Johns Hopkins University, Dept. of Biostatistics Working Papers

9-1-2004

On Time Series Analysis of Public Health and Biomedical Data

Scott L. Zeger
The Johns Hopkins Bloomberg School of Public Health, szeger@jhsph.edu

Rafael A. Irizarry
Johns Hopkins Bloomberg School of Public Health, Department of Biostatistics, rafa@jhu.edu

Roger D. Peng
Johns Hopkins Bloomberg School of Public Health, Department of Biostatistics, rpeng@jhsph.edu

On Time Series Analysis of Public Health and Biomedical Data

Scott L. Zeger, Rafael Irizarry and Roger D. Peng

Department of Biostatistics
The Johns Hopkins University Bloomberg School of Public Health
Email: szeger@jhsph.edu, ririzarr@jhsph.edu, rpeng@jhsph.edu

Running head: On Time Series Analysis of Public Health

Key words: stochastic process; smoothing; autocorrelation; periodogram; spectrum; regression; autoregressive model; ARMA; non-linear time series

Abstract: A time series is a sequence of observations made over time. Examples in public health include daily ozone concentrations, weekly admissions to an emergency department or annual expenditures on health care in the United States. Time series models are used to describe the dependence of the response at each time on predictor variables including covariates and possibly previous values in the series. Time series methods are necessary to account for the correlation among repeated responses over time. This paper gives an overview of time series ideas and methods used in public health research.

Corresponding author: Scott L. Zeger
615 N. Wolfe Street
Baltimore MD 21205
szeger@jhsph.edu
ph: 410 955 3067
fax: 410 955 0958

Contents

Introduction

A time series is a sequence of measurements equally-space through time or along some other metameter. The numbers of pregnancies each week in the JiVitA Project study in Bangladesh are shown in Figure 1 as an illustration.

The study goal is to reduce maternal and infant mortality through population-based randomized trial of vitamin A or beta-carotene supplementation (Joanne Katz, personal communications). Nearly 45,000 pregnant women have been randomized to receive different supplements. They and their babies are followed to observe their health outcomes.

The pregnancy time series in Figure 1 has several interesting features. Most apparent is the seasonality in the numbers of pregnancies that tend to be higher in the early summer than winter months. We can also see some evidence that the weekly number of pregnancies is decreasing over time. Finally, on closer examination, it appears that the number of pregnancies in a week may be negatively associated with the number the few weeks before. That is, if there were more than expected pregnancies in a given week, fewer may occur in the succeeding weeks. This is another way of saying that nearby observations tend to be correlated in a time series. This correlation can be of interest, for example, when we are trying to predict the near future of the series or it can be a nuisance as it would be if we were trying to determine whether the downward trend in Figure 1 is likely to be a chance event.

Figure 2 plots the correlation coefficient between the number of pregnancies for one week and the number *u* weeks later against the lag *u* from 1 to 20.

As detailed below, this *autocorrelation function* (ACF) demonstrates that the observations in the time series can not be assumed to be independent as is done in most standard statistical analyses. The autocorrelation requires that special time series methods be used instead.

The goals of time series analysis include simple *description, explanation, prediction* or *control* (9). The plots above are examples of *descriptive* analysis used to uncover patterns of potential scientific import. We typically use regression analysis to *explain* the dependence of a response time series Y_t on predictor series $X_{1t}, \ldots, X_{pt}$ while taking appropriate account of the lack of independence among the time series observations. In the JiVitA example, we might use simple functions of time to explain the seasonality in the pregnancy data and to ask whether there is evidence of a systematic downward trend. If we are *predicting* a future response Y_{n+u} using an observed series, we can regress the response at a given time on preceding responses and possibly also on covariates that are known into the future such as week in the example above. Finally, when the goal is to *control* a process, we use time series analysis to determine whether the process is systematically deviating from expectation and to identify changes to the process that will bring it back to the desired trajectory. In our example, we might be interested to know whether the downward trend is systematic and if so, how it might be reversed. Process control is particularly important in industrial applications and will not be considered here further.

The interested reader who is familiar with regression analysis will find the book by Diggle (13) an excellent introduction to time series analysis for biomedical and public health research. For technical details about time series models and forecasting, the classic

text by Box and Jenkins (4) is recommended. A technical treatment of time series theory is given by Brockwell and Davis (5).

This paper provides an overview of key ideas and methods for time series analysis of public health and biomedical data. We illustrate the problems and methods with basic analyses of the pregnancy time series and with reference to articles from the public health literature. We begin with a review of some ideas that underpin the statistical reasoning and methods for analyzing time series data. The next two sections discuss descriptive and explanatory methods in turn. We then summarize some of the recent public health and biomedical literature that uses times series methods and finally present two recent examples from our time series research that extend the simple methods in novel directions.

Time Series as a Single Observation of a Stochastic Process

Probability theory underlies statistical reasoning and practice. A central concept is the *random variable* Y that represents the outcome of an experiment or observational process. The adjective *random* implies that the variable can take different values, each with its own probability. The *probability distribution* P(y) for the random variable Y is just a listing of the probability the variable takes each of its possible values y. That is, $P(y) = \Pr(Y=y)$. For example, if Y is the number of live pregnancies in the first week of 2004 in the JiVitA project, then Y can take any value from among the non-negative integers: 0, 1, 2, … Here, P(150) is the chance that exactly 150 babies will be born that week.

In thinking about random variables, it is helpful to think of an urn with a large number of beads. Each bead has a value y written on it. When we do an experiment or make an observation, we choose a bead at random from the urn. The probability P(y) is then just the fraction of beads that have the particular value y written on them.

A *stochastic process* $\{Y_0, Y_1, Y_2, \ldots\}$ is a possibly infinite sequence of random variables ordered in time (39, 35). A time series is a single realization of a stochastic process. In the urn model, instead of a single number written on each bead, imagine a whole time series written there. With a random variable, we are interested in the probability the variable takes each of its possible values. With a stochastic process, we can ask the same question about the random variable at every time or about combinations of them. For example, we might be interested in the probability that Y_7 is greater than 200 or the joint probability that none of $Y_0, Y_1, \ldots, Y_{10}$ exceeds 300.

We typically obtain multiple independent observations on a random variable. For example, in National Medical Expenditure Survey (34), the annual medical expenditures for roughly 30,000 persons were sampled. Johnson et al. (27) and many others have used these data to estimate the mean expenditure for the population or for persons who suffer from a major smoking-caused disease.

In contrast, a time series is a single observation on the stochastic process. We want to make an inference about the properties of the underlying stochastic process from a single realization, a single observation at each time. With longitudinal data, we observe a short time series for each of a large number of subjects (14).

For example, in the pregnancies example, there is only one series to contemplate. Then how can we talk about the probability that the number of pregnancies in week 7 exceeds 200? It either did or did not.

There are two related concepts that make possible the use of probability in time series analysis. The first is the assumption of *stationarity* (35). A stochastic process is *stationary* if the probability distribution of the random variables at a set of times (t_1, t_2, …, t_n) is the same as the distribution of the variables at the times ($t_1+\tau$, $t_2+\tau$,, …, $t_n+\tau$), that is, if the times are all shifted by the same amount τ. Stationarity says that the probability distribution of any set of n variables depends only on their relative, not absolute times. If we let n=1, the stationarity implies that the individual observations all have the same univariate distribution, in particular, same mean and variance. Under this assumption, we can use the replication over time in a time series to make inferences about the common mean, variance and other statistics. The ability to make valid probability statements by looking over time rather than across replicates at one time is called *ergodicity* (35), the second, closely-related concept.

A key consequence of the assumption of stationarity is that the degree of dependence between two random variables in the stochastic process decreases as the time interval between them increases. That is, two observations far enough apart in time are essentially independent. This implies that by following a process longer in time, new independent information will be accumulated.

Longitudinal data comprise many, usually shorter, time series. There is a very important distinction between time series and longitudinal data analysis (e.g. 14). With a single sequence, we must rely on the assumption that observations far enough apart in time are approximately independent. When this is true, the amount of information about a parameter increases proportional to the number of observations.

With longitudinal data, we assume that the short time series for individuals are independent. We do need to make the stronger assumption that the correlation between repeated observations on the same person dies away to 0.0 with increasing time separation. The key sample size is the number of people, not observations for each person. The analysis of longitudinal data is closer to multivariate analysis, in this sense.

For many regression problems, one can think of creating longitudinal data from one long time series by breaking the series into shorter blocks that are long enough so that most of the correlation exists within blocks and observations from different blocks are approximately uncorrelated. Then inferences can be made robust by borrowing strength across independent blocks rather than relying on an assumed autocorrelation structure within blocks. This is the idea behind bootstrapping of time series. See for example Li & Madalla (28).

Why Time Series Analysis

The most important assumption is standard regression analysis is that the multiple observations are independent of one another. For example, in a study of smoking as a risk factor (X) for cardiovascular disease (Y), we assume that the presence of absence of disease for one participant is independent of this response for every other participant. The idea is that each observation provides new evidence about the association of Y with X. We quantify the amount of evidence when we calculate standard errors, confidence intervals or hypothesis tests for regression coefficients. If the data are not independent, the usual inferences are not correct.

With time series data, we expect that neighboring observations are correlated with one another, so each observation does not provide entirely independent information and the usual regression formulae for standard errors and other inferences are not valid.

To illustrate, consider the question of whether the pregnancy process underlying the time series in Figure 1 has a decreasing mean. We address this question through the simple model $Y_t = \beta_0 + \beta_1 t + \varepsilon_t$, $t = 0, 1, \ldots, n$ where β_1 is the rate of decrease in the expected number of pregnancies per week. If $\beta_1=0$, there is no trend. We will assume for the moment that the obvious seasonality in the series is the result of autocorrelation in the ε_t, not a systematic component as is actually more realistic given our understanding of seasonal influences on pregnancies in Bangladesh.

Suppose we use ordinary least squares to regress pregnancy counts on time, thereby ignoring the correlation among the residuals. The estimate of β_1 is -0.43 with reported standard error 0.15. Hence, there appears to be strong evidence of a downward trend. While the slope estimate from least squares is unbiased in the presence of autocorrelated errors, the estimate of its standard error is incorrect. A valid estimate can be obtained if we assume that the residuals are a realization of a stationary stochastic process. In this case, we obtain a correct estimate of 0.51, more than 3 times as large. Using the correct standard error, we reach a different conclusion about the trend. Given that the series ε_t is clearly autocorrelated, we can not conclude that there is a real downward trend.

So the first reason to use time series methods is to obtain valid inferences. The second reason is to obtain efficient estimates and inferences. The least squares estimator of the slope is the unbiased estimator with the smallest possible variance when the residuals are uncorrelated and have equal variance over time. In the presence of correlation, there is another linear estimator with smaller variance. If we assume the estimated autocorrelation function is the true one for this problem, then the optimal estimate of the trend is -0.42 with standard error 0.45. This standard error is 0.06 smaller than the standard error (0.51) for the least squares slope. Hence, by using optimal time series methods, we can reduce the standard error of the slope by 10% percent. This is like having roughly 30 $[=(.51/.45)^2 * 149]$ more observations. In summary, using time series methods produces valid and more efficient inferences.

Trend or Autocorrelation; Fixed or Random Variation

The need to assume stationarity does not prevent us from modeling processes that change over time. As illustrated above, we can decompose the stochastic process $Y_t = S_t(\beta) + \varepsilon_t$ where S_t is assumed to be a systematic change in the level of the process over time that can be represented by a small number of parameters β and ε_t is a stochastic process of deviations about the trend that we more reasonably assume form a stationary process.

In the regression above, we decomposed the time series into a linear trend and residual series that we assume is stationary. But there is also obvious seasonality in the residuals. Trends and seasonal fluctuations can arise from either systematic or random events. Whether we treat these variations as part of the systematic or random component depends on our beliefs about the underlying process during the period of observation and beyond.

Figure 3 shows four monthly data sets simulated from the simplest time series model $Y_t = \rho Y_{t-\tau} + a_t$, $t=1, 2,\ldots,n$ where the a_t are independent and identically

distributed normal variates with mean 0 and common variance. The four series correspond to the values of $\rho = 0.5$ and 0.99; $\tau = 1$ and 12.

Note in panel (a) ($\rho = 0.5$; $\tau = 1$), the series looks stationary because the autocorrelation is modest. In panel (b) ($\rho = 0.99$; $\tau = 1$), the correlation at lag one is near 1.0 and the series meanders with trends that persist for extended periods. Panels (c) and (d) demonstrate that seasonality can appear if the current monthly value depends on the one 12 observations ago.

Whether a longer-term trend represents the "meandering" of a stationary process that will eventually reverse itself or whether it represents a trend that will persist can usually not be clearly established from a finite record, alone. This is the core of the argument about global warming (26). The preferred model must be determined by external evidence about the mechanisms that give rise to the observed trends.

Descriptive Time Series Analysis

In *descriptive analysis*, the objective is to create data displays and summary statistics that lead to a better understanding of the variation in the response over time. For example, we want to understand factors that influence the pregnancy rate in the JiVitA Project population. The time series plot shown in Figure 1 is the simplest and often most effective tool. It makes apparent a downward trend, seasonality and possibly autocorrelated residuals.

A time series plot can be enhanced by adding a *smooth curve* that highlights trends amidst the variation. Inherent in smoothing is a decomposition: $Y_t = S_t + R_t$ where S and R are the smooth and rough parts of the series, respectively. The simplest smoothing methods are the running mean and running median of length 2m+1. Here, the smooth value at time t is just the average (mean or median) of the Y_ts from t-m up to t+m. The degree of smoothness of the S_t that is produced increases as m increases. The averaging interval 2m+1 is called the *bandwidth* of the smoother. Running medians are insensitive to outliers but tend to produce undesirable discrete jumps in the resulting smooth curve. An iterated combination of running medians, then means produces the resistance to outliers of the median and the smoothness of the mean (45). Running means are a special case of *kernel estimators* of S_t obtained by taking weighted averages of neighboring observations. Usually, we choose weights that decrease as we move away from t. Other popular methods of smoothing include *smoothing splines* and *wavelets*. Hastie et al. (23) present an overview of smoothing methods.

In many problems where there are different mechanisms operating at different time scales, it is natural to decompose a time series into more than two components, that is we let $Y_t = S_{1t} + S_{2t} + \ldots + S_{pt} + R_t$. In the pregnancy time series, it makes more sense to assume the pregnancy process is the sum of three terms: a long-term trend that reflects the changing population, seasonality that results from the demands of an agrarian society and the effects of weather and residuals that can be assumed to be stationary (11). Figure 4 shows this decomposition. If we calculate the autocorrelation of the residuals after removing the long-term trend and seasonality, we discover a statistically significant negative autocorrelation at a lag of 3 weeks. This discovery, not readily apparent in the original series, is discussed further below.

In studies of the association of daily mortality with air pollution (e.g. 2), the daily time series can usefully be decomposed into annual trends, seasonality, monthly

variation, weekly variation and daily variation. One motivation for this decomposition is that air pollution and confounders have different influences at the different time scales. For example, annual trends are influenced by changing: medical practice, population size, smoking rates and other slowly varying factors. The seasonal component reflects the effects of infectious disease epidemics and weather. The daily variation includes the effects of acute but not chronic exposure to air pollution. This decomposition has been used to estimate a pollution-mortality association that is less sensitive to confounding and robust to "harvesting" where only the very frail are affected by pollution (49).

It is possible and often desirable to decompose a time series into nearly as many component series as there are observations. The *discrete Fourier transform* (DFT) (3) and *discrete wavelet transform* (7) are examples. To obtain a DFT, a time series of length n (here assumed odd) is re-expressed exactly as its mean plus the sum of (n-1)/2 cosine waves with frequencies 1, 2, …, (n-1)/2 cycles in the length of the series, each cosine having arbitrary amplitude and phase. The *periodogram* is the squared amplitude of each cosine plotted against frequency as shown in Figure 5 for the pregnancy data. The j^{th} value is the amount of variation that can be explained by regressing the time series on a cosine wave that completes exactly j cycles in the length of the data.

The *spectrum* of the underlying stochastic process is defined as the expected value of the squared amplitude. It is estimated poorly by the periodogram each of whose values is approximately proportional to the spectrum times an independent χ^2 random variable with 2 degrees of freedom. Because the χ^2 component is so noisy and because the true spectrum for a stationary process is likely smooth, a better estimate of the spectrum is obtained by using a running mean of the periodogram ordinates (3). The spectrum is the Fourier transform of the true autocorrelation function for the process.

In Figure 5, the utility of the periodogram and estimated spectrum is apparent. Notice the relatively higher values at the lowest frequencies (longest periods), in particular at the annual frequency (1 cycle per each year of data). These large squared amplitudes reflect the trend and the seasonality in the data set. However, there are also smaller peaks corresponding to periods of roughly five and two and a half weeks. The first peak reflects some near-monthly process in the data. The 2.5 week peak is likely its first harmonic indicating that its shape is not exactly like a single cosine. These spectral peaks raise the question: is there a mechanism in this population that is producing a five-week cycle in the numbers of pregnancies? The investigators have an explanation. The data collection system divides women into 5 administrative groups for ascertaining pregnancies. Each group is visited once per 5 weeks. The numbers of women in the groups are not exactly equal. Hence more births are discovered in some batches than others producing this cycling in the total pregnancies. While this finding is not of scientific interest, it does demonstrate the usefulness of decomposing the data into components and of frequency domain methods for uncovering patterns of potential interest.

Frequency domain methods like these are commonly used in biomedical and public health research. Recent papers illustrate applications to the study circadian rhythms (e.g. 38), brain function (e.g. 36) and heart performance (e.g. 47).

Overview of Time Series Regression Models

When the objective of statistical analysis is *explanation*, we model the dependence of a response process Y_t on one or more predictor time series X_t that are assumed to be fixed, not random. Explanation is done using *regression analysis*. A key assumption of the standard regression model is that the responses are independent of one another after adjusting for predictor variables. With time series data, neighboring values of Y tend to be correlated. The correlation must be taken into account to make valid inferences about the parameters of the process.

With a time series rather than a scalar response, we can predict Y_t using the past responses Y_{t-1}, $Y_{t-2},\ldots,Y_1$ alone or in combination with covariate series X_t. *Autoregressive models* (AR) are an example of the former. *Distributed lags models* are an example of the latter. AR models are discussed in more detail below. Distributed lags models are used in studies of air pollution and mortality. See the recent paper by Bell et al. (2) for examples and further references.

Public health research gives rise to discrete as often as continuous responses. For example, in an intervention study to reduce smoking, the response might be a time series of daily binary indicators of whether a participant smoked or not on that day. Or, it might be the number of cigarettes smoked each day, a counted response.

With a scalar response for each subject, generalized linear models (GLMs; 31) have unified regression analysis for binary, count and continuous outcomes. Linear, logistic, probit, log-linear and inverse regression models are special cases of GLMs. Most of the time series ideas and methods discussed here for contiinous outcomes have extensions to the GLM family. Below, we briefly illustrate how time series models can be formulated for discrete as well as continuous responses.

In some problems, the response Y_t is not only influenced by the predictor series X_t, but it also influences X_{t+1} and/or other future values of the predictor series. For example, suppose we are interested in the effect of vitamin A deficiency (X) on the risk of acute respiratory infection (ARI) (Y). The question is whether vitamin A deficient children are at increased risk of ARI? But in this case, scientists believe that ARI may deplete the stores of vitamin A in the liver and cause deficiency. So there is a feedback loop whereby poor nutritional status gives rise to a greater risk of infection that further depletes the stores of micro-nutrients producing more infection, poorer health and possibly death (41). With feedback like this, a full understanding of the relationship of the two series requires their joint modeling, not only a regression of Y on X. One way to build a joint model is in two parts: a regression of Y_t on X_t; and a regression of X_t on Y_{t-1} and other variables from the past of the Y process. Work on multivariate time series is summarized by Reinsel (37). In the context of many short time series, also see the recent book by van der Laan & Robins and references therein (46).

Strategies for Regression Models with Time Series Data

There are two primary types of time series models: *marginal models*; and *conditional models* given past observations. In both, we condition on the history of the X_t series so we suppress the conditioning on the predictor variables to simplify the notation below.

A marginal model has as its primary target, the expectation of Y_t (given the covariate time series). We do not condition on the past responses Y_{t-1}, $Y_{t-2},\ldots,Y_1$. Because the time series of observations are autocorrelated, we must simultaneously model the correlation between all pairs of observations. Hence, the model has two parts:

the usual regression model for the mean as is done for independent observations; and a second model for the autocorrelation function.

A simple marginal linear model is given by:

$$E(Y_t) = X_t\beta$$
$$Cov(Y_t, Y_{t-u}) = \sigma^2\rho^u. \quad (1.)$$

The model for the mean is a standard regression analysis. Its parameters, β, have the same interpretation as in ordinary linear regression analysis. The model for the covariance among pairs of responses *u* time intervals apart is assumed to decrease exponentially with *u*. Here, ρ is the correlation of two response variables one time unit apart. This particular *autocorrelation function* is for a *first order autoregressive process* (AR-1) because it arises if the data are generated by:

$$Y_t = X_t\beta + \varepsilon_t$$
$$\varepsilon_t = \rho\varepsilon_{t-1} + a_t \quad (2.)$$

where a_t is a "white noise" time series comprised of independent observations with mean 0 and common variance $\sigma^2/(1-\rho^2)$. Note this is an ordinary regression model but each residual can be expressed as a linear combination of the one that came before plus an independent error. The name "autoregressive process" refers to the second equation above in which the residuals at one time are a function of them previous one(s).

With a binary time series, the analogous marginal logistic model is given by:

$$\log\frac{\Pr(Y_t = 1)}{\Pr(Y_t = 0)} = X_t\beta$$
$$\log OR(Y_t, Y_{t-u}) = \theta_0 + \theta_1 u \quad (3.)$$

where $OR(Y_t, Y_{t-u})$ is the odds ratio, a measure of association for two binary responses *u* times apart that is assumed to be a linear function of *u* characterized by unknown parameters, θ. For a marginal model like this one, the interpretation of the regression coefficients, β, does not depend on the assumption made about the log odds ratio. With binary responses, it is better to specify the model for the associations in terms of log odds ratios rather than correlations that are constrained by the mean. He we assume the odds ratio decays exponentially with the lag *u.* This binary equivalent of the autocorrelation function has been investigated by Heagerty & Zeger (24) who termed it the *lorelogram*.

An alternative formulation is to model the conditional expectation of the response variable as a function of: the covariates, X_t, but also the past responses, Y_{t-1}, $Y_{t-2},\ldots,Y_1$. In this case, we are combining the model for the dependence of Y on X with the model for the autocorrelation by including the prior Y values as predictors. We typically assume that the response at time t depends on only the most recent responses, for example from times t-1 and t-2.

In the linear regression case, a basic conditional model has the form:

$$Y_t = X_t\beta + \alpha_1 Y_{t-1} + \cdots + \alpha_p Y_{t-p} + a_t \quad (4.)$$

Where the a_t are assumed to independent and identically distributed, mean zero random variables with common variance. Because the model above assumes that Y_t depends only on the p previous responses, it is an example of a *Markov chain* of order p.

This formulation is commonly used in economics. Statisticians tend to prefer a different parameterization of the same model:

$$Y_t = X_t\beta + \alpha_1(Y_{t-1} - X_{t-1}\beta) + \cdots + \alpha_p(Y_{t-p} - X_{t-p}\beta) + a_t . \quad 5.$$

In this version, we regress the response on the deviation of prior responses from their mean, rather than on the responses themselves. By doing so, the marginal expectation $E(Y_t|X_t) = X_t\beta$ so that the regression coefficients, β, in this conditional model also have a marginal model interpretation. The two models immediately above give exactly the same predictions. The distinction between these two formulations has been discussed in detail by Louis (29).

With binary responses, an example of a conditional model is

$$\log\frac{\Pr(Y_t = 1 \mid Y_{t-1},\ldots,Y_1)}{\Pr(Y_t = 0 \mid Y_{t-1},\ldots,Y)} = X_t\beta + \alpha_1 Y_{t-1} + \cdots + \alpha_p Y_{t-p} . \quad 6.$$

The interpretation of a coefficient β_j is the log odds ratio comparing the probability of response for persons whose X_j values differ by one unit and who have the same past p responses. Since in this model there is no interaction of the predictor variables X with the past responses, the interpretation of β_j is the same for every history, Y_{t-1}, $Y_{t-2},\ldots,Y_{t-p.}$ If we consider the case where the outcome is a disease presence/absence and all past responses 0s, then this can be seen to be a model for disease incidence. In this case, a marginal model is for disease prevalence. Diggle et al. (14) discuss the use of this conditional model for modeling disease incidence and maintenance in the context of many shorter time series.

To illustrate the difference in the two parameterizations of the linear time series model above, we have regressed the JiVitA weekly number of births on: a linear trend; sine and cosine terms with 52 week and 26 week periods to represent seasonality; and on the number of births in each of the two previous weeks. We fit the model using the past birth numbers themselves and then their residuals corrected for trend and seasonality. In the former, the trend estimate is -0.50 (95% CI: -0.72, -0.27). It is the decrease per week in the expected number of births but comparing weeks whose recent histories are identical. In the second model, the estimate is -0.79 (95% CI: -1.08, -0.50). Once the seasonality is controlled for using the harmonic terms, this model allows us to conclude that the births in this study population are decreasing by 0.79 per week or roughly 40 per year. Once seasonality is modeled, the evidence is strong that the trend is not due to chance alone.

Autoregressive Moving Average (ARMA) Models for Gaussian Processes

The autocorrelation function (ACF) for a first order autoregressive model decays exponentially with lag. In general, the ACF for an AR-p model is mixtures of exponential functions that decay or oscillate to zero with increasing lag. The regression form of the AR-1 model above makes it clear that, given Y_{t-1}, the correlation of Y_t with Y_{t-u}, u>1 will be 0.0. In general, this conditional correlation of Y_t with Y_{t-u}, given the intervening observations Y_{t-1}, …, Y_{t-u+1}, is called the partial autocorrelation function (PACF). It is useful to determine whether to increase the order of an autoregressive model because for an AR-p model, the PACF is non-zero for $u \leq p$ and 0 otherwise. Autoregressive models adequately describe the autocorrelation functions observed in a

wide range of public health and biomedical problems and are easy to use with standard regression programs.

In some problems, however, the autocorrelation is large for only a few lags and then drops to near 0.0, rather than decaying exponentially with lag. Here, we use a moving average (MA-q) model defined

$$Y_t = a_t + \theta_1 a_{t-1} + \cdots + \theta_q a_{t-q} \qquad 7.$$

where the a_t are independent Gaussian (normal) variables with mean 0 and constant variance. The MA model describes the observed process as a finite linear combination of independent innovations. Its ACF is non-zero for exactly the first q lags and zero beyond. Its PACF is a mixture of exponentials or damped harmonics, like the ACF for the AR model.

Finally, the ARMA-(p,q) model is a combination of the two models. The ARMA-(1,1) is particularly useful because it can arise when an AR-1 process is observed with error (13). The classic text on ARMA modeling is by Box & Jenkins (4) who introduced the modern approach to time series modeling.

ARMA models are stationary by their definition and hence do not have long-term trends that take the process far from its mean level. When time series data display a clear trend, a stationary model is not appropriate. We discussed above including predictor variables, for example simple functions of time, to deal with trends. But another approach is to assume process change, not the process itself, is stationary. The simplest model of this type is the *random walk* where the change from one time to the next is a Gaussian variable with mean zero and constant variance. This process can meander far from its zero mean. The random walk is a special case of an ARIMA-(p,d,q) model or autoregressive, integrated, moving-average model where an ordinary ARMA-(p,q) model is assumed, not for the data, but for its d^{th} difference. See Box & Jenkins (4) for details.

Non-linear Time Series Models

The term "non-linear" time series refers to models where there are systematic departures from the Gaussian, ARMA framework described above. There are many interesting ways to formulate such models. We have already mentioned the generalized linear model (31) extensions of autoregressive models for binary, count and other non-Gaussian data. Diggle et al. (14) discuss GLMs for time series in detail. Log-linear models for time series of counts were developed by Zeger (48) and have been modified and applied by many investigators in studies of air pollution and morbidity and mortality (e.g. 2).

Another simple way to extend autoregressive models and create non-linearity is to adopt an AR-p model but allow the conditional variance of Y_t given the past to depend on the past observations as the conditional mean does. These *autoregressive conditional heteroscedastic or ARCH models*, first introduced by Engle (16), can account for periods of increased variation in a process that might reflect a sudden increase in instability of the underlying process.

Also in the AR family, a non-linear model results if we assume that the past responses influence the current one through a non-linear function. For example, we might assume that the two previous responses predict the current one by:
$\alpha_1 Y_{t-1} + \alpha_2 Y_{t-2} + \alpha_{12} Y_{t-1} Y_{t-2}$. The interaction would allow two successive extreme values to have a greater influence than would be predicted by their individual effects.

There is close connection between non-linear time series models and *chaos theory* or *dynamical systems* (19). The simple deterministic equation $Y_t = rY_{t-1}(1-Y_{t-1})$ was first discovered by Robert May (30) to produce stochastic looking time series for a certain set of values of r. This equation is a deterministic analogue of a non-linear first order autoregressive process. The connections between non-linear time series models and dynamical systems are developed by Tong (44) and references therein.

Recent Applications of Time Series Analysis

Papers reporting time series analyses are wide spread in the biomedical and public health literature. A search of PubMed for the term "time series analysis" in the titles or abstracts of English publications identifies 999 papers, 798 since 1990. This section gives a brief overview of clusters of topics or disciplines where time series methods have been most commonly employed in recent years to demonstrate the flavor and diversity of applications and to provide the reader with entrée to specific areas that may be of further interest.

The explosion of biotechnology has increased the rate and complexity of data acquisition producing single time series or large collections of time series in the basic sciences. For example, gene expression is now repeatedly measured through time creating tens of thousands of time series that are to be analyzed simultaneously to describe molecular and cellular processes (42; 17). To date, most investigators have applied ordinary least squares regression to the data. Because autocorrelation is ignored, standard errors of regression coefficients will not be valid, but if the nature of the autocorrelation is similar across genes, the relative ranking of coefficients across genes will be correct.

Time series analysis is also common in physiologic studies. For example, DiPietro et al. (15) used advanced sonography to monitor fetal neuro-development. They have studied the psychophysiology of the maternal-fetus relationship by monitoring maternal heart rate and skin conductance in tandem with fetal heart rate and motor activity at various times during gestation. As detailed in the example in the next section, they use cross-correlation functions to quantify the evolving interaction of mother and fetus.

In physiologic research, image analysis has dramatically increased the demand for time series studies. Positron emission tomography (PET) and functional magnetic resonance imaging (fMRI) produce what is inherently time series data. The signal to noise ratio is small for both technologies making more sophisticated times series models attractive. With fMRI, the scientific focus is on oxygen consumption while oxygen transport is directly observed. This necessitates adjustment for hemodynamic delay using time series techniques. With fMRI, it is common to obtain a time series at each of 10^5 or more voxels (positions) in the brain. Hence, time series analyses are repeated this many times, once at each position and the parameters, usually regression coefficients are then summarized in maps. This approach, implemented as Statistical Parametric Mapping or SPM was introduced by Friston and colleagues (18) and is supported by software from the Wellcome Trust in England.

In critical care medicine, new technology has made it possible to simultaneously acquire at high sampling frequency, time series on many physiologic processes, to store and analyze these data to monitor patient health status and to predict outcome. For

example, Goldstein et al. (20) describe a system in a pediatric intensive care unit that samples and stores 11 different physiologic variables once per second. Goldstein and co-authors (21) have previously illustrated the application to heart rate variability of time and frequency domain time series methods to track patient status and to predict outcomes. Clinical guidelines using such analyses have been developed (43). Non-linear dynamic models and the associated chaos theory has even attracted the attention of critical care physicians monitoring patient status (22).

Time series methods have had substantial use in basic epidemiologic studies of infectious and chronic diseases. A recent monograph of statistical methods including time series models for monitoring population health is by Brookmeyer & Stroup (6). An illustration of times series methodology is given by Checkley et al. (10) who studied the effects of el Nino-driven variations in temperature on hospital admissions for diarrhoeal diseases in Peru. Allard (1) reviews the use of time series methods for tracking infectious disease processes.

In environmental epidemiology, time series models are used to characterize the variation in environmental conditions and to investigate the relationship of exposures to health outcomes. Modern time series studies of air pollution and health began in the 1970s. For example, Hexter & Goldsmith (25) studied the association of daily mortality in Los Angeles county with levels of carbon monoxide. Since then, there has been an explosion of time series research on the association of particulate and other air pollution measures with daily hospitalizations and deaths in individual cities and across the U.S., Europe, and Canada. See Bell et al. (2) for an excellent review of the history of this work and for an extensive bibliography.

Health services researchers employ time series to evaluate planned and unplanned interventions. For example, Meara and colleagues (32) used the autoregressive regression analysis described above to estimate the change in time trends for percentages of newborns receiving medical services associated with national and state legislation that mandated such services be provided. They show that legislation in the Ohio Medicaid population was associated with a dramatic change from an increasing to decreasing trend in the fraction of newborns discharged from the hospital within one day. Similar methods were used to quantify the effects of the SARS epidemic on medical expenditures in Taiwan (8) and to test the effects of Florida's repeal of its motorcycle helmet laws on traffic fatalities (33).

Finally, demographic analyses of population health status often involve time series methods. For example, Shmueli (40) fit a time series regression model to estimate the effects of income inequality as measured by the Gini Index on life expectancy and infant mortality over a 21 year period in Israel. A short distributed lags model was assumed so that health outcomes could depend on the degree of inequality from multiple recent years.

Recent Examples

This section presents two recent examples of exploratory time series analyses using time and frequency domain methods. In each case, simple displays of the data using the methods described above make apparent characteristics of the underlying process that are worth investigating further.

Fetal monitoring: DiPietro and colleagues (15) acquired measurements of fetal heart rate (FHR) and fetal movement (FM) 5 times per second for a 50 minute period on 120 mother-fetal pairs who were monitored in this way at 20, 24, 28, 32, 36 and 38-39 weeks of gestation. Since FHR and FM are simultaneously recorded, this is an example of a multivariate time series where Y_t is a vector with two entries, call them Y_1 and Y_2. Figure 6 displays these data for one monitoring session.

It is well known clinically that in the third trimester, large fetal heart accelerations are associated with fetal activity. A relatively straight-forward time series technique, the *cross-correlation function*, provides a visual description of how these associations develop with weeks of gestation. As described above, the autocorrelation function $\rho(u)$ is the correlation for values of a series at time t with those from the same series at time t-u. By the stationarity assumption, we estimate $\rho(u)$ by the sample correlation coefficient for the pairs (Y_t, Y_{t-u}) for $t=u+1, u+2,\ldots,n$.

The cross-correlation function $\rho_{12}(u)$ between two series Y_{1t} and Y_{2t} is just the correlation between observations from Y_1 at time t with those from Y_2 at time t-u, that is, u times before. We estimate $\rho_{12}(u)$ by the sample correlation coefficient from the pairs $(Y_{1t}, Y_{2\,t-u})$. Note that $\rho_{12}(u)$ can be different for positive and negative lags while the autocorrelation function satisfies $\rho(u) = \rho(-u)$.

Figure 7 is a descriptive plot of the average, over individuals, of the cross-correlation functions for FM and FHR for each gestation week. Notice that a peak cross-correlation at lag of approximately -6 seconds (FM leads FHR) starts to appear at 24 weeks of gestation. As the fetus gets older, this peak grows and becomes more clearly defined. This result, first discovered by DiPietro et al. (15) captures a potentially important characterization of the development of the cardiovascular system of the fetus.

Mapping Brain Function: Electrocorticographic (ECoG) signals are brain electrical potentials recorded by electrodes placed on the cerebral cortex of subjects undergoing surgery for intractable epilepsy. In this study, ECoG signals were recorded on a grid of 48 electrodes to localize seizure foci and to map brain functions (12). These signals were recorded 1000 times per second while subjects performed cognitive and motor tasks. Each session produced a time series vector with an entry for each of the 48 electrodes. In Figure 8, we see time series obtained for one electrode for 10 trials of the same motor task. In general, the first 1.4 seconds "look different" from the remaining 1.6 seconds; the signals appear to be *non-stationary*. Notice the average of the signals is not a useful quantity. This is because the signals are out of phase, that is, have their minima and maxima at different times and therefore average out to 0. This is a case where it is more convenient to consider the periodogram and spectra, collectively called *frequency domain* statistics.

Each of the 47 2-dimensional plot in Figure 9 corresponds to an electrode. The horizontal axis is time from 0 to 3 seconds. Because of the change in the look of the series over time, the periodogram and spectrum for each electrode was estimated for bins of time moving from left to right. The vertical axis represents frequency with slow fluctuations at the top and faster ones at the bottom. The color represents the estimated value of the spectrum which measures the size of the fluctuations at each frequency according to the scale on the right.

This descriptive display shows that the time series for the electrodes corresponding to the lower right part of the array in Figure 9 have substantial variations at the longer-time scales (red color at top of the plots). Of particular interest is the dramatic change in activity in column 6, rows 3 and 4 between 0.5 and 1.0 seconds into the trial at frequencies between 8 and 13 Hz. Variation at these frequencies is "turned off" as the subject does the task. These frequencies are obviously essential to this particular brain function (12). This part of the brain should not be disturbed if the skill to perform this particular task is to be preserved after the surgery.

List of Figures

Literature Cited

1. Allard R. 1998. Use of time-series analysis in infectious disease surveillance. *Bull. World Health Organ.* 76:327--33

2. Bell ML, Samet JM, Dominici F. 2004. Time-Series Studies of Particulate Matter. *Annu. Rev. Public Health* 25:247--80

3. Bloomfield P. 2000. *Fourier Analysis of Time Series: An Introduction*. New York: Wiley

4. Box GEP, Jenkins GM. 1976. *Time Series Analysis: Forecasting and Control*. San Francisco: Holden-Day

5. Brockwell PJ, Davis RA. 1987. *Time Series: Theory and Methods*. New York: Springer

6. Brookmeyer R, Stroup DF. 2004. *Monitoring the Health of Populations*. New York: Oxford University Press

7. Bruce A, Gao H-Y.1996. *Applied Wavelet Analysis with S-plus*. New-York: Springer

8. Chang H-J, Huang N, Lee C-H, Hsu H-J Hsieh C-J, Chou Y-J. 2004. The impact of the SARS epidemic on the utilization of medical services: SARS and the fear of SARS. *Am. J. Public Health* 94:562--64

9. Chatfield C. 1984. *The Analysis of Time Series; An Introduction*. London: Chapman and Hall

10. Checkley W, Epstein LD, Gilman RH, Figuero D, Cama RI, et al. 2000. Effects of El Niño and ambient temperature on hospital admissions for diarrhoeal diseases in Peruvian children. *Lancet* 355:442--50

11. Cleveland WP, Tiao GC. 1976. Decomposition of seasonal time series: a model for the census X-11 Program. *J. Am. Stat. Assoc.* 71:581--87

12. Crone NE, Hao L, Hart J Jr, Boatman D, Lesser RP et al. 2001. Electrocorticographic gamma activity during word production in spoken and sign language. *Neurology* 57:2045--53

13. Diggle PJ. 1990. *Time Series; A Biostatistical Introduction*. Oxford: Clarendon Press

14. Diggle PJ, Heagerty PJ, Liang K-Y, Zeger SL. 2002. *Analysis of Longitudinal Data, second edition*. New York: Oxford University Press

15. DiPietro JA, Irizarry RA, Hawkins M, Costigan KA, and Pressman EK. 2001. Cross-correlation of fetal and somatic activity as an indicator of antenatal neural development. *Am. J. Obstet. Gynecol* 185:1421--28

16. Engle RF. 1982. Autoregressive conditional heteroscedasticity with estimates of the variance of U.K. inflation. *Econometrica* 50:987--1008

17. Filkov V, Skiena S, Zhi J. 2002. Analysis techniques for microarray time-series data. *J. Comput. Biol.* 9:317--30

18. Friston KJ, Jezzard PJ, Turner R. 1994. Analysis of functional MRI time-series. *Hum. Brain Mapp.* 1:153--71

19. Gleick J. 1987. *Chaos*. New York: Viking

20. Goldstein, B, McNames J, McDonald, Ellenby M, Lai, S, et al. 2003. *Crit. Care Med.* 31:433--41

21. Goldstein B, Fiser DH, Kelly MM, Mickelsen D, Ruttimann U, et al. 1998. Decomplexification in critical illness and injury: the relationship between heart rate variability, severity of illness, and outcome. *Crit. Care Med.* 26:352--57

22. Goldberger AL. 1996. Non-linear dynamics for clinicians: chaos theory, fractals, and complexity at the bedside. *Lancet* 347:1312--14

23. Hastie T, Tibshirani R, Friedman J. 2001. *Elements of Statistical Learning*. New York: Springer-Verlag

24. Heagerty PJ, Zeger SL . 1998. Lorelogram: a regression approach to exploring dependence in longitudinal categorical responses. *J. Am. Stat. Assoc.* 93:150--62

25. Hexter AC, Goldsmith JR. 1971. Carbon monoxide: association of community air pollution and mortality. *Science* 172:265--67

26. Intergovernmental Panel on Climate Change. 2001. Contribution of working group to the third assessment report of the intergovernmental panel on climate change. Houghton J, Ding Y, Griggs M, Noguer M, van der Linden P, Dai X, et al, eds. *Climate Change 2001: the Scientific Basis*. Cambridge: Cambridge University Press

27. Johnson E, Dominici F, Griswold M, Zeger SL. 2003. Disease cases and their medical costs attributable to smoking: an analysis of the National Medical Expenditure Survey. *J. Econom.* 112:135--52

28. Li H, Maddala GS. 1996. Bootstrapping time series models. *Econom. Rev.* 15:115--58

29. Louis TA. 1988. General methods for analyzing repeated measures. *Stat. Med.* 7: 29--45

30. May RM. 1976. Simple mathematical models with very complicated dynamics. *Nature* 261: 459--67

31. McCullagh P, Nelder JA, 1989. *Generalized Linear Models*. New York: Chapman and Hall

32. Meara E, Kotagal UR, Atherton HD, Lieu TA. 2004. Impact of early newborn discharge legislation and early follow-up visits on infant outcomes in a state Medicaid population. *Pediatrics* 113:1619--27

33. Muller A. 2004. Florida's Motorcycle Helmet Law Repeal and Fatality Rates. *Am. J. Public Health* 94:556--58

34. National Center for Health Services Research.1987. National Medical Expenditure Survey. Methods II. Questionnaires and data collection methods for the household survey and the Survey of American Indians and Alaska Natives. National Center for Health Services Research and Health Technology Assessment.

35. Parzen M. 1999. *Stochastic Processes*. Philadelphia: Society for Industrial and Applied Mathematics

36. Pesenti A, Rohr M, Egidi M, Rampini P, Tamma F, et al. 2004. The subthalamic nucleus in Parkinson's disease: power spectral density analysis of neural intraoperative signals. *Neurol. Sci.* 24:367--74

37. Reinsel G. 1997. *Elements of Multivariate Time Series Analysis*. New York: Springer-Verlag

38. Refinetti R. 2004. Non-stationary time series and the robustness of circadian rhythms. *J. Theor. Biol.* 227:571--81

39. Ross SM. 1970. *Applied Probability Models with Optimization Applications*. San Francisco: Holden-Day

40. Shmueli A. 2004. Population health and income inequality: new evidence from Israeli time-series analysis. *Int. J. Epidemiol.* 33:311—17

41. Sommer A, Tarwotjo I, Katz J. 1987. Increased risk of xerophthalmia following diarrhea and respiratory disease. *Am. J. Clin. Nutr.* 45:977-80

42. Spellman P, Sherlock G, Zhang M, Iyer V, Anders K, et al. 1998. Comprehensive identification of cell cycle-regulated genes of the yeast *Saccharomyces cerevisiae* by microarray hybridization. *Mol. Biol. Cell* 9: 3273--97

43. Task Force of the European Society of Cardiology and the North American Society of Pacing and Electrophysiology. 1996. Heart rate variability: standards of measurement,physiological interpretation, and clinical use. *Circulation* 93:1043--65

44. Tong H. 1990. *Non-linear Time Series*. New York: Springer-Verlag

45. Tukey JW. 1977. *Exploratory Data Analysis*. Reading: Addison-Wesley

46. van der Laan, MJ, Robins JM. 2003. *Unified Methods for Censored Longitudinal Data and Causality.* Springer-Verlag, New York.

47. Zamarron C, Gude F, Barcala J, Rodriguez JR, Romero PV. 2003. Utility of oxygen saturation and heart rate spectral analysis obtained from pulse oximetric recordings in the diagnosis of sleep apnea syndrome. *Chest* 123:1567--76

48. Zeger SL. 1988. Regression model for time series of counts. *Biometrika* 75:621--30

49. Zeger SL, Dominici F, Samet J.1999. Harvesting-resistant estimates of pollution effects on mortality. *Epidemiology* 10:171--75

Figures

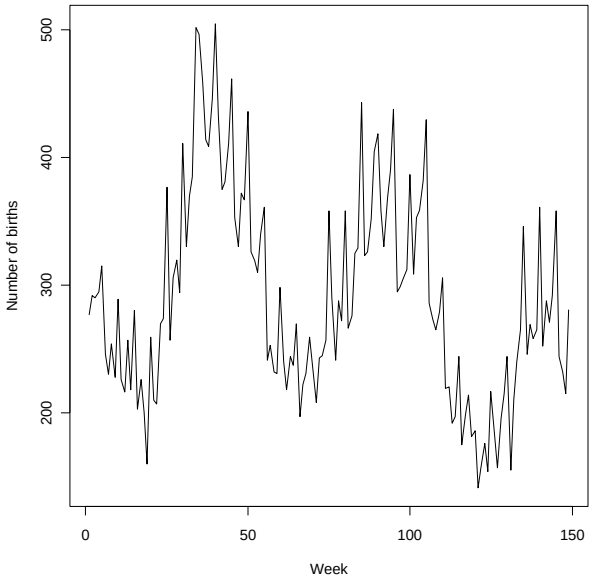

Figure 1

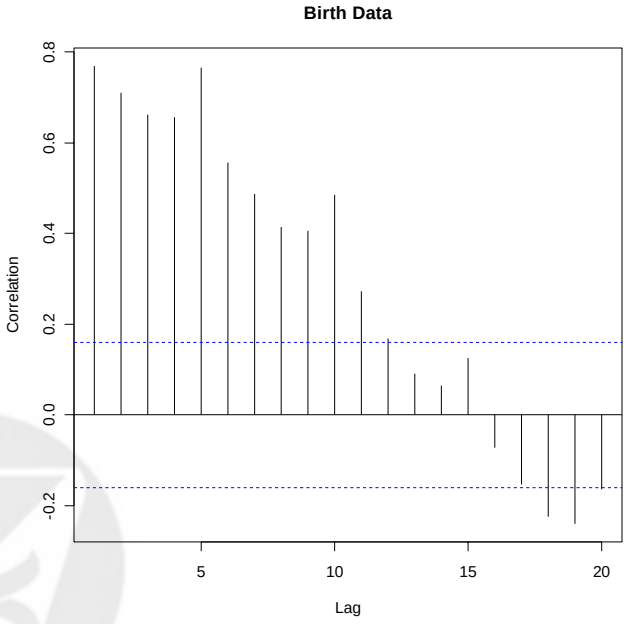

Figure 2

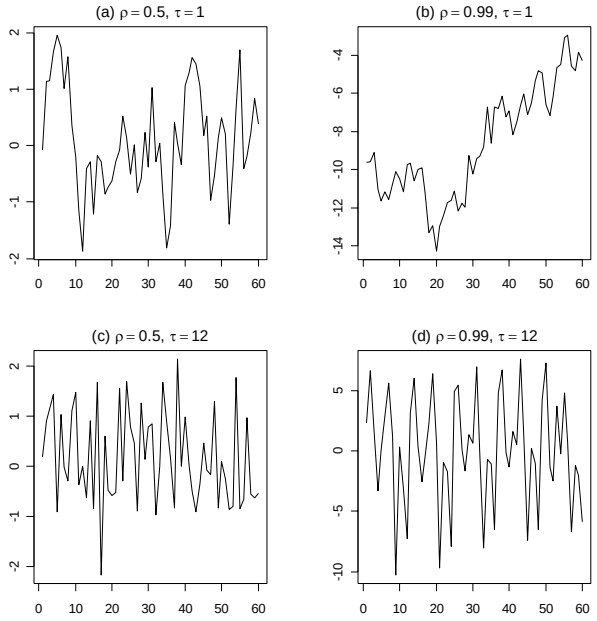

Figure 3

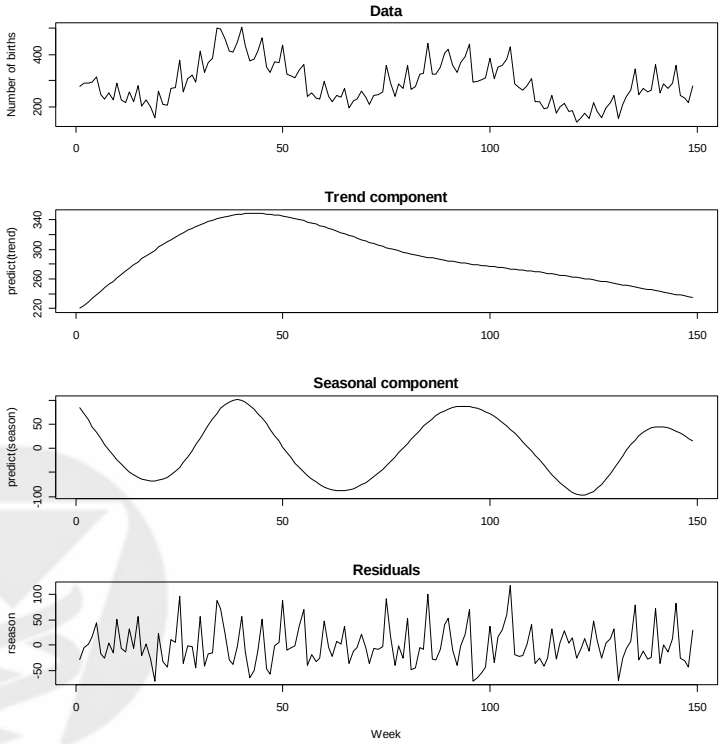

Figure 4

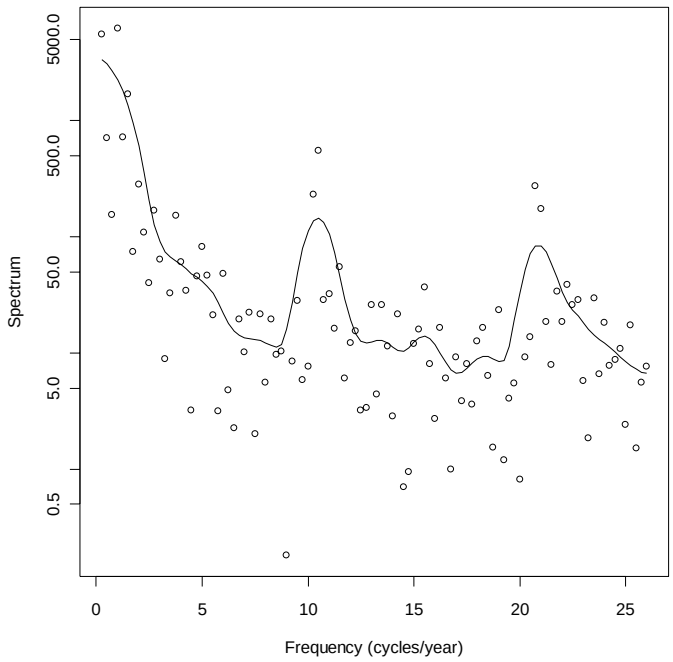

Figure 5

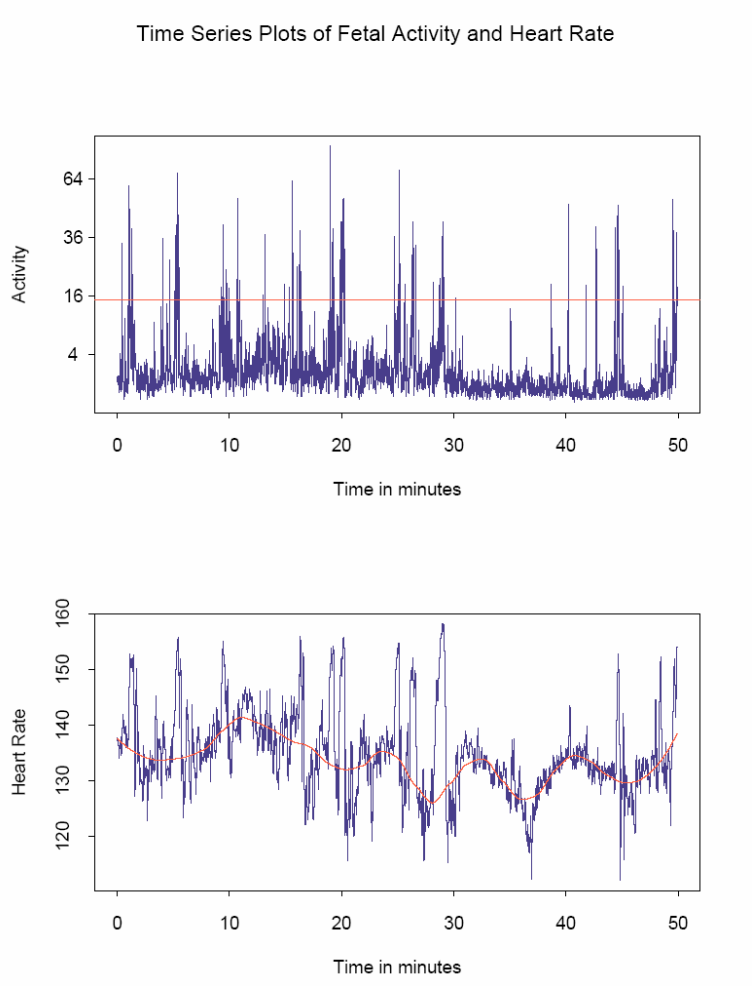

Figure 6

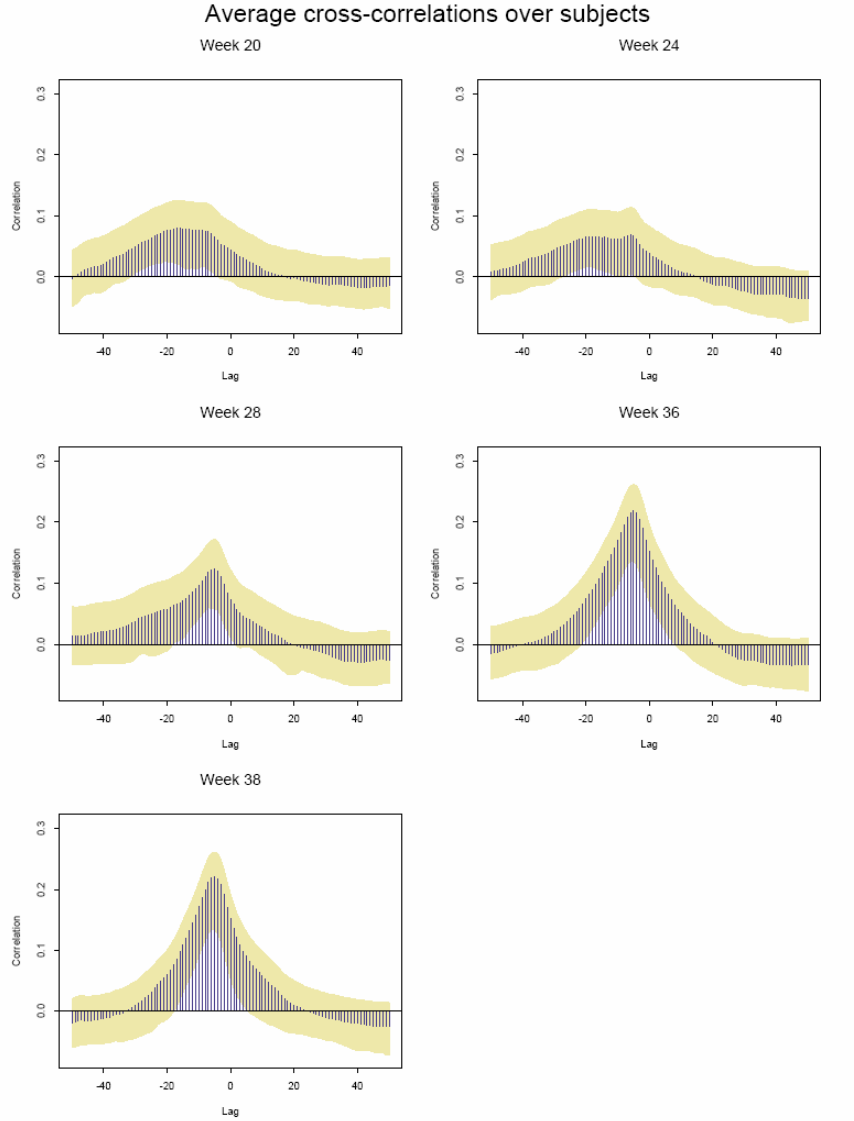

Figure 7

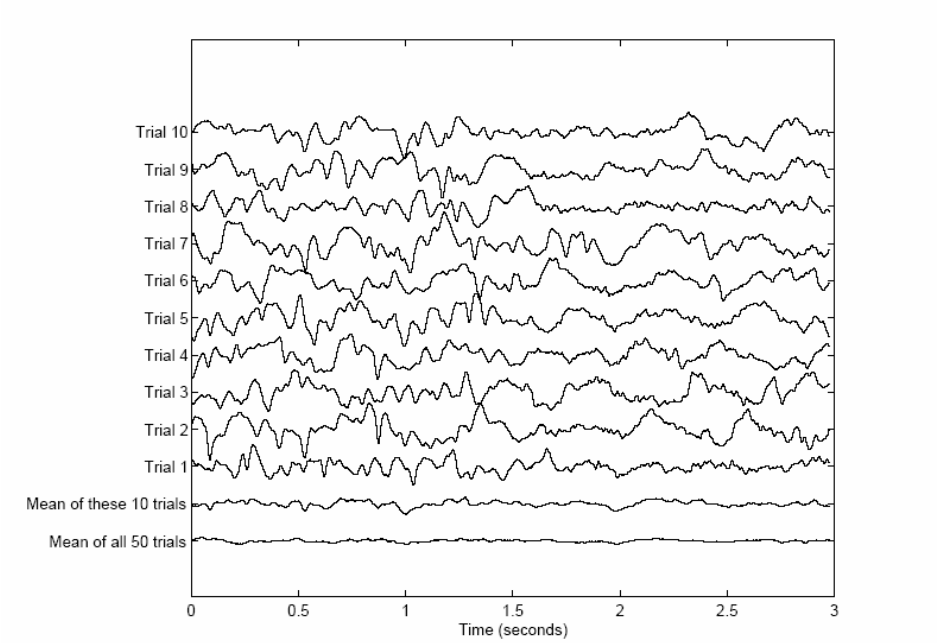

Figure 8

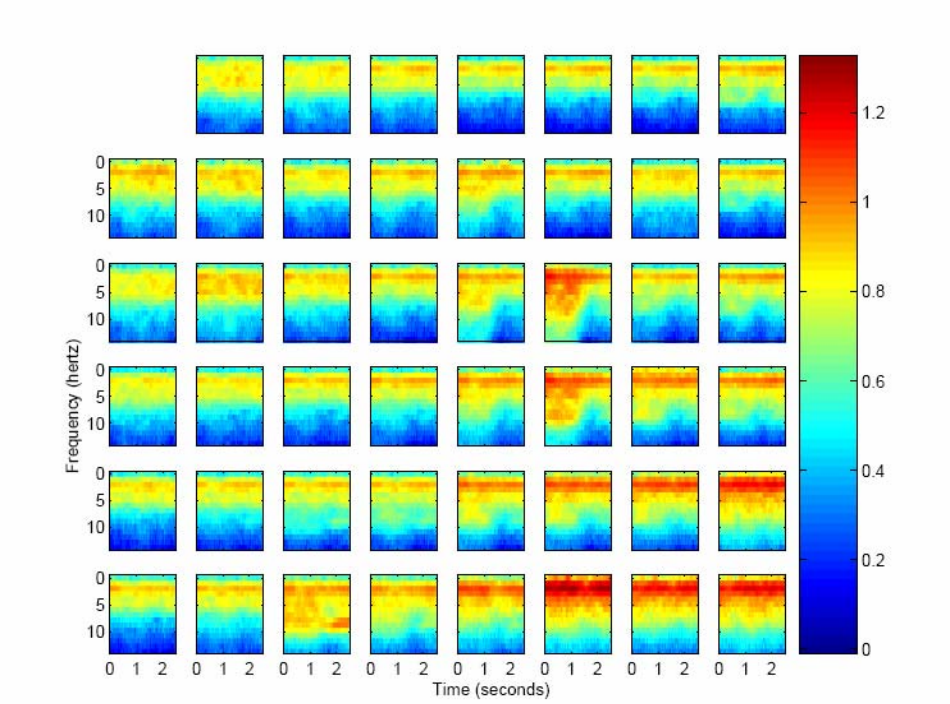

Figure 9