Empirical Analyses of BOLD fMRI Statistics

II. Spatially Smoothed Data Collected under Null-Hypothesis and Experimental Conditions

G. K. Aguirre, E. Zarahn, and M. D'Esposito

Department of Neurology, Hospital of the University of Pennsylvania, Philadelphia, Pennsylvania 19104-4283

Received September 9, 1996

In the companion to this paper (E. Zarahn, G. K. Aguirre, and M. D'Esposito, 1997, NeuroImage, 000-000), we describe an implementation of a general linear model for autocorrelated observations in which the voxel-wise false-positive rates in fMRI “noise” datasets were stabilized and brought close to theoretical values. Here, implementations of the model are tested for use with statistical parametric mapping analysis of spatially smoothed fMRI data. Analyses using varying models of intrinsic temporal autocorrelation and either including or excluding a global signal covariate were conducted upon human subject data collected under null hypothesis as well as under experimental conditions. We found that smoothing with an empirically derived impulse response function (IRF), combined with a model of the intrinsic temporal autocorrelation in spatially smoothed fMRI data, resulted in a map-wise false-positive rate which did not exceed a 5% level when a nominal $\alpha = 0.05$ tabular threshold was applied. Use of other models of intrinsic temporal autocorrelation resulted in map-wise false-positive rates that significantly exceeded this level. fMRI data collected while subjects performed a behavioral task were used to examine (a) task-dependent global signal changes and (b) the dependence of sensitivity on the temporal smoothing kernel and inclusion/exclusion of a global signal covariate. The global signal changes within an fMRI dataset were shown to be influenced by the performance of a behavioral task. However, the inclusion of this measure as a covariate did not have an adverse affect upon our measure of sensitivity. Finally, use of an empirically derived estimate of the IRF of the system was shown to result in greater map-wise sensitivity for signal changes than the use of a broader (in time) Poisson (parameter = 8s) kernel.

INTRODUCTION

Blood oxygenation level-dependent (BOLD), functional MRI (fMRI) detects changes in a signal which reflects underlying neural activity (Kwong et al., 1992; Ogawa et al., 1993). This technique can be used to provide multiple observations (images) over time from multiple locations (voxels) within the brain. Typically, the data acquired are used to make statements regarding the effect of behavioral manipulations upon the fMRI signal. Ideally, these statements should be made with knowledge of their statistical significance; i.e., the likelihood that the findings are the result of chance alone. In order to provide this information, the analysis of fMRI data frequently begins with the application of a univariate, parametric statistical model to the time series of each voxel within the brain, providing a map of statistical effect (e.g., t value). This step addresses only a portion of the analytical challenge, however. Unanswered is the question of the map-wise significance of these effects in the face of the multiple comparisons performed (due to the large number of voxels in a given dataset). One class of solutions to this challenge, used for several years with positron emission tomography (PET) data, is to treat the entire statistical parametric map as the unit of hypothesis, the false-positive rate of which is to be controlled (Friston et al., 1991; Worsley et al., 1992).

Statistical parametric mapping (SPM) (Friston et al., 1991; Worsley et al., 1992) allows the selection of a map-wise significance threshold for an entire, multidimensional dataset which meets certain criteria. Among these criteria is that the components of the fields must have a multivariate Gaussian distribution and that the data provide a good lattice approximation to a continuous random field with stationary spatial autocovariance (Friston et al., 1991; Worsley et al., 1992). Because the intrinsic spatial smoothness of most fMRI data is small compared to the individual voxel size (Xiong et al., 1995; Forman et al., 1995), the latter requirement is met by smoothing the dataset. To the extent that the assumptions of SPM are met by fMRI data, the SPM model should predict (expected) false-positive rates for fMRI SPMs.

In the companion to this paper (Zarahn et al., 1997), several different parametric models were used to ana-
lyze spatially unsmoothed, human fMRI “noise” datasets—fMRI data acquired while subjects did not perform any time-locked behavior. These models were implemented within the framework of Worsley and Friston’s (1995) modified general linear model (GLM) for autocorrelated observations. Those analyses examined, in part, the effect of varying the expected intrinsic autocorrelation upon the observed voxel-wise, and Bonferroni corrected map-wise, false-positive rate. The general purpose of this paper is to modify and apply these parametric models, developed with spatially unsmoothed fMRI data, to the analysis of smoothed fMRI data, evaluated on the map-wise level using SPM methodology. In particular, the specificity and relative sensitivity of different implementations of Worsley and Friston’s GLM were examined. Two primary features of the models were varied: temporal autocorrelation (including the representation of intrinsic temporal autocorrelation and temporal smoothing) and the presence or absence of a global signal covariate.

As was demonstrated in the companion paper, temporal autocorrelation is present in fMRI data collected under the null hypothesis. This intrinsic autocorrelation was reasonably modeled by a function with a 1/frequency (hereafter referred to as 1/f) term, i.e., there was increasing power at lower frequencies. In addition, a putative impulse response function (IRF) of the fMRI system was empirically derived. It was shown that temporal smoothing with this IRF, coupled with the 1/f model, brought voxel-wise false-positive rates to levels slightly, but significantly, lower than tabular values. Alternatively, temporal smoothing alone (ignoring the 1/f autocorrelation) with a relatively broadband Poisson kernel (parameter = 8 s) also controlled the false-positive rate. In contrast, analyses which failed to account for intrinsic temporal autocorrelation [akin to the classic “boxcar” correlation (Bandettini et al., 1993)] produced false-positive rates which were significantly greater than tabular values. Here, these models were applied to the analysis of smoothed, null-hypothesis fMRI data to determine if false-positive rates are well controlled on the map-wise, SPM level. Furthermore, the models were used to analyze fMRI data collected while subjects performed a behavioral task in order to assess relative sensitivity. It has been proposed that, all else being equal, the optimal temporal smoothing kernel should match the hemodynamic response of the system (Friston et al., 1995a). Given that the two IRFs differ slightly in shape, and given that smoothing with the Poisson kernel affords fewer effective degrees of freedom (eff df) for the analysis, we anticipated that differences in relative sensitivity would be observed.

Also examined here were the properties and map-wise statistical effects of global signals (the average time series across all brain voxels) present in fMRI data. The companion paper demonstrated that fMRI datasets are spatially coherent. This spatial coherence was greater at lower temporal frequencies. In the analysis of spatially unsmoothed data, the inclusion of a global signal covariate within the examined analytical models stabilized (i.e., reduced the between-dataset variance of) false-positive rates. Additionally, a global signal covariate might reduce spatially coherent noise and thus increase power. Advocating the use of a global signal covariate [as performed in most PET analyses and some fMRI analyses (Friston et al., 1995a)] should be tempered, however, by the possibility that the global signal could reflect task activation. If this is the case, inclusion of a global signal regressor might reduce sensitivity and increase the incidence of spurious, negatively correlated voxels (Ramsay et al., 1993). This possibility was examined here by determining if the global signals from a behavioral dataset show a systematic relationship with the ongoing task. Next, the effect of a global signal covariate on our measure of sensitivity for task activation was examined.

Finally, the transition to spatially smoothed data and evaluation on the map-wise level requires the consideration of a few additional issues. First, spatial smoothing may change the intrinsic temporal autocorrelation (1/f model) present in fMRI data under the null hypothesis, as the spatial structure of the data may not be uniform across temporal frequency. As a result, we determined the appropriate 1/f model for the human subject noise data under conditions of spatial smoothing. These estimates are presented and used in further analyses. Second, an estimate of the expected spatial smoothness of the component process of the statistical map under analysis is essential for the assessment of SPM significance. This estimate of spatial smoothness is not a trivial problem, given the features of the datasets discussed below. We examine here the smoothness of our noise maps under different conditions of analysis.

**METHODS**

**MRI Technique**

Imaging was carried out on a 1.5-T SIGNA scanner (GE Medical Systems) equipped with a prototype fast gradient system for echoplanar imaging. A standard radiofrequency (RF) head coil was used with foam padding to comfortably restrict head motion. High-resolution sagittal T1-weighted images were obtained in every subject. A gradient echo-echoplanar sequence was used to acquire data sensitive to the BOLD signal at a $T_R = 2000\text{ ms}$, $T_E = 50\text{ ms}$. Resolution was $3.75 \times 3.75\text{ mm}$ in plane and $5\text{ mm}$ between planes (16 axial slices acquired). A total of 160 gradient echo-echoplanar images in time were obtained per slice in each 320-s run. Twenty seconds of “dummy” gradient and RF pulses preceded the actual data acquisition.
Motion Compensation

A slice-wise motion correction method was utilized which removed spatially coherent signal changes via the application of a partial correlation method to each slice in time. For each axial slice at each time, a difference image between that slice at Time t and that slice at Time 0 (a motion image) was correlated with an image composed of the difference between the slice at Time 0 shifted to the right 1 voxel and that same slice shifted to the left 1 voxel (an X-shift image). The same operation was performed for y shifts (using Y-shift images). The X-shift and Y-shift images, weighted by the strength of their respective correlations with the motion image, were subtracted from the image of the slice at Time t. Thus, the rationale of this method was to subtract out signal changes that correlated with small (on the order of a voxel) translations in the x and y dimensions. A conceptually similar method for motion in the z dimension was then applied to each axial image. A pseudo-Z-shift image was computed for each axial slice by subtracting the average of the first 10 images (in time) from the average of the last 10 images. The Z-shift image, weighted by its correlation with a motion image (computed after applying the corrections described above for X and Y shifts) was then subtracted from the slice image at Time t. The rationale for the pseudo-Z-shift image was that translation in the z dimension would occur steadily throughout the acquisition of images within a run. For additional comments on the motion correction method used, and how it compares with other motion correction techniques, see the companion paper.

fMRI Datasets

Two separate fMRI datasets were employed in these analyses.

Human Subject Resting ("Noise") fMRI Data

Seventeen healthy, young (ages 22–34) subjects were scanned using the above parameters. Subjects were instructed to relax with eyes open. The room was dim, but lights from the control room were visible. In the creation of statistical maps and assessment of false-positive rates, an "assumed" behavioral task was employed. The behavioral paradigm effect in these analyses is assumed because the subjects were not actually engaged in an experimental paradigm. The temporal structure of the assumed paradigm was designated to be a boxcar with an 80-s period (40 s off/40 s on) which falls within the range of typical task paradigm structures used in our, and other, labs.

Mental Rotation ("Task") fMRI Data

Fourteen healthy, young (ages 20–37), right-handed subjects were scanned during their performance of a mental rotation task. Briefly, each run consisted of 40-s blocks of a visuospatial task alternated with 40-s blocks of a sensorimotor control task repeated four times. Thus, a boxcar with an 80-s period (40 s off/40 s on) represented the ideal neural activity in this experiment. The visuospatial task was adapted from Haxby and colleagues (1991) and involved the matching of a stimulus with a rotated choice. Subjects alternated left and right button presses to empty stimulus presentations during the sensorimotor control task. Stimuli were presented every 3 s. The companion paper describes the derivation of an IRF from a dataset acquired under very similar experimental conditions (except for the rate of stimulus presentation). Importantly, the dataset used here to examine sensitivity was independent from that used to derive the IRF. The functional activity observed within this dataset, and its subsequent interpretation, were virtually identical to that previously reported (Shin et al., 1995) and are not further described here.

Determination of 1/f Model of Intrinsic Temporal Autocorrelation

The comparison paper presents an analysis of the power spectra of noise data. A 1/f model of the noise was adopted [see Eq. (1) of the companion paper]. Spatial smoothing and removal of global signal components could change the nature of the average power spectrum. Average power spectra of the human subject "noise" data were analyzed, as described in the companion paper, with and without spatial smoothing and with and without removal of the global signal to determine which coefficients of the 1/f model would provide the optimum fit.

Creation of Statistical Maps

Both noise and task datasets were subjected to voxel-wise analysis performed using the GLM for autocorrelated observations, as described by Worsley and Friston (1995). Ten versions of SPM analysis that varied in their generation of the voxel-wise univariate statistics were applied to each of the 17, single-subject noise datasets. These different analytical models varied primarily in their inclusion or exclusion of a global signal covariate and in the particular model of the intrinsic temporal autocorrelation used. As described in the companion paper, the K matrix of the modified GLM was used to model both the smoothing kernel and the intrinsic temporal autocorrelation present in the noise. The contents of the K matrix used in the different analyses are briefly described here:

Assumed independence. No temporal smoothing was conducted. The K matrix contained only an impulse.

1/f model. As described above, a curve of 1/frequency shape was found to fit the square root of the average
power spectrum of the null-hypothesis data. A time-
domain form of this representation of the distribution
of power under the null hypothesis was included in the
\( K \) matrix in this analysis.

Our impulse response function. In the previous
paper we described an empirically derived IRF. This
IRF was used to temporally smooth the data in this
analysis and was included in the \( K \) matrix.

Poisson IRF. Friston and colleagues (1994a) pre-
sented an IRF modeled as a Poisson parameter = 8 s
kernel. We were interested in comparing this IRF to our
empirically derived model. The Poisson IRF was used
to temporally smooth the data and was included in the
\( K \) matrix.

Our IRF and 1/f model. This analysis combined
both our empirically derived IRF and the aforementioned 1/f model of intrinsic autocorrelation.

In addition, all models included discrete time sine
and cosine covariates periodic with the length of the
timeseries (160 images) and with fundamental frequen-
cies below that of the fundamental frequency of the
paradigm. Their inclusion high-pass filters the data
(Friston et al., 1995b).

Spatial Smoothing

Single-subject data were smoothed by convolving
with a 3-voxel (11.25 × 11.25 × 15 mm) FWHM 3-D
Gaussian kernel. By the matched-filter theorem (Rosen-
feld and Kak, 1982; cited in Worsley et al., 1996b),
optimal signal detection occurs when the size of the
filter matches the size of the expected signal. The size of
this kernel is based upon our observation of activity
across different subjects and cognitive paradigms, but
still represents a relatively arbitrary selection. While
Worsley and colleagues have presented a general solu-
tion for searching for activation across whole space
(1996b), those calculations are computationally inten-
sive and unnecessary for the present purposes.

Calculation of Map-wise Significance Level

The t field result of Worsley (1994) was used to
calculate a critical t threshold corresponding to a
two-tailed \( \alpha = 0.05 \) (i.e., a one-tailed \( \alpha = 0.025 \)). While
this result has since been superseded by a more general
solution (Worsley et al., 1996a), the more recent method
was not used in these studies as a "continuity correct-
ion" is still required to account for the voxelation of
the dataset at low degrees of smoothness (K. J. Worsley,
personal communication).

Empirical Measurement of Map Smoothness

The assessment of map-wise significance requires an
estimate of the smoothness of the component processes
of the map under study. A measure of map smoothness
is given by Poline et al. (1995),

\[
W = \prod_i \frac{\sum_x t(x)^2}{\sum_x t_i(x)^2},
\]

where \( W \) is the smoothness, \( t(x) \) is the mean corrected
process, and \( t_i(x) \) is the partial derivative of \( t(x) \) in the \( i \)th
dimension.

Our smoothness estimation was based on the ratio of
the variances of the statistical processes and their first
derivatives in the usual way [Eq. (1) above]. However,
this estimation is confounded by increased variance in
the numerator relative to increases in the denominator
due to deviations from the null hypothesis. We there-
fore elected to estimate the smoothness in the null
datasets only; the smoothness of the task maps was
assumed to be equivalent to the mean smoothness
obtained for the noise maps for a given analytical
model. Furthermore, instead of estimating smoothness
for each individual subject, the voxel t scores were
combined across subjects, and the variance was esti-
mated from this combined population. In situations
where a global covariate was not included we antici-
pated that the variance of the t values based on
individual subjects would be less than the variance
based upon these combined data. This is because, as
noted in the companion paper (Zarahn et al., 1997),
spatial dependence is present within fMRI datasets as
reflected in the global signal. Spatially coherent global
activity introduces bias into the t values from each
subject, the direction and magnitude of which is indi-
cated by the correlation of the global signal with the
assumed behavioral wave form, which will in turn
contribute to the variance of the t values when com-
bined over subjects. This is indeed what we found.
Having estimated the smoothness of the t maps we
converted these estimates to the smoothness of the
underlying component fields (Worsley et al., 1992)
using the appropriate coefficient (Holmes, 1994).

Examination of Map-wise False-Positive Rate

The single-subject noise datasets were thresholded
as above and examined for suprathreshold voxels. Any
noise map containing a voxel with a t value exceeding
the critical level, thus rejecting the null hypothesis,
was scored as a false-positive result. As described
above, a total of 10 different treatments of the noise
datasets were conducted. The binomial distribution
was used to assess if the null hypothesis of a map-wise
false-positive rate of \( \leq 0.05 \) could be rejected for the set
of 17 maps obtained under each treatment.
Characterization of the Global Signal

The global signal of each dataset was determined by computing the average time series across all brain voxels. Two aspects of the global signal were examined. First, the global signals were correlated voxel-wise with their smoothed datasets of origin for three randomly selected subjects. These maps represent the spatial distribution of the relative presence of the global signals throughout the volumetric datasets. Second, the relationship of the structure of the global signal with the behavioral paradigm was examined by comparing the correlations of the global signal with the IRF-convolved behavioral reference function. These comparisons were conducted for both the task and the noise datasets and evaluated with an independent samples t test.

Examination of Sensitivity

Those models (of the original 10 described above) which yielded a specificity not less stringent than expected in the noise analysis were used to analyze each of the 14, single-subject task datasets. Because an estimate of the specificity of these models was known, any difference in the amount of activation detected in the task datasets by the different models could likely be attributed to differences in sensitivity. SPMs generated for each of the 14 subjects with each of the models were individually thresholded at their appropriate critical t level and the number of suprathreshold voxels was counted. This measure was compared between maps generated by analyses which differed in a single model component in order to determine the effect of that component. The number of subjects with an increase in the measure was identified, and Clopper–Pearson 95% confidence intervals (CIs) were used to determine if the different treatments had a significant effect.

RESULTS AND DISCUSSION

As several different analyses are reported below, discussion follows each result for the sake of clarity.

Determination of 1/f Model

The null-hypothesis frequency distribution of power within fMRI data can greatly impact inferential statistics. Specifically, accurate characterization of the intrinsic temporal autocorrelation of fMRI data is crucial for obtaining a desired false-positive rate. In the companion paper, we found that the average power spectra of noise data contains increasing power at lower frequencies. A 1/f function was found to reasonably model the shape of the square root of the average power spectrum. As expected, manipulations of the dependent data, such as spatial smoothing and removal of the global signal, had an effect upon the shape of the 1/f function.

Table 1 provides the coefficients of the 1/f equation used to model the intrinsic autocorrelations for spatially smoothed and unsmoothed noise data, with and without removal of the global signal. The effect of the temporal autocorrelation upon inferential statistics does not depend on the absolute magnitude of these values, but upon the relative shape of the curves that they define. Figure 1 provides a graph of the curves defined by these coefficients, each curve scaled to its own k3 value [see Eq. (1) of companion paper], which represents the “white noise” asymptote. Of importance

<table>
<thead>
<tr>
<th>TABLE 1</th>
</tr>
</thead>
<tbody>
<tr>
<td>Coefficients of 1/f Model of Intrinsic Autocorrelation</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Spatial smoothness of dependent data</th>
<th>Removed</th>
<th>Intact</th>
</tr>
</thead>
<tbody>
<tr>
<td>k1, k2, k3</td>
<td>k1, k2, k3</td>
<td></td>
</tr>
<tr>
<td>Smoothed (3-voxel FWHM)</td>
<td>5981, 0.0029, 0.01</td>
<td>2527, 0.01, 0.009</td>
</tr>
<tr>
<td>Unsmoothed</td>
<td>113.56, −0.0006, 1.92</td>
<td>76.98, 0.00023, 1.91</td>
</tr>
</tbody>
</table>

Note. Coefficients of the 1/f model [Eq. (1) of Zarahn et al., 1997] for intrinsic autocorrelations in fMRI “noise” data under different treatments.

FIG. 1. The 1/f curves defined by the coefficients presented in Table 1. These curves were derived by nonlinear least-squares fitting to the square root of the average power spectrum of 13 null-hypothesis scans (see Zarahn et al., 1997, for details). Two manipulations of the raw signal data were undertaken: spatial smoothing and removal of the effect of the global signal. The curves representing the four possible combinations of these manipulations are presented here. These curves have been normalized to their k3 (horizontal asymptote) value in order to facilitate comparison. The arrow on the x axis indicates the frequency of the assumed paradigm (0.0125 Hz).
is the relative power present at the assumed paradigm frequency (indicated by an arrow on the x axis) compared to the power present at higher frequencies. As can be seen, removing the effect of the global signal from the noise data reduces, but does not eliminate, the relative power of the lower frequencies. This indicates that some low-frequency signal fluctuations do not share phase over a large spatial scale. In addition, spatial smoothing increases the relative power of low frequencies (including the assumed paradigm frequency). This suggests that low frequencies share phase over at least a local spatial scale to a greater extent than do high frequencies, in accord with the observations of spatial coherency in Zarahn et al. (1997). The impact of this effect will be an increase in the false-positive rate in spatially smoothed fMRI data if these changes in temporal autocorrelation are not taken into account.

Previous studies have reported the power spectra of noise data which also appear to follow a 1/f function (Weisskoff, 1993, 1996; Boynton et al., 1996, Fig. 9b; Jezzard and Song, 1996, Fig. 7). It should be noted, however, that while there is evidence that the 1/f relationship is common to BOLD fMRI data, the specific values listed in Table 1 are likely idiosyncratic to our laboratory. The application of these analytical techniques at other sites will therefore require an initial examination of human subject noise datasets in order to determine a site-appropriate approximation of intrinsic temporal autocorrelation.

The models of temporal autocorrelation presented here were derived from single fMRI scans. It is also possible to obtain multiple fMRI scans from a single subject and either concatenate (e.g., Breiter et al., 1996) or average (e.g., Buckner et al., 1996) them in order to improve sensitivity. It is important to note that the 1/f pattern observed here in single scans is expected to be present, unchanged, following the combination of multiple scans using either method. Under the null hypothesis, there is no expectation that frequencies will share phase across scans. Combining separate scans will be expected to reduce the absolute power at each frequency. The power at one frequency relative to others will, however, be unchanged, preserving the 1/f relationship. As a result, the autocorrelation functions derived from single, null-hypothesis scans should be suitable for use with data which are composed of multiple scans, concatenated or averaged, derived from single or multiple subjects.

**Empirical Measurement of Map Smoothness**

An estimate of map smoothness, important for assessment of significant changes within a volume, requires either an assumed or a measured estimate of the expected map variance [reflected by the numerator in Eq. (1)]. Table 2 shows the mean map-wise variance measured for the 17 noise putative t statistic maps under different treatments. Note that a Gaussian distribution (which a t statistic approaches at high degrees of freedom) has a variance equal to 1. As can be seen, different models of temporal autocorrelation produced fairly different map-wise variances. Notable is the increase in map-wise variance which appears to accompany the inclusion of a global signal regressor within the model. This effect is perhaps explained by the degree to which the global signal is shared across space, as shown by Fig. 2. As can be seen, the global signal is present in a patchy, yet extensive, distribution over the observed volume. The presence of this shared signal across space will tend to make the t values of all voxels more similar, thus reducing the variance of the voxels within the map. Inclusion of a global signal covariate within the model will reduce the spatial dependence of the dataset, thus increasing the variance of the map.

It is not the case, however, that the variance of the process (i.e., the population of noise maps from which this set of 17 maps is drawn) is increased by inclusion of a global signal covariate. Because the t maps that were created without a global signal covariate have large-scale spatial coherency, their map-wise variances are biased to be smaller than the variance of the underlying process, i.e., the variance of any one realization (map) is smaller than the variance of the underlying process. A method of examining this proposal involves measuring the variance of data combined across multiple realizations. This value, with increasing numbers of realizations, will approach the expected variance of the underlying process. Table 3 presents the variance of the data combined across the 17 subjects for each treatment. As can be seen, the measure of variance obtained from the combined data is not greatly different from the pooled map variance obtained for models which include a global signal regressor (compare Tables 2 and 3). This is the expected result if any one realization accurately represents the entire distribution of the underlying process. Conversely, the variance measures of the com-
TABLE 3
Combined Voxel Variance

<table>
<thead>
<tr>
<th>Model of temporal autocorrelation</th>
<th>Global signal covariate</th>
<th>Present</th>
<th>Absent</th>
</tr>
</thead>
<tbody>
<tr>
<td>Independence assumed</td>
<td>2.51 (1.01)</td>
<td>3.51 (1.01)</td>
<td></td>
</tr>
<tr>
<td>1/f</td>
<td>1.90 (1.02)</td>
<td>1.99 (1.02)</td>
<td></td>
</tr>
<tr>
<td>Our IRF</td>
<td>1.69 (1.04)</td>
<td>2.31 (1.04)</td>
<td></td>
</tr>
<tr>
<td>Poisson IRF</td>
<td>1.44 (1.08)</td>
<td>1.83 (1.08)</td>
<td></td>
</tr>
<tr>
<td>1/f and our IRF</td>
<td>0.89 (1.05)</td>
<td>1.08 (1.05)</td>
<td></td>
</tr>
</tbody>
</table>

Note. Voxel from the 17 maps generated under each treatment were combined and the variance of the test statistic measured. Provided in parentheses is the expected variance of the test distribution for each treatment, given the eff df ($V = df/(df - 2)$).

Combined data are greater than those obtained from individual maps for analyses conducted without a global signal regressor. These results may be attributed to the different temporal autocorrelation structures which are expected under treatments which include or exclude the global signal covariate. Because relative power at the assumed paradigm frequency is greater in datasets which do not remove global signal effects (see Fig. 1), the distribution of the test statistic will be broader in those analyses which ignore intrinsic temporal autocorrelation when the global signal is not removed. Analyses which included a representation of intrinsic temporal autocorrelation (i.e., the 1/f model) account for the effect of the inclusion or exclusion of the global signal covariate. Notably, these analyses yielded combined voxel variance measures which were more similar across models which included or excluded the global signal covariate.

The variance measure reflects the “width” of the distribution present under the null hypothesis. If the model of autocorrelation accurately predicts the power at the task frequency under the null hypothesis, then the distribution should approach the expected variance. As the eff df of a t test increases, the expected distribution should approach a Gaussian distribution and thus possess a variance nearly equal to 1. All treatments produced maps with at least 26 eff df, thus the combined voxel variance of all of the treatments should have been below 1.08. As can be seen from Table 3, the use of a representation of temporal autocorrelation which included our empirically derived IRF and a 1/f model resulted in t map variance measures (obtained from the combined data) which were the closest to their expected levels. The other models of temporal autocorrelation have variances which substantially exceed their expected levels. This suggests that these other models are systematically underrepresenting the power present at the assumed task frequency under the null hypothesis.

The values presented in Table 3 were used in estimating the smoothness of all 17 noise maps generated under each of 10 different treatments. Table 4 shows the mean smoothness estimations (expressed as FWHM).

Modifications of the temporal autocorrelation model had little effect upon the measured map smoothness. This result is consistent with the simple “scaling” effect which the different models have upon the test statistic (as the smoothness measure incorporates an estimate of the variance of the process). Alternatively, inclusion of the global signal regressor reduced the smoothness of every map generated under every treatment. This finding argues that at least some of the spatial coherency which is measured by the global signal can be described by a stationary, continuously differentiable spatial autocovariance function (i.e., is reflected in our measure of smoothness of this type).

In Zarahn et al. (1997), the smoothness of statistical maps generated from unsmoothed, null-hypothesis data was estimated to be less than a single voxel (FWHM = 0.90 pixels). The convolution of this small intrinsic kernel with our exogenous filter (FWHM = 3 voxels) would be expected to lead to maps with a smoothness on the order of that of our kernel. Interestingly, the estimated map smoothness was substantially greater than the width of the kernel with which the data were smoothed (see Table 4).

Finally, it should be noted that the method used for estimation of smoothness employed here is only valid for data collected under the null hypothesis. We make the assumption, however, that data collected during the execution of a behavioral paradigm should have the same expected null-hypothesis smoothness as was calculated for the treatments described here. Thus, the average smoothness values estimated for the null-hypothesis datasets were used in the later analysis of all task statistical maps. Techniques for estimating the smoothness of statistical maps for task data, by analysis of GLM residuals, may provide a more direct measure (Kiebel et al., 1996), though this method relies on the assumption of the same spatial smoothness across all temporal frequencies. The companion paper presents

TABLE 4
Mean Estimated Map Smoothness

<table>
<thead>
<tr>
<th>Model of temporal autocorrelation</th>
<th>Global signal covariate</th>
<th>Present</th>
<th>Absent</th>
</tr>
</thead>
<tbody>
<tr>
<td>Independence assumed</td>
<td>4.45 ± 0.54</td>
<td>5.97 ± 0.96</td>
<td></td>
</tr>
<tr>
<td>1/f</td>
<td>4.37 ± 0.42</td>
<td>5.74 ± 0.84</td>
<td></td>
</tr>
<tr>
<td>Our IRF</td>
<td>4.34 ± 0.45</td>
<td>5.93 ± 0.81</td>
<td></td>
</tr>
<tr>
<td>Poisson IRF</td>
<td>4.44 ± 0.53</td>
<td>5.96 ± 0.97</td>
<td></td>
</tr>
<tr>
<td>1/f and our IRF</td>
<td>4.37 ± 0.48</td>
<td>5.99 ± 0.82</td>
<td></td>
</tr>
</tbody>
</table>

Note. The mean (±SD) FWHM of the 17 “noise” SPMs generated under each treatment. The smoothness of each t map was derived using the combined voxel variance measures provided in Table 3.
FIG. 2. Spatial distribution of the global signal. The global signals from three randomly selected subjects were cross-correlated with their smoothed datasets of origin. Also shown (bottom row) is the result of an equivalent analysis upon computer-generated Gaussian noise, smoothed to FWHM = 3 voxels. The resulting correlation maps are presented here, arbitrarily thresholded at $|R| > 0.2$. As can be seen, the global signal in smooth data is nearly ubiquitous, but varies spatially in its intensity (compare with Fig. 6 of Zarahn et al., 1997). Notably, the extent and magnitude of the global signal is far greater than what would be expected as a simple result of spatial smoothing (compare with bottom row). The spatial distribution of the global signal does not suggest any obvious anatomical/vascular/functional division to the authors.
evidence (the dependence of the spatial coherency measure upon frequency) which at least argues for an empirical validation of this assumption.

**Examination of Map-wise False-Positive Rate**

Sets of statistical maps were generated for each of the 17 smoothed noise datasets and thresholded at their appropriate critical t value. These sets of maps were generated with models which included or excluded the global signal regressor, and with varying models of intrinsic temporal autocorrelation. Because these datasets were collected while subjects rested quietly, any map with a t value exceeding the critical t was designated a “false-positive” map. Table 5 presents the number of false-positive maps observed with each treatment.

Manipulation of the model of intrinsic temporal autocorrelation had a large effect upon the gross number of false-positive maps. Except for the removal of low frequencies, the analysis conducted without a global signal covariate under the assumption of temporal independence is equivalent to a simple correlation between the dependent data and a boxcar reference function, a popular method of analyzing fMRI data (Bandettini et al., 1993). As can be seen, the observed map-wise false-positive rate (94%) is far in excess of what would be expected if the true rate was 0.05. The inclusion of the 1/f model of temporal autocorrelation reduced the number of false-positive maps, but this rate was still unacceptably high. As was argued in the companion paper, subsets of voxels within a dataset may have a distribution of power which differs significantly from the average power spectrum of the data. Some of these populations might have greater power at the paradigm frequency under the null hypothesis than predicted by the 1/f model and thus produce a false-positive map.

The introduction of temporal smoothing in the form of the two IRFs decreased the false-positive rates considerably. The two IRFs differ in their bandwidth, with our empirically derived IRF passing more high frequencies. The lower false-positive rates observed with the Poisson IRF likely reflect the increased stringency (passing fewer high frequencies) of this kernel. Use of both the Poisson IRF and a global signal regressor resulted in a false-positive rate which, by a narrow margin, did not reject the null-hypothesis assumption of a true map-wise false-positive rate of 0.05.

Finally, use of our IRF in conjunction with the 1/f model resulted in no false-positive maps.

Inclusion of the global signal regressor within the model had a variable effect upon the incidence of false-positive maps. Several countervailing effects upon relative significance might accompany the use of a global signal covariate, including changes in temporal autocorrelation (Fig. 1) and map-wise variance (Table 2). Notably, removal of the effect of the global signal can be expected to produce a map which better satisfies one of the assumptions of SPM analysis: stationary spatial autocovariance. As can be seen from Fig. 2, the global signal is present in a patchy distribution over the map. The presence and distribution of this signal might be expected to result in a nonstationary field, complicating estimation of map-wise smoothness. The inclusion of a global signal regressor, therefore, might be expected to better satisfy this assumption.

Overall, failure to accurately model the temporal autocorrelation present within fMRI data under the null hypothesis resulted in an unacceptable map-wise false-positive rate. Because only a single suprathreshold voxel was required for a false-positive map, one might argue that requiring a cluster of suprathreshold voxels (Friston et al., 1994b; Forman et al., 1995; Xiong et al., 1995) would have reduced our false-positive rate. In fact, the excursion sets of the false-positive maps were regularly composed of spatially extended “blobs” of several voxels. In analyses conducted with a size requirement (data not shown), the incidence of false-positive maps actually increased, as the critical t value dropped with the cluster criterion. This effect, while not

---

**TABLE 5**

Map-wise False-Positive Rates

<table>
<thead>
<tr>
<th>Model of temporal autocorrelation</th>
<th>Global signal covariate</th>
</tr>
</thead>
<tbody>
<tr>
<td>Independence assumed</td>
<td>Present</td>
</tr>
<tr>
<td>1/f</td>
<td>15 of 17 (P &lt; 0.0001)</td>
</tr>
<tr>
<td>Our IRF</td>
<td>13 of 17 (P &lt; 0.0001)</td>
</tr>
<tr>
<td>Poisson IRF</td>
<td>8 of 17 (P &lt; 0.0001)</td>
</tr>
<tr>
<td>1/f and our IRF</td>
<td>3 of 17 (P = 0.0503)</td>
</tr>
<tr>
<td>Note. The number of false-positive maps observed under each of 10 different analytical treatments. The number in parentheses beside each proportion is the P value associated with a one-tailed binomial distribution test of a true false-positive rate of 0.05.</td>
<td></td>
</tr>
</tbody>
</table>

---

**FIG. 3.** The variable effect of the global signal regressor upon task data. Shown are four axial slices from three SPMs, thresholded at $\alpha = 0.05$ (two-tailed, corrected for multiple comparisons) analyzed under two different conditions. The top row for each subject (GCV) shows the map generated from a model which did not include a global signal covariate. The bottom row (+GCV) did include this regressor. The three subjects shown were selected to illustrate the range of effects of the global signal covariate upon task-correlated activity. Other model components included the appropriate 1/f model and our IRF in the $K$ matrix and the inclusion of low-frequency confounds. Suprathreshold t scores are present in bilateral parietal and left premotor cortices for all three subjects.
rigorously examined here, might be expected due to the presence of nonstationary spatial autocovariance within the statistical maps.

Of the 10 different models used to analyze the noise data, only three produced false-positive rates which were not significantly greater than 0.05:

- (A) the Poisson IRF with a global signal covariate,
- (B) our IRF with 1/f model with a global signal covariate, and
- (C) our IRF with 1/f model excluding a global signal covariate.

The three models differ along two axes. The first axis is the inclusion or exclusion of a global signal covariate. Two of the models (B and C) contain our 1/f and IRF representations of temporal autocorrelation but differ on the inclusion of this covariate. The second axis is the model of temporal autocorrelation. Two of the models (A and B) contain a global signal covariate, but one uses the Poisson IRF previously presented by Friston and colleagues (1994a), while the other employs our empirically derived IRF along with a 1/f model of temporal autocorrelation. We next examined the application of these sufficiently specific models to fMRI data collected while subjects engaged in a cognitive task. Two features of task datasets were examined. First, the effect of task behavior upon the global signal was explored. Second, the relative sensitivity of these models for task effects was analyzed.

**Global Signal and Task Effects**

Global (whole-brain) changes in signal have long been recognized as potential confounds within PET experiments. While at times a contentious issue, several PET studies have suggested that the local change in blood flow which accompanies a change in neural activity is additive to, and independent of, global changes in blood flow (Friston et al., 1990; Ramsay et al., 1993). This property of global blood flow change has led to its inclusion as a covariate within ANCOVA models used to analyze PET data. It is often assumed that the global cerebral blood flow will not be substantially affected by an ongoing behavioral paradigm. We examined if this assumption is correct for fMRI data.

The global signals from the 14 subjects who performed a visuomotor task were obtained. These were then correlated with the task reference function which had been convolved with our IRF. The mean value of these 14 correlations was $0.10 \pm 0.18$ SD with values ranging from $-0.19$ to $0.46$. The mean of these values was significantly greater than 0 ($t$ test, one-tailed, $P = 0.012$). Global signals were also obtained for the 17 noise datasets. These were correlated with the same reference function. As the subjects did not engage in any experimentally time-locked behavior during the acquisition of the noise data, the reference function represents an experimental artifact. The mean $R$ value of the global signals with the assumed reference function was $0.0097 \pm 0.19$ for the noise data, with values ranging from $-0.36$ to 0.31. These values were not significantly different from 0 ($t$ test, two-tailed, $P = 0.84$). Thus, the structure of the global signal reflects the presence of a time-locked cognitive behavior for this task dataset.

The companion paper considered several possible contributions to the spatial coherency of the global signal. The presence of an ongoing behavioral task introduces potential sources of spatially coherent signal change in addition to those mentioned previously. A change in local BOLD signal, induced by changes in neural activity, is certainly one possible source. As the global signal is a simple average of all the voxels within the brain at each point in time, local, phase-locked signal changes (i.e., “activation”) will be represented within this global measure. Of course, the magnitude of this effect could vary between behavioral paradigms. In particular, if a behavioral task induces roughly equivalent magnitudes of positively and negatively correlated changes or if a task positively activates only a small proportion of the tissue being imaged, the global signal may be less correlated with the task. Finally, task-correlated motion (Hajnal et al., 1994) could contribute to the global signal. It should be noted, however, that motion correction was observed to have a minor effect upon the character of the global signal in Zarahn et al. (1997).

Situations in which global signal change covaries with the ongoing task have been previously considered in the analysis of PET data. The proposal has been made that in these cases the inclusion of a global signal covariate would be expected to “downgrade or eliminate the effect in which one was interested” (Ramsay et al., 1993). More explicitly, because the global signal tends to reflect the behavioral paradigm, its inclusion as an independent variable within the regression model can be expected to explain some amount of task variance within the dependent data. This might reduce sensitivity for positively correlated signal changes, thus decreasing the incidence of positively correlated, suprathreshold voxels. Despite this possibility, the inclusion of the global regressor might still effect a reduction in otherwise unexplained variance large enough to cause a net increase in sensitivity. A second potential effect of a task-correlated global signal covariate is an increase in negatively correlated voxels. Voxel which share some non-task-related components of the global signal, but correlate less well with the behavioral paradigm than the global signal itself, may appear as artifactual negative correlations.
Examination of Relative Sensitivity

In the analyses presented above, three models were found to yield map-wise false-positive rates not significantly greater than the nominal 0.05 level. Here, we examine these sufficiently specific models to see if they can be further differentiated on the basis of a measure of relative sensitivity. As described above, these models differ along two axes: model of temporal autocorrelation and presence of a global signal covariate.

Because no absolute measure of the presence or absence of task-induced signal change within our fMRI datasets is available, we evaluated the effect of changes along these two axes (ignoring any possible interactions) using a relative measure of sensitivity which examined the number of suprathreshold voxels in the SPMs. In addition to sensitivity for the task effect, this measure is impacted by changes in map-wise characteristics, such as smoothness and the eff df of the analysis. This quantification was conducted for both positively and negatively correlated voxels.

Models B and C were compared to examine the effect of the global signal covariate upon relative sensitivity. Inclusion of this regressor increased the number of positively correlated, significant voxels for 9 of the 14 subjects (not significant, 95% CI = [0.39–0.89]). This trend, while not significant, is in the direction opposite of that which may have been predicted from the global signal analysis presented above. Although the global signal covariate explains task variance, these findings suggest that the loss is offset by reductions in the unexplained variance of the model. It should be noted, however, that individual subjects showed considerable variability in the effect of the global signal regressor.

Figure 3 presents the statistical maps from three representative subject analyses, without (top row for each subject) and with (bottom row) the inclusion of a global signal regressor. These subjects represent the range of the effects of the global signal independent variable upon task-correlated activity: reductions, minimal change, and increases, respectively.

The inclusion of the global signal regressor had a marked effect upon the incidence of negatively correlated, suprathreshold voxels ("deactivations"). Thirteen of the 14 subjects showed an increase in negatively correlated voxels for analyses which included the global signal covariate (significant, 95% CI = [0.79–0.99]). This effect is understandable, given the observation that the global signal correlates with the task. This finding suggests that the use of the global signal covariate may decrease specificity (though not strictly statistical specificity) for negatively correlated changes in a way that could not be detected using noise datasets. Thus, for experiments in which the global signal is correlated with the task and included as a covariate, the most conservative option may be to analyze only one tail of the distribution, restricting hypothesis testing to positively correlated changes. As was mentioned above, not all fMRI experiments may face this difficulty. Other datasets, with more closely matched task and control conditions, may not possess enriched power in the global signal at the task frequency. In those cases, inclusion of the global signal covariate would be expected to have primarily beneficial effects.

Models A and B were compared to examine the effect of the model of intrinsic temporal autocorrelation upon relative sensitivity. Eleven of the 14 maps (significant, 95% CI = [0.57–0.99]) demonstrated an increase in significantly correlated voxels when our empirically derived IRF and 1/f model was used, compared to the Poisson IRF. Because of the greater stringency of the Poisson kernel, Model A (Poisson IRF) yielded maps with fewer eff df than Model B (our IRF and 1/f). A conceptually equivalent statement is that the analysis which employed the Poisson IRF possessed lower power. This difference most likely contributed to the observed difference in the sensitivity measure. An additional, but perhaps not unrelated, possibility is that our empirically derived IRF better modeled the evoked hemodynamic responses, resulting in larger parameter estimates and smaller error terms. The measure of sensitivity used here, however, is incapable of separately weighing these alternatives.

CONCLUSIONS

We have considered here several implementations of the GLM for autocorrelated observations presented by Worsley and Friston (1995). Our first purpose was to determine which, if any, of these models provide appropriate map-wise specificity when used with SPM analysis of smoothed BOLD fMRI data. Smoothed fMRI datasets collected under the null hypothesis, i.e., while subjects rested quietly, were used to test the models. These datasets, unlike computer-generated noise, will not necessarily adhere to the assumptions of the analyses. Analyses which employed either a Poisson kernel (parameter = 8 s) or our empirically derived IRF coupled with a 1/f model of intrinsic temporal autocorrelation produced false-positive maps at a rate not significantly greater than 5%.

These sufficiently specific models were then used to analyze fMRI “task” data. It was found that map-wise sensitivity was significantly different between the two models of temporal autocorrelation. Analyses which employed the Poisson IRF displayed reduced sensitivity compared to those which used our empirically derived IRF and 1/f model of temporal autocorrelation. It is not possible to state at this time if the observed differences in sensitivity are attributable entirely to differences in eff df, or if one IRF better models the hemodynamic response than another.

The global signals of smoothed fMRI data were also
examined. It was demonstrated that the global signal reflected the presence of an ongoing cognitive task. Because of this characteristic, it is difficult to provide a conclusive recommendation either for or against the use of a global signal regressor in the analysis of fMRI data. Given the spatial distribution of the global signal (Fig. 2), models which include a global signal covariate are likely to produce maps with a more stationary spatial autocovariance structure, thus better satisfying one of the assumptions of SPM analysis. The inclusion of a global signal covariate, however, had little effect upon the map-wise specificity of the model (Table 5). Finally, trends toward improvements in sensitivity were observed in the presence of a global signal regressor. This last observation suggests that while the global signal regressor explains some amount of the task-related variance, the inclusion of this covariate results in a reduction in unexplained variance for the model as a whole. The consistent increase, however, in negatively correlated ("deactivated") voxels caused by the global signal regressor might argue against its use.

The specific effects (either beneficial or detrimental) of inclusion of a global signal regressor are likely to vary greatly from experiment to experiment and to depend upon the extent to which the global signal reflects the task paradigm. Data collected from paradigms with very different task and control conditions (e.g., motor responses versus rest) might be expected to possess global signals strongly correlated with the task. Indeed, preliminary analyses of other experimental datasets in our lab, beyond the cognitive task dataset presented here, suggest this is the case.

In this paper, and its companion, we have developed an implementation of the GLM of Worsley and Friston (1995) for the analysis of fMRI data which attempts to independently and accurately model both the impulse response function of the system and the temporal autocorrelation present under the null hypothesis. Of the several implementations examined here, this empirically derived model was shown to be both the most specific and the most sensitive.

ACKNOWLEDGMENTS

This work was supported by grants from the National Institutes of Health (NS01762 and AG13483), the McDonnell-Pew Program in Cognitive Neuroscience, and the Charles A. Dana Foundation.

REFERENCES


