earlier (see also [15]), the lagged values of atmospheric pressure and/or temperature can be viewed as possible GDV etiological factors and thus as additional explanatory variables

as well. Thus, the number of explanatory variables that can be considered becomes quite large and the question of variable selection becomes important. Of the many tests that can be used to verify the statistical significance of the model covariates, the likelihood-ratio type tests are usually the best for a binary time series with a logit link model that is considered here. The older and easier to use Wald test is often less reliable, especially when the values of coefficients are large (see [1]). It is useful to recount the definition of the likelihood-ratio test; to do so requires adopting a simple terminology. Suppose the null hypothesis of interest is $H_0 : \beta_j \equiv 0$, $1 \leq j \leq p$. The model that contains the coefficient β_j, $1 \leq j \leq p$ is called the *unrestricted* and the model without it (t.i., the model with an imposed constraint $\beta_j \equiv 0$) is called *restricted.* The log-likelihood is a function of the coefficient vector β; if the coefficient vectorwith β_j removed is denoted $\tilde{\beta}$, the likelihood ratio statistic is equal to twice the difference between the log-likelihood of unrestricted model $l(\beta)$ and the one for a restricted coefficient vector under the null hypothesis $l(\tilde{\beta})$: $L = 2(l(\beta) - l(\tilde{\beta}))$. As a consequence, it uses twice the amount of information used by the Wald test statistic. Its other advantage is that its form allows easy testing of the composite hypotheses that check the equality of the multiple coefficients to zero. Note that the difference in likelihood ratios can also be used as a goodness-of-fit criterion to compare the performance by a series of nested models. Another possibility is to use one of the classical model selection criteria such as AIC criterion due to [2].

3 Results

As mentioned before in the previous section of the paper, the plot of the population-at-risk against time seems to show a noticeable downward trend that is modeled using a simple linear regression approach; in other words, we assume that $T_t = a + b * t$ within the time range is being considered. The result is illustrated in the Figure (1) where the trend is superimposed on top of the dog census data.

The trend is highly significant with the value of F-statistic about 76.72 on 1 and 49 df. However, residuals are clearly autocorrelated which can be seen from the Figure (2). This means that there is some remaining autocorrelation between the residuals that has to be explained. It is also easy to notice that sample autocorrelation of the residuals decreases in absolute value very slowly while exhibiting periodic behavior (see Figure (3)); this suggests possible lack of stationarity and/or periodicity.

As usual, it is a good idea to begin with the exploratory analysis of the data. It is difficult to say if the data is truly periodical although, based on the values at 20 and 40

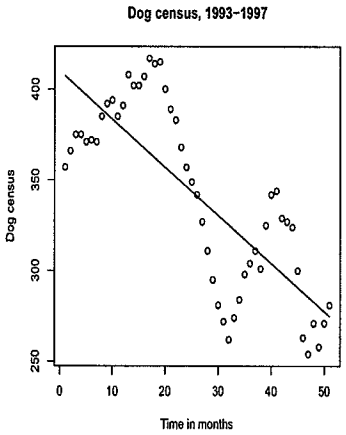

Figure 1:

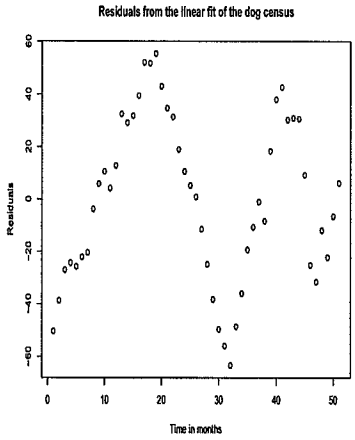

Figure 2:

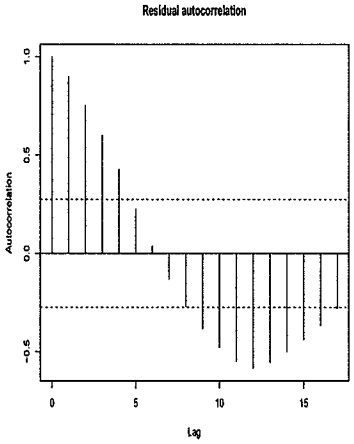

Figure 3:

months after the start of the observation period, one can suspect that the period of about 20 months may be present. There is a dearth of methods that can be used to address this problem given a relatively short length (51 months) of the data available. Of the possible methods, differencing with the lag 20 cannot be used since the series itself is only 51 observations long. Thus, a loss of 20 observations means the loss of more than 1/3 of the sample. Other options, such as smoothing the census data using the seasonal moving average, also result in a huge loss of data. On top of that, to the best of our knowledge, there are no known reasons as to why the dog population should exhibit such a periodic behavior. All of this suggests that it is best not to model the seasonal component in the model for the dog census data. In order to check for stationarity, the first difference of residuals is considered. Let us denote the residual time series $\hat{\varepsilon}_t \doteq N_t - \hat{a} - \hat{b} * t$ where $\hat{a}$ and $\hat{b}$ are the estimates of the linear trend coefficients. Then, the first difference of the residual time series is $\Delta\hat{\varepsilon}_t = \hat{\varepsilon}_t - \hat{\varepsilon}_{t-1}$. The autocorrelation plot of $\Delta\hat{\varepsilon}_t$ shows quickly decreasing autocorrelation pattern that, except the first lag and the lag number 12, stays within the regular 95% confidence interval (see Figure (4)). The presence of somewhat significant twelfth lag may stem from the presence of a seasonal effect that has not been accounted for. For reasons enumerated earlier, the seasonality is not modeled here.

This suggests the possibility of fitting a simple autoregressive model of order 1 (AR(1)). Choosing AR(1) means, effectively, fitting ARIMA (1,1,0) (AutoRegressive Integrated Moving Average) model to the estimated residuals $\hat{\varepsilon}_t$. The autocorrelation function of

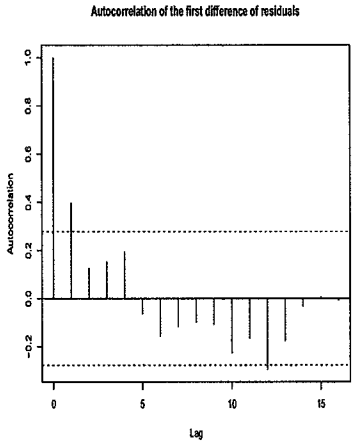

Figure 4:

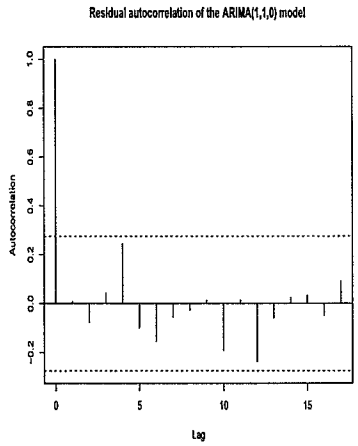

Figure 5:

the residuals of thusly fit model fits entirely (beginning with lag 1) within the standard 95% interval (see Figure (5)). We use this model to predict the missing values of the dog census for the year 1998. The plot that includes both original and imputed values of the dog census is shown in Figure (6). The imputed values are obtained by using the regular

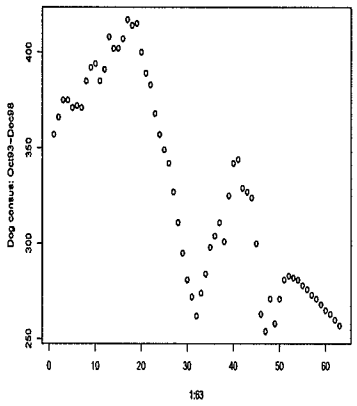

Figure 6: The complete plot of the dog census data

prediction based on the ARIMA(1,1,0) model established above.

The known and imputed values of the dog census put together give the full time series Y_t for the population-at-risk for each month from Oct 1993 through Dec 1998. It is used to define the probability of GDV occurrence π_t on a given day which serves as a response variable in the binary time series model (2).

Let us denote the minimum daily atmospheric pressure on a given day $pmin_t$, maximum daily atmospheric pressure $pmax_t$, maximum daily atmospheric temperature $tmax_t$, minimum daily atmospheric temperature $tmin_t$. Past values (a day before) of the above are $pmin_{t-1}$, $pmax_{t-1}$, $tmax_{t-1}$ and $tmin_{t-1}$, respectively. The maximum daily rise/drop in the atmospheric pressure on the day of GDV event is denoted rp_t and dp_t, respectively while the maximum daily rise/drop in the temperature on the day of GDV event is denoted rt_t and dt_t. These are the covariates to be considered in the analysis. The dogs' breed and weight are not used as covariates since the dog population is rather homogeneous - it consists of three large breeds routinely used as MWD. Also note that sex and age of the dogs that were not diagnosed with GDV during the study period are not known. Because of that, sex and age are not considered as possible covariates as well. All of the dogs are

Table 1: Possible binary time series models

Model	Systematic part
1	$\beta_0 + \beta_1 Y_{t-1} + \beta_2 Y_{t-2} + \beta_3 Y_{t-3} + \beta_4 pmin_t$
2	$\beta_0 + \beta_1 Y_{t-1} + \beta_2 Y_{t-2} + \beta_3 Y_{t-3} + \beta_4 pmin_{t-1}$
3	$\beta_0 + \beta_1 Y_{t-1} + \beta_2 Y_{t-2} + \beta_3 Y_{t-3} + \beta_4 tmax_t$
4	$\beta_0 + \beta_1 Y_{t-1} + \beta_2 Y_{t-2} + \beta_3 Y_{t-3} + \beta_4 tmax_{t-1}$
5	$\beta_0 + \beta_1 Y_{t-1} + \beta_2 Y_{t-2} + \beta_3 Y_{t-3} + \beta_4 tmin_t$
6	$\beta_0 + \beta_1 Y_{t-1} + \beta_2 Y_{t-2} + \beta_3 Y_{t-3} + \beta_4 tmin_{t-1}$
7	$\beta_0 + \beta_1 Y_{t-1} + \beta_2 Y_{t-2} + \beta_3 Y_{t-3} + \beta_4 pmax_t$
8	$\beta_0 + \beta_1 Y_{t-1} + \beta_2 Y_{t-2} + \beta_3 Y_{t-3} + \beta_4 pmax_{t-1}$
9	$\beta_0 + \beta_1 Y_{t-1} + \beta_2 Y_{t-2} + \beta_3 Y_{t-3} + \beta_4 rp_t$
10	$\beta_0 + \beta_1 Y_{t-1} + \beta_2 Y_{t-2} + \beta_3 Y_{t-3} + \beta_4 dp_t$
11	$\beta_0 + \beta_1 Y_{t-1} + \beta_2 Y_{t-2} + \beta_3 Y_{t-3} + \beta_4 rt_t$
12	$\beta_0 + \beta_1 Y_{t-1} + \beta_2 Y_{t-2} + \beta_3 Y_{t-3} + \beta_4 dt_t$

fed a standard diet, housed in outdoor runs and are under the observation 24 hrs a day. The following candidate models are considered:

We do not include more than one temperature-or pressure-related variable in a model because of their strong correlation. Recall that our main goal is identification of the possible etiological factors of GDV; thus, the presence of many strongly correlated predictors would clash with that goal. As an example, it is easy to realize that all of the temperature- and atmospheric pressure- related covariates are correlated; as a rule, both change simultaneously in response to passing atmospheric fronts. To illustrate this fact, we attach a table of Pearson correlations between temperature- and pressure- related variables on the day of GDV event, together with 95% confidence intervals. Except the correlation between the minimum daily pressure $pmin_t$ and the minimum daily temperature $tmin_t$, all the other pairwise correlations have a 95% confidence interval whose both ends are larger in absolute value than 0.5; this suggests, of course, that with probability 95% almost all of the pairwise correlations exceed 50%. It goes without saying that the values of the air temperature and atmospheric pressure on the GDV day and 1- or 2 - days before are also strongly correlated. With this in mind, no more than one temperature- or pressure-related covariate is included in any model.

Table 2: Covariate Correlation Table

	$pmax_t$	$pmin_t$	$tmax_t$	$tmin_t$
$pmax_t$	1.000	0.882 (0.871, 0.891)	-0.622 (-0.649, -0.594)	-0.606(-0.633, -0.577)
$pmin_t$		1.000	-0.517 (-0.549, -0.484)	-0.462(-0.497, -0.426)
$tmax_t$			1.000	0.781(0.762, 0.797)
$tmin_t$				1.000

In the absence of any information suggesting that climatologic events that occurred a long time ago influence GDV occurrence, it seems prudent to limit the number of daily lags of Y_t included in the model; this research limits them to 3. This also helps to reduce the dimensionality of the problem. As is the case in classical time series, the residuals of each model can later be analyzed for autocorrelation patterns; if a significant amount of unexplained autocorrelation still remains, additional lags can be added.

It turns out that for each one of the above models, the values of Y_{t-1} and Y_{t-3} have very low values of the loglikelihood ratio test statistic. The computations show that for all of the models 1-12 the loglikelihood ratio statistic value for the coefficient of Y_{t-1} is 0.06 (the p-value is 0.806) while the loglikelihood ratio for Y_{t-3} is 0.04 (the p-value is 0.841). This suggests that both of these predictors should be removed from all models; we only retain Y_{t-2} in all of them. The revised models are given in the Table (3).

All of the 12 models presented in the Table (3) are now fit using the iterative reweighted least squares algorithm commonly applied to fitting of generalized linear models. Given that the canonical link function log is used for the binary data, the iterative reweighted least squares algorithm coincides with the regular Newton-Raphson algorithm in this case. Each of these models has 2 parameters and 1914 degrees of freedom. Given that the logistic regression considered here has the binary response, it is not possible to use the classical Pearson X^2 goodness-of-fit statistic since the χ^2 approximation does not work even for large sample sizes. Among many possible choices, the le Cessie and van Houwelingen test (see [5]) based on smoothed standardized residuals is probably the most commonly used today. If the total number of observations is N, the probability of GDV event at time t is π_t and the fitted probability is $\hat{\pi}_t$, denote the standardized residual $\hat{r}_t = \frac{y_t - \pi_t}{\sqrt{\hat{\pi}_t(1-\hat{\pi}_t)}}$, $t = 1, \ldots, N$. As a next step, a set of weights $\{w_{it}\}$ is chosen and the

Table 3: Revised Models

Model	Systematic part
1	$\beta_0 + \beta_1 Y_{t-2} + \beta_2 pmin_t$
2	$\beta_0 + \beta_1 Y_{t-2} + \beta_2 pmin_{t-1}$
3	$\beta_0 + \beta_1 Y_{t-2} + \beta_2 tmax_t$
4	$\beta_0 + \beta_1 Y_{t-2} + \beta_2 tmax_{t-1}$
5	$\beta_0 + \beta_1 Y_{t-2} + \beta_2 tmin_t$
6	$\beta_0 + \beta_1 Y_{t-2} + \beta_2 tmin_{t-1}$
7	$\beta_0 + \beta_1 Y_{t-2} + \beta_2 pmax_t$
8	$\beta_0 + \beta_1 Y_{t-2} + \beta_2 pmax_{t-1}$
9	$\beta_0 + \beta_1 Y_{t-2} + \beta_2 rp_t$
10	$\beta_0 + \beta_1 Y_{t-2} + \beta_2 dp_t$
11	$\beta_0 + \beta_1 Y_{t-2} + \beta_2 rt_t$
12	$\beta_0 + \beta_1 Y_{t-2} + \beta_2 dt_t$

smoothed residuals are defined as $\hat{r}_{si} = \sum_{t=1}^{N} w_{it}\hat{r}_t$ are obtained where s in the subscript stands for "smoothed." The weights w_{it} can be generated by a variety of kernel functions. The original test of le Cessie and van Houwelingen uses the uniform kernel and this is the one that is used in this research as well. The test statistic itself is defined as the normalized sum of squared smoothed residuals

$$T = \sum_{i=1}^{N} \frac{\hat{r}_{si}^2}{\hat{V}(\hat{r}_{si}^2)}.$$

where $\hat{V}$ in the denominator denotes the estimated variance of the squared smoothed residual $\hat{r}_{si}^2$. For more details about the test, see [5, 11]. For each model, the Table (4) contains the p-value for the le Cessie and van Houwelingen's test (CH for short) and a value of AIC (Akaike information criterion).

Note that all of the p-values of the smoothed residuals test are in excess of 0.1 which suggests that the null hypothesis cannot be rejected at a reasonable level of significance for any of the 12 models; thus, all of them seem to be adequate in explaining the data. It is, of course, a matter of considerable interest to identify models where the coefficients of the external covariates (temperature- and pressure-related variables) are significant. The significance is measured using both Wald and loglikelihood-ratio p-values. Only

Table 4: Model Selection

Model	CH	AIC
1	0.1377	516.89
2	0.8650	517.46
3	0.6272	519.33
4	0.6272	519.57
5	0.6272	520.37
6	0.6272	518.63
7	0.3218	519.53
8	0.9557	517.02
9	0.3489	520.13
10	0.2658	520.26
11	0.4595	522.00
12	0.4595	521.64

the models where the log-likelihood ratio p-values do not exceed 0.15 for both external covariate and the response lag Y_{t-2} are shown in the Table (5). The Table (5) contains the values of odds ratios for the lagged response Y_{t-2} and the external covariate together with their loglikelihood- and Wald- p-values. The cutoff 0.15 is chosen in order not to make an abrupt break at at the standard threshold 0.1, thus somewhat artificially separating the model with p-values of a coefficient being, for example, 0.95 and 0.105.

In order to confirm that the 6 models shown in the Table (5) are adequate, it is useful to check their raw residual autocorrelation plots. For reasons of brevity, only those for models 1, 2 and 8 are shown in The Figures (7), (8) and (9), respectively. The results for the other three models are the same.

The raw residuals autocorrelation plots are nearly identical and neither shows any substantial autocorrelation remaining. Thus, all of the models given in Table(5) can be assumed to be adequate.

4 Discussion and Conclusions

A study of the GDV data set originally reported in [10] and later used in [15] has been conducted in order to provide a more general view of the GDV occurrence process. The

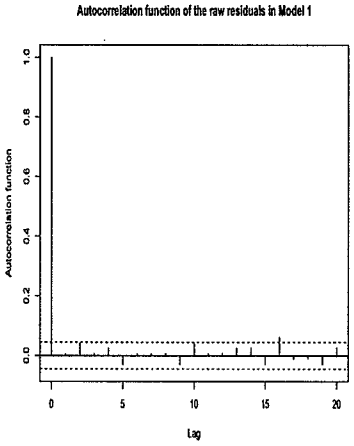

Figure 7:

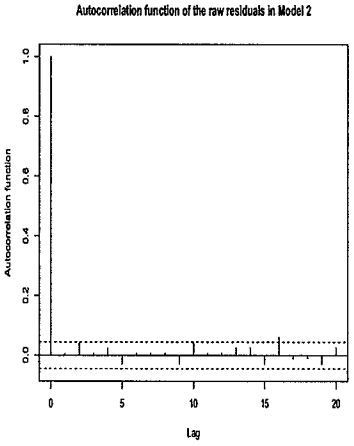

Figure 8:

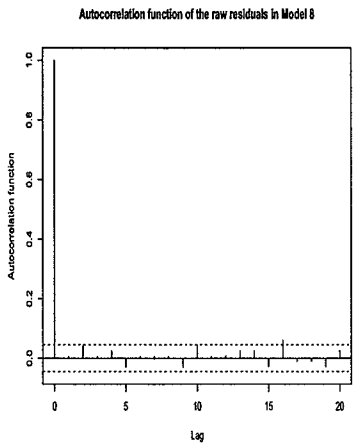

Figure 9:

Table 5: Model Selection

Model	OR of Y_{t-2}	Wald p-value	LL p-value	Ext. Covariate OR	Wald p-value	LL p-value
1	2.397556	0.0919	0.1182762	1.053660	0.0204	0.02298022
2	2.437079	0.08333	0.1182762	1.050609	0.02886	0.03197196
8	2.369692	0.09648	0.1182762	1.049160	0.02093	0.02476849
6	2.452964	0.0836	0.1182762	0.9848775	0.0576	0.0640221
3	2.43702	0.0859	0.1182762	0.9846333	0.0934	0.09847928
4	2.430174	0.087	0.1182762	0.985313	0.109	0.1145717

GDV occurrence is viewed as a binary time series process. It has been fit as a logistic regression model with covariates that include both extraneous climatological variables and the lags of the response process Y_t. This view of the data can be supported by noticing that identifying all of the possible etiological factors of GDV is difficult [18, 16] and thus there is an uncertainty over which ones of them should be included in the model. It is likely that unknown omitted factors are, in fact, climatological variables that are correlated over time; their omission induces temporal autocorrelation in the response.

A number of variables were identified as factors significantly influencing the rate of GDV occurrence. They are the maximum temperature on the day of GDV and the day before, minimum temperature on the day before the GDV, minimum daily pressure on the day of GDV and the day before as well as the maximum daily pressure on the day before GDV. Out of these, maximum temperature on the day of GDV and the day before and minimum daily pressure on the day of GDV and the day before were also identified as important and statistically significant factors in [15].

Unlike [15], we take into account the temporal correlation between the GDV events; as a result, fairly significant correlation between the current likelihood of GDV and that 2 days before the observation day is detected. More specifically, the coefficient of the lagged response Y_{t-2} translates into the odds ratio that is consistently above 2 for all of the models considered. In all cases, the respective Wald p-values of Y_{t-2} coefficients are all below 0.1 and the loglikelihood p-values are close to 0.12. This indicates that possible case of GDV 2 days before the current observation signifies a much higher probability of GDV occurring on the day of observation. This suggests that all of the results for the external covariates should be interpreted conditionally on a value of y_{t-2}.

All of the odds ratios of the pressure-related covariates are slightly greater than 1; this indicates positive association between these factors and the probability of the GDV event on a given day. For example, the minimum daily atmospheric pressure on the day of GDV event has an odds ratio of 1.053660 conditionally on the known value of Y_{t-2}. This means that for each change of the minimum daily atmospheric pressure by 1 unit, the odds of the GDV case occurring on that day increase by the factor of 1.053660 for a given outcome 2 days before the observation day.

The situation is reversed with the temperature-related covariates. In the case of models 3, 4 and 6, when the covariates are the maximum temperature on the day of GDV, the maximum temperature on the day before GDV and the minimum temperature on the day before the GDV, respectively, all of them have the odds ratios that are below 1. This means that the increasing temperature reduces risks of GDV. All of these 3 odds ratios are

numerically close to each other. As an example, observe that the maximum temperature on the day of GDV has the odds ratio of 0.9846333. This means that with the increase in one degree of the maximum temperature on a given day the odds of GDV on that day decrease by a factor 0.9846333 for a given outcome 2 days before the observation day.

The approach using the logistic model for binary time series used in this research appears to be adequate in the case where there are at most a few daily observations with more than 1 GDV case. In that case those days can be simply ignored. However, this may not be the case if a larger group of dogs is observed and, as a consequence, the number of days with more than one case of GDV becomes sizable. If this happens, instead of considering the binary time series setting, it is necessary to consider a categorical time series model where the possible range of responses Y_t is $0, 1, 2, \ldots, k$ for some positive $k > 1$. Given that these are ordinal data, a suitable model is, probably, the cumulative odds model applied to the categorical time series. The authors intend to obtain a more extensive data in order to pursue this research direction.

References

[1] Agresti, A. *Categorical Data Analysis* Wiley, 2nd edition, 2002

[2] Akaike, H. A new look at the statistical model identification *IEEE Transactions on Automatic Control*, 1974, 19 (6): 716723.

[3] Archer, D.C., Pinchbeck, G.L., Proudman, C.J., Clough, H.E. Is equine colic seasonal? Novel application of a model based approach *BMC Veterinary Research*, 2006, 2:27

[4] Brockman, D.J., Holt, D.E., Washabau, R.J. Pathogenesis of acute canine gastric dilatation-volvulus syndrome: Is there a unifying hypothesis? *Compend Contin Educ Pract Vet*, 2000, 22:11081113

[5] le Cessie, S., and van Houwelingen, J.C. Goodness-of-Fit Test for Binary Regression Models, Based on Smoothing Methods *Biometrics*, 1991, 47(4):1267-1282

[6] Dennler, R., Koch, D., Hassig, M., Howard, J., Montavon, P. M. Climatic conditions as a risk factor in canine gastric dilatation-volvulus *The Veterinary Journal*, 2005, 169(1):97-101

[7] Fokianos, K. and Kedem, B. Regression theory for categorical time series *Statistical Science* 2003 18(3):357-376

[8] Godshall D., Mosallam, U., Rosenbaum, R. Gastric volvulus: case report and review of the literature *J. Emerg. Med.*, 1999, 17(5):837-840

[9] Guttorp, P. *Stochastic Modelling of Scientific Data* Chapman& Hall, London, 1995

[10] Herbold, J.R., Moore, G.E., Gosch, T.L., Bell, B.S. Relationship between incidence of gastric dilatation-volvulus and biometeorology events in a population of military working dogs *American Journal of Veterinary Research*, 2001, 3:47-52

[11] Hosmer. D.W., Hosmer, T., le Cessie, S., and Lemeshow, S. A comparison of goodness-of-fit tests for the logistic regression model *Statistics in Medicine*, 1997, 16(8):965-980

[12] Kedem, B. and Fokianos, K. *Regression Models for Time Series Analysis*, Wiley, 2002

[13] Mayo A, Erez I, Lazar L, Rathaus V, Konen O, Freud E. Volvulus of the stomach in childhood: the spectrum of the disease *Pediatr. Emerg. Care* 2001, 17(5): 344-348

[14] McCullagh, P. and Nelder, J.A. *Generalized Linear Models*, Chapman& Hall, 2nd edition, 1989

[15] Moore, G.E., Levine, M., Anderson, J.D. and Trapp, R.J. Meteorological influence on the occurrence of gastric dilatation-volvulus in military working dogs in Texas, *International Journal of Biometeorology*, 2008,52(3), 219-22

[16] Raghavan, M., Glickman, N., McCabe G., Lantz G., Glickman L. Diet-related risk factors for gastric dilatation-volvulus in dogs of high-risk breeds, *J. Am. Anim. Hosp. Assoc.*, 2004, 40:192203

[17] Enscore, R.E., Biggerstaff, B.J., Brown, T.L., Fulgham, R.F., Reynolds, P.J., Engelthaler, D.M., Levy, C.E., Parmenter, R.R., Montenieri, J.A., Cheek, J.E., Grinnell, R.K., Ettestad, P.J., and Gage, K.L. Modeling relationships between climate and the frequency of human plague cases in the southwestern united states, 1960-1997 *Am.J. Trop. Med. Hyg.*, 66(2):186-196

[18] Schellenberg, D., Qilong Y., Glickman N., Glickman L. Influence of thoracic conformation and genetics on the risk of gastric dilatation-volvulus in Irish setters, *J. Am. Anim. Hosp. Assoc.*, 1998, 34:64-73

[19] Sevcik, WI, Steiner, IP. Acute gastric volvulus: case report and review of the literature *CJEM* 1999; 1(3):200-203

[20] Upadhyaya VD, Gahgopadhyay AN, Pandey A, Kumar V, Sharma SP, Gupdat DK Acute gastric volvulus in neonates - a diagnostic dilemma *Eur. J. Pediatr. Surg.* 2008, 18(3): 188-191

[21] Van Kruiningen, H.J., Gregoire, K., Mieuten, D.J. "Acute gastric dilatation: a review of comparative aspects by species, and a study in dogs and monkeys", *J. Am. Anim. Hosp. Assoc.*, 1974, 10:294339