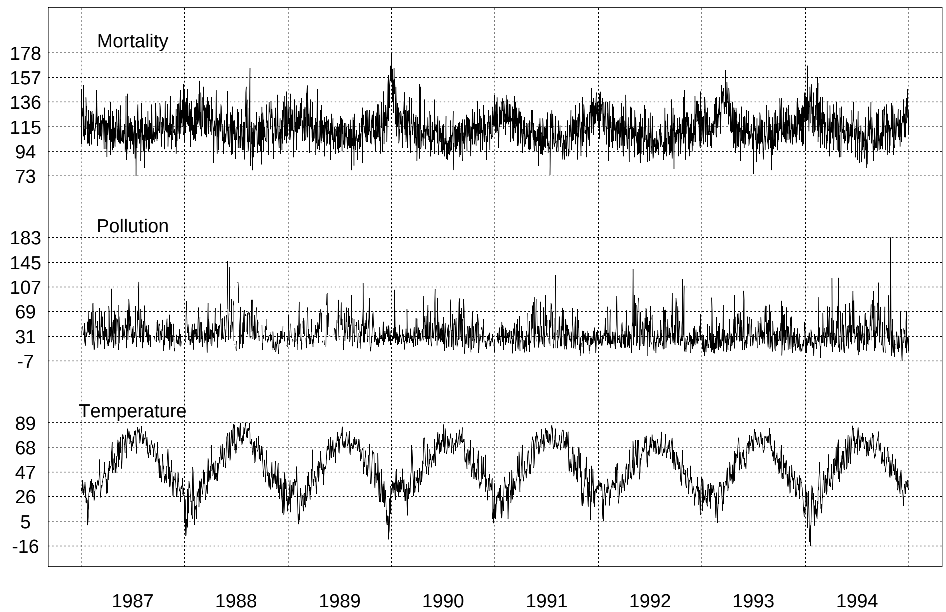
Chapter 10

Introduction to Time Series Analysis

A *time series* is a collection of observations made sequentially in time. Examples are daily mortality counts, particulate air pollution measurements, and temperature data. Figure 1 shows these for the city of Chicago from 1987 to 1994. The public health question is whether daily mortality is associated with particle levels, controlling for temperature.



We represent time series measurements with $Y_1, \ldots, Y_T$ where T is the total number of measurements. In order to analyze a time series, it is useful to set down a statistical model in the form of a *stochastic process.* A stochastic process can be described as a statistical phenomenon that evolves in time. While most statistical problems are concerned with estimating properties of a population from a sample, in time series analysis there is a different situation. Although it might be possible to vary the length of the observed sample, it is usually impossible to make multiple observations at any single time (for example, one can't observe today's mortality count more than once). This makes the conventional statistical procedures, based on large sample estimates, inappropriate. Stationarity is a convenient assumption that permits us to describe the statistical properties of a time series.

10.1 Stationarity

Broadly speaking, a time series is said to be *stationary* if there is no systematic trend, no systematic change in variance, and if strictly periodic variations or seasonality do not exist. Most processes in nature appear to be non-stationary. Yet much of the theory in time-series literature is only applicable to stationary processes.

One way of describing a stochastic process is to specify the joint distribution of the observations $Y(t_1), \dots, Y(t_n)$ for any set of times $t_1, \dots, t_n$ and any value of n. A time series is said to be strictly stationary if the joint distribution of $Y(t_1), \dots, Y(t_n)$ is the same as that of $Y(t_1 + h), \dots Y(t_n + h)$ for all $t_1, \dots, t_n$ and h. To see how this is a useful assumption, notice that the above condition implies that the expected value and covariance structure of any two components, $Y_a(t)$ and $Y_b(t)$, of a time series are constant in time

$$\mathrm{E}\{Y_a(t)\} = \mu_a \,,\ \mathrm{var}\{Y_a(t)\} = \sigma_a^2 \text{ and } \mathrm{corr}\{Y_a(t), Y_b(t+h)\} = \gamma_{ab}(h). \quad (10.1)$$

The function $\gamma_{ab}(h)$ is called the cross-correlation function if $a \neq b$ and the autocorrelation function if $a = b$.

In practice it is often useful to define stationarity in a less restricted way than that described above. In many cases, the statistical structure of the processes can be completely described with the second-order properties of equation (10.1). We can estimate the quantities in (10.1) using standard statistical procedures, for example we may estimate the cross-correlation at *lag* h, $\gamma_{a,b}(h)$ with the sample correlation of $Y_a(1), \dots, Y_a(T-h)$ and $Y_b(h+1), \dots, Y_b(T)$.

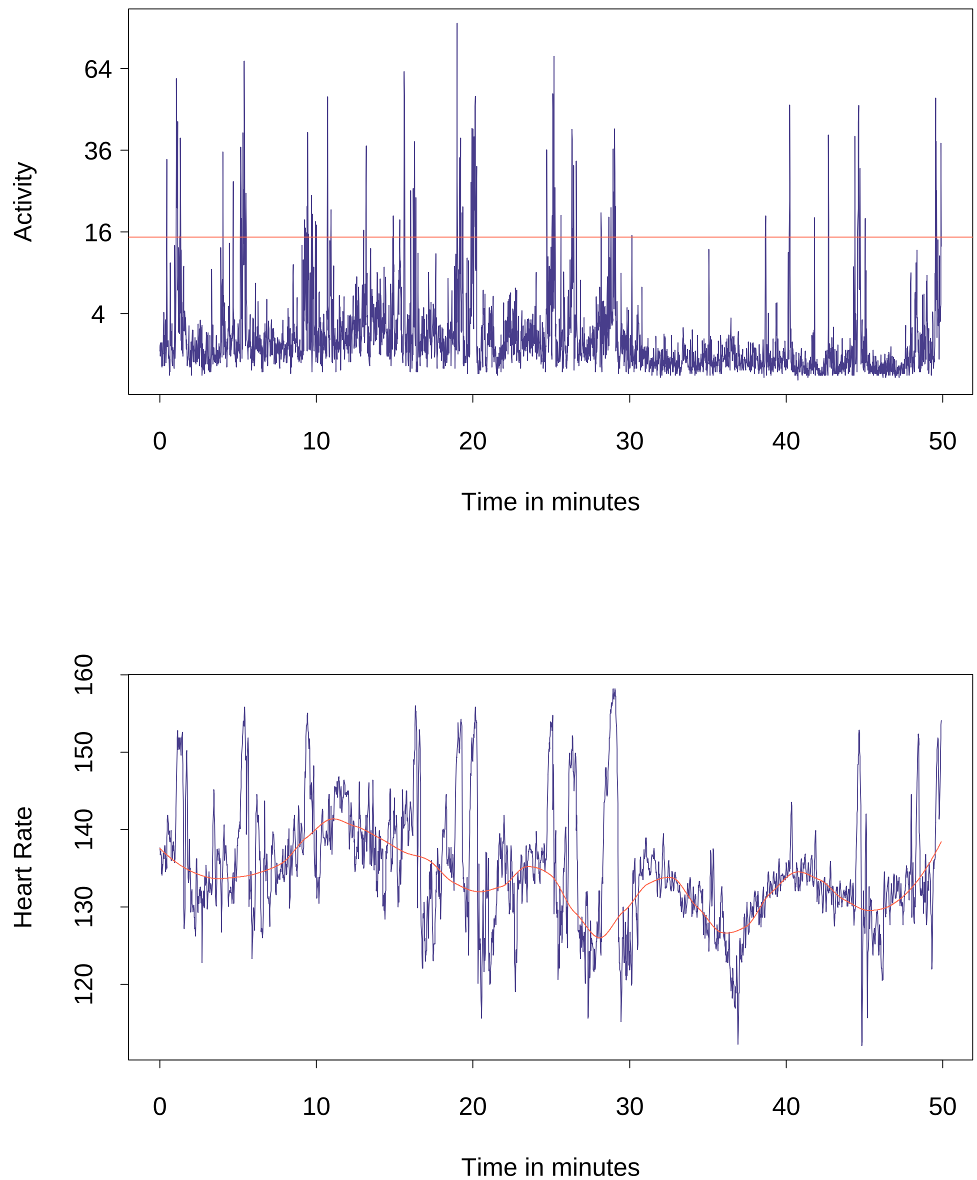
10.1.1 An example: Fetal Monitoring

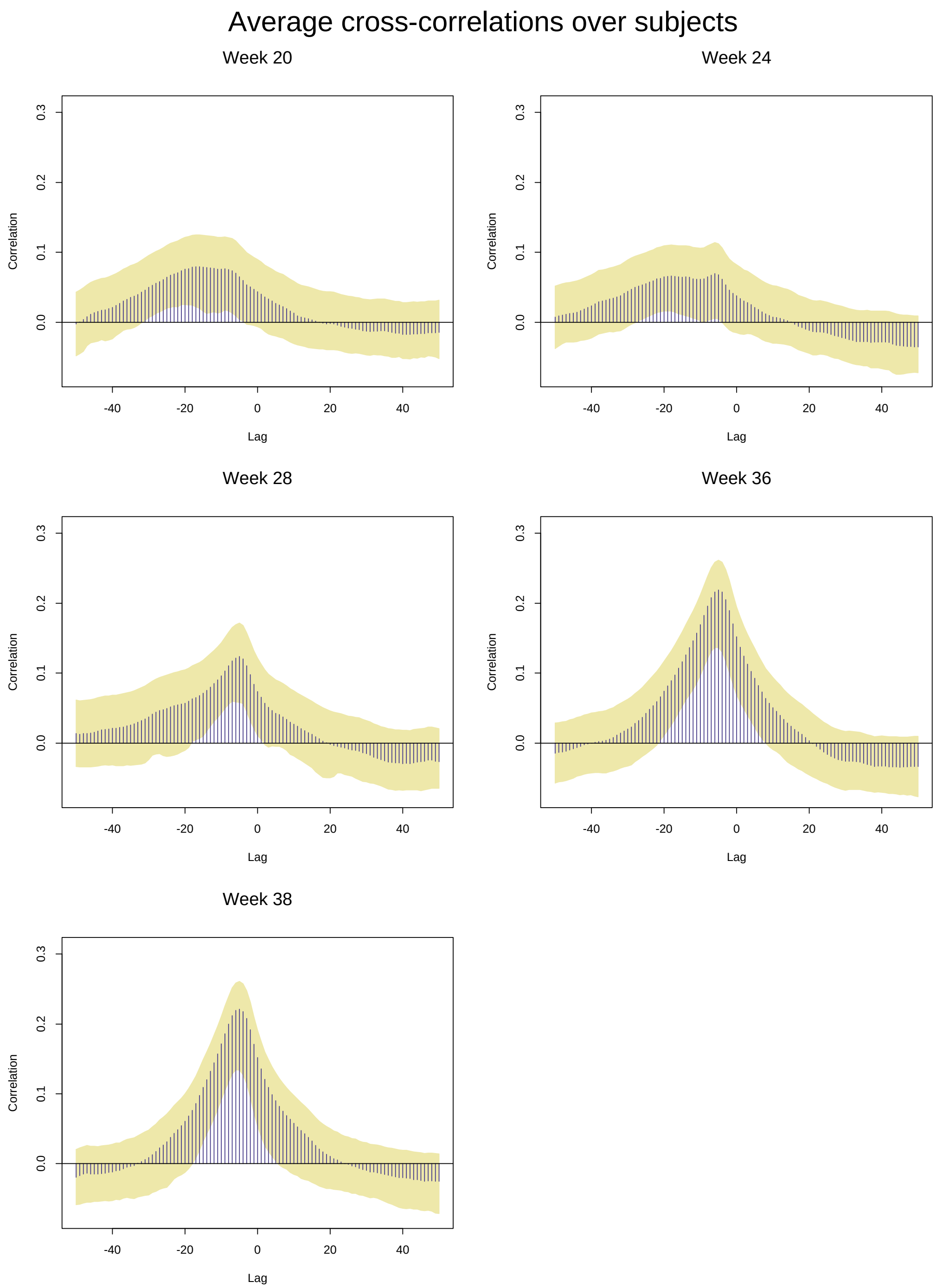
Measurements of fetal heart rate (FHR) and fetal movement (FM) are generated by maternal-fetal monitoring. Approximately 5 measurements per second are taken during 50 minutes on 120 subjects that are monitored at 20,24,28,32,36, and 38-39 weeks of gestation. Both FHR and FM are recorded giving us a multiple time series $Y(t), t = 1, \dots, 50 \times 60 \times 5$, where $Y(t)$ is a vector with 2 entries.

The association between accelerations of FHR and FM has been documented since the 1930s. For example, it has been observed that in the third trimester most large fetal heart accelerations are associated with fetal activity. In Section 4.2 we will describe how relatively straight-forward time series techniques provide a visual descriptions of how these associations vary with weeks gestation. These description have motivated a methodology that will provide us is with a more rigorous assessment of this relationship.

If we consider the FM and FHR measurements, seen in Figure 5, as outcomes from a two component time series, we may consider the cross-correlation function of these two components as a description of the association between these two processes. Notice that the measurements taken for each fetus at each gestation week has a cross-correlation function associated with them. In Figure 6, as a descriptive plot, we show the average, over individuals, of these functions for each gestation week. Notice that a peak at around the -6 second lag starts to appear in the plot for the 24 week gestation. As the fetus gets older, this peak grows and becomes more defined. This result can be considered a first step in the characterization of the relationship between FM and FHR.

Time Series Plots of Fetal Activity and Heart Rate




Average cross-correlations over subjects
Week 20
Week 24
Week 28
Week 36
Week 38
Correlation
Lag
0.0
0.1
0.2
0.3
-40
-20
0
20
40

10.2 Spectral Analysis

Sometimes it is useful to describe the properties of the time series in a frequency domain. The spectrum is defined as

$$f_{ab}(\lambda) = \frac{\sigma^2}{2\pi} \sum_{h=-\infty}^{\infty} \gamma_{ab}(h) \exp(-i\lambda h)$$

There is a one-to-one correspondence between the spectrum and the autocovariance function

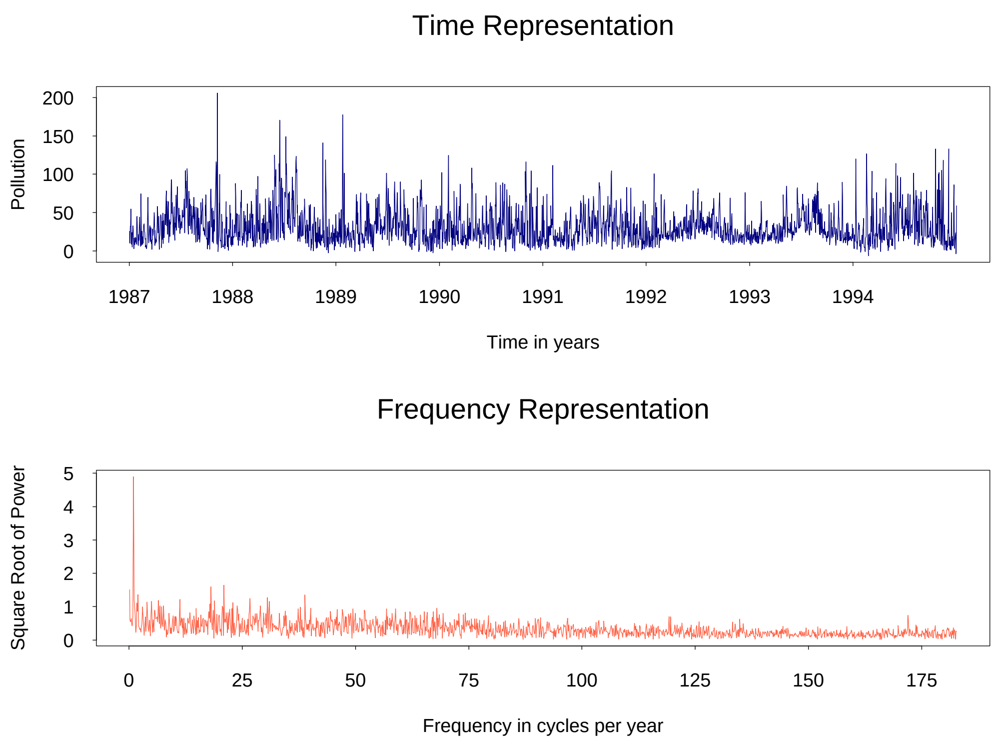
$$\sigma^2 \gamma_{ab}(h) = \int_{-\pi}^{\pi} f(\lambda) \exp(i\lambda h) d\lambda$$

We call $|f_{aa}|^2$ the power spectrum. A natural way of estimating a power spectrum is using the periodogram which is the modulus of the Fourier transform of the data

$$I(\lambda) = \frac{1}{2\pi T} |\sum_{t=1}^{T} Y_t \exp(-i\lambda t)|^2$$

We usually compute the periodogram at the Fourier frequencies $\lambda_j = (2\pi j)/T, j = 1, \ldots, T/2$. These have desirable statistical properties

The periodogram is also useful for detecting periodicities (deterministic ones) in the signal. It is a mathematical fact that if the data $Y_1, ...Y_T$ has a period p, the the periodogram will have peaks at frequencies $\lambda = 2\pi T/p$ and its multiples.



10.2.1 An Application

Figured 4a and 4c shows recorded ECoG signal for two channels for a subject that has received a sensory stimulus at some point during the recording. A straightforward way of estimating the spectrum of a stationary process is the periodogram

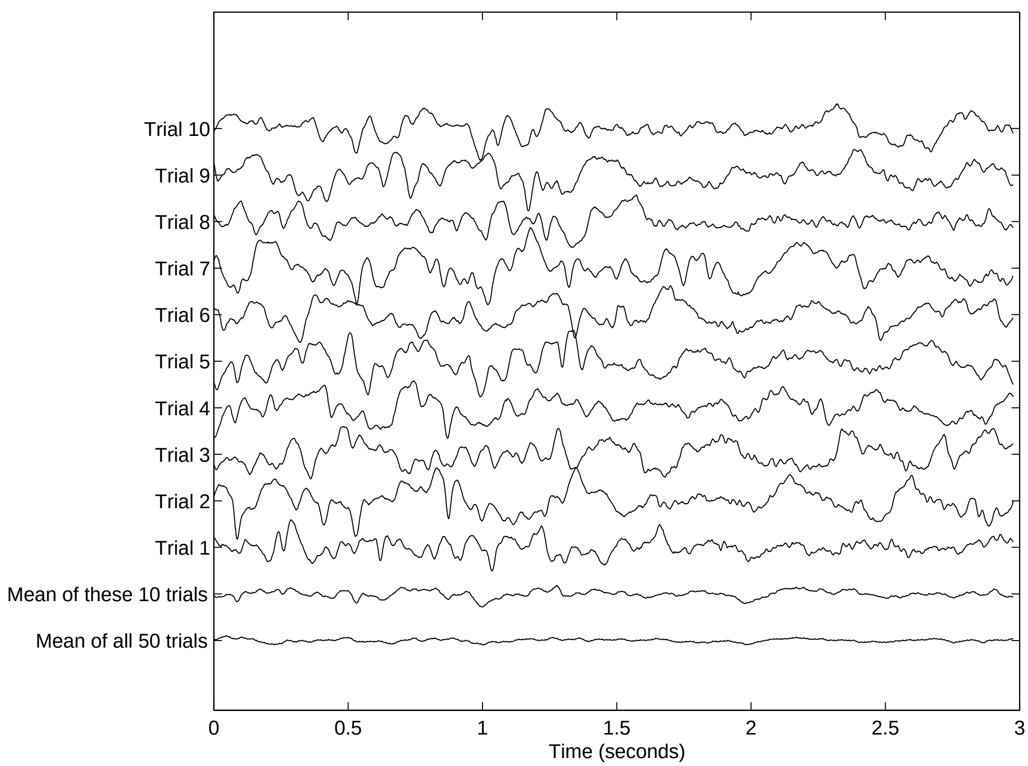
$$I(\lambda) = \frac{1}{2\pi T} \left| \sum_t Y(t) \exp(i\lambda t) \right|^2 .$$

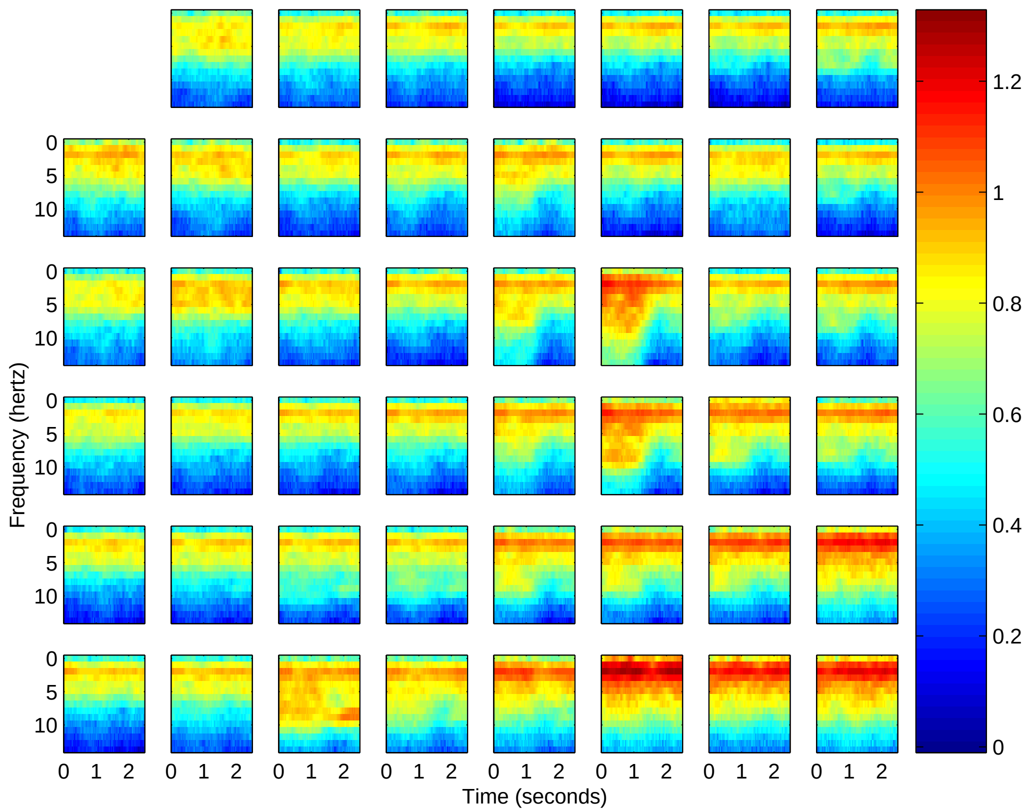
In Figures 4b and 4c the periodogram of this data is shown. Brain researchers have speculated that the so-called α $(8 - 13Hz.)$, β $(15 - 25Hz.)$, and γ $(> 30Hz.)$ bands of human brain signals can indicate functional activation of sensorimotor cortex. Notice that the periodogram exhibits a peak around frequencies 10 Hz., 20 Hz. and 60 Hz.. If we were to approximate the ECoG signal as a stationary processes, we would describe it as having periodic components around these frequencies. However, we are interested in learning how the signal changes when the subjects are given a stimuli. Thus it seems more appropriate to model the signal as a non-stationary processes and study the time-varying spectral density.

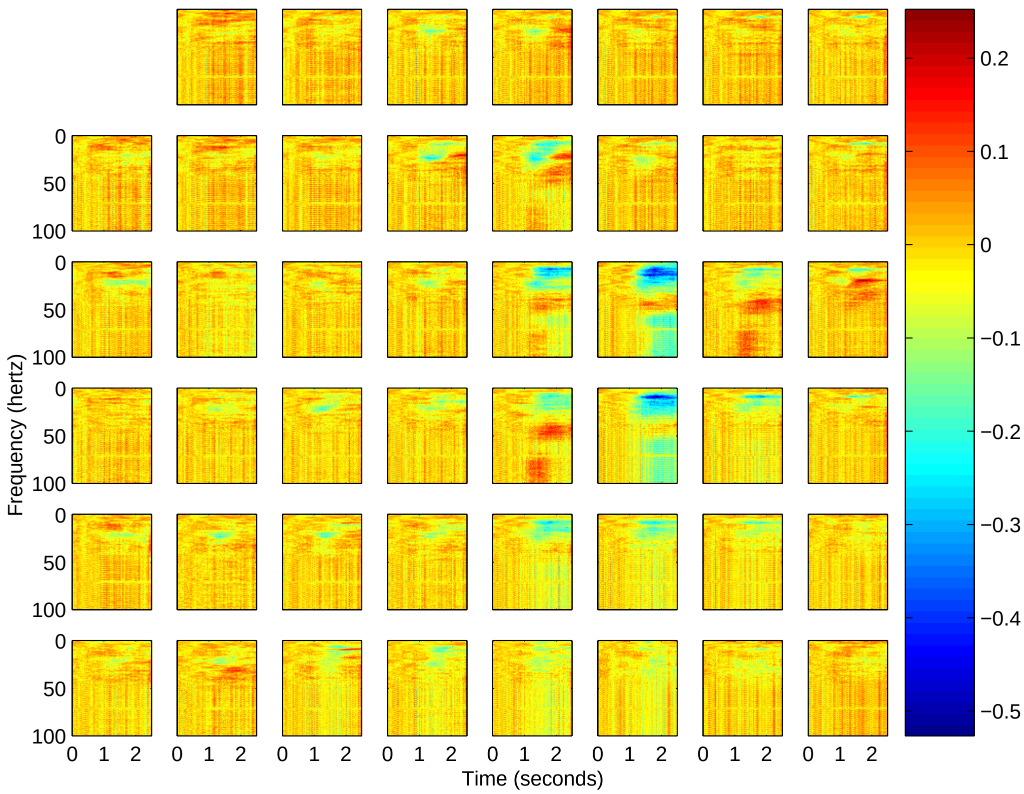
A straightforward estimate of a time-varying spectral density would be the dynamic periodogram. Basically, for each time t_0 we consider a window around that point of size $h(t_0)$ and estimate a weighted periodogram

$$I(t_0;\lambda) = \frac{1}{2\pi h(t_0)} \left| \sum_t w\left(\frac{t-t_0}{h(t_0)}\right) Y(t)\exp(i\lambda t)\right|^2 .$$

Figures 4c and 4e show the estimated time-varying spectral densities for the signals of channels 19 and 20 (lighter colors represent higher values) The figure seems to suggest that the α band changes power and frequency after the stimulus (time 0).




Frequency (hertz)
0
5
10
Time (seconds)
0 1 2
0
0.2
0.4
0.6
0.8
1
1.2



Frequency (hertz)
Time (seconds)
0
50
100
0
1
2
0.2
0.1
0
−0.1
−0.2
−0.3
−0.4
−0.5