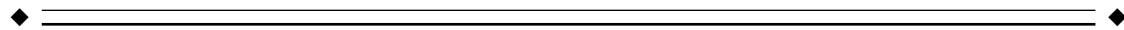


Time Series Analysis in the Time Domain and Resampling Methods for Studies of Functional Magnetic Resonance Brain Imaging

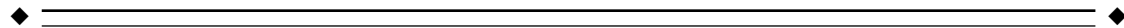
Joseph J. Locascio,* Peggy J. Jennings, Christopher I. Moore,
and Suzanne Corkin

*Department of Brain and Cognitive Sciences and the Clinical Research Center,
Massachusetts Institute of Technology, Cambridge, Massachusetts*



Abstract: Although functional magnetic resonance imaging (fMRI) methods yield rich temporal and spatial data for even a single subject, universally accepted data analysis techniques have not been developed that use all the potential information from fMRI of the brain. Specifically, temporal correlations and confounds are a problem in assessing change within pixels. Spatial correlations across pixels are a problem in determining regions of activation and in correcting for multiple significance tests. We propose methods that address these issues in the analysis of task-related changes in mean signal intensity for individual subjects. Our approach to temporally based problems within pixels is to employ a model based on autoregressive-moving average (ARMA or “Box-Jenkins”) time series methods, which we call CARMA (Contrasts and ARMA). To adjust for performing multiple significance tests across pixels, taking into account between-pixel correlations, we propose adjustment of P values with “resampling methods.” Our objective is to produce two- or three-dimensional brain maps that provide, at each pixel in the map, an estimated P value with absolute meaning. That is, each P value approximates the probability of having obtained by chance the observed signal effect at that pixel, given that the null hypothesis is true. Simulated and real data examples are provided. *Hum. Brain Mapping 5:168–193, 1997.* © 1997 Wiley-Liss, Inc.

Key words: fMRI; ARMA; ARIMA; CARMA; Box-Jenkins; bootstrap; permutation; autoregression; moving-average; autocorrelation



INTRODUCTION

The purpose of this report is to present a method for estimating the statistical significance of task-related changes in functional magnetic resonance imaging (fMRI) signal intensity in the brain for individual human subjects. This method will produce a two- or three-dimensional brain map that provides at each

pixel a P value estimate with absolute meaning as opposed to its being only a relative index of strength of effects. That is, the P at each given pixel approximates the probability of having obtained by chance the putative experimental condition effect at that pixel assuming the null hypothesis is true. The validity of these P values rests on correctly modeling and removing extraneous temporal effects within each pixel, as well as adjusting for the multiple significance tests performed across all pixels taking into account the between-pixel correlational structure. We have attempted to make our presentation accessible to fMRI researchers with background in conventional statistics only.

*Correspondence to: Joseph J. Locascio, Ph.D., Dept. of Brain and Cognitive Sciences, Massachusetts Institute of Technology, E10-003A, Cambridge, MA 02139. E-mail: jomit@wccf.mit.edu
Received for publication 3 March 1997; accepted 7 March 1997

Our objectives are limited. We restrict ourselves to looking primarily for changes in mean levels of signal intensity as opposed to searching for other moments of change or assessing networks of activation [McIntosh and Gonzalez-Lima, 1994], although our methods can be extrapolated to address these other aims. Also, we are concerned with fMRI data analysis, only at the level of assessing the nature and statistical significance of task-related changes in signal. Various components of “extraneous variation” in signal are estimated but are viewed essentially as noise to be removed so that the component of interest, experimental condition effects, can be assessed without bias or obfuscation. Determining the nature and source of the “noise” is important in ultimately establishing the physical and biological underpinnings of fMRI signals, but for our limited purposes, this variability is tentatively relegated to being considered just a statistical nuisance. Further, we describe the analysis of signal changes for individual subjects only, but these methods can be augmented with more conventional multivariate statistical methods to provide between-subject or group tests. In addition, we do not explicitly deal with gross signal contaminants such as subject movement, vascular effects, and global blood oxygenation. We assume that correcting for these problems can be accomplished antecedent to or as a facet of application of our methods. For example, steady linear or curvilinear signal change due to subject movement will be detected and removed as a byproduct of our technique. Last, we present our data analysis methods as a heuristic general approach rather than a finalized, specialized, end-product with detailed algorithms. We invite criticisms, refinements, additions, and modifications of our core strategy to make it more suitable to specific objectives. We especially encourage further testing of our methods with real and simulated data beyond what we have done.

CRITIQUE OF EXISTING fMRI DATA ANALYSIS METHODS

Analysis methods for fMRI data are still being developed. They vary widely in terms of their objectives and approach [Bandettini et al., 1993; Binder and Rao, 1994; Le Bihan and Karni, 1995]. The focus here is on methods for individual subjects. A commonly employed method of assessing the statistical significance of task-related effects at each of many pixels is to perform, at each pixel, conventional parametric (e.g., t-tests, analysis of variance [ANOVA], F tests) or nonparametric (e.g., Kolomogorov-Smirnov [KS], Wilcoxon-rank sum test [WRS]) [Siegel and Castellan,

1988] tests of the significance of group differences. Each signal value is treated as an observation, and the various task or control condition epochs are the groups whose distributions of signal values are being compared. For example, the KS method determines the value where two cumulative distribution functions of signal values differ most and computes the probability of a difference that large or larger under the null hypothesis. After these tests are run, a stringent cutoff for significance (e.g., $P = .0001$) is then applied across pixels to compensate for the fact that multiple testing inflates type I error (probability of rejecting the null hypothesis when it is true) [Kim et al., 1993a,b, 1994; Tyszka et al., 1994]. These methods assume that observations within pixels, i.e., the signal values, are independent of each other, which they generally are not; those close in time can typically be expected to be more highly correlated than those farther apart. Further, task-related effects are confounded with linear or curvilinear trends across time that are unrelated to the tasks. The P value cutoff used to correct for multiple testing is often arbitrary, not based on probability theory. Sometimes a Bonferroni correction for multiple testing is applied [Myers, 1979], which multiplies the P value at each pixel by the number of pixels analyzed, but this correction tends to be overly conservative. It assumes erroneously that no between-pixel correlations exist.

Bandettini et al. [1993] and others [DeYoe et al., 1994; Disbrow et al., 1995] employ experimental paradigms consisting of sequences of on/off conditions and then find pixels whose patterns of signal changes across time are relatively congruent with the condition patterns. Such methods can provide powerful evidence of condition effects, but data analysis here is wedded to experimental design. These methods do not constitute an independent data analysis technique that can be applied post hoc to any arbitrary study paradigm, and especially not to one with a small number of lengthy experimental and control conditions. Designs with many repetitions of short on/off conditions would in many cases appear to be better suited to sensory/motor studies where the presence or absence of the “condition,” i.e., stimuli or movement, can be brief and clearly demarcated. In contrast, for cognitive studies, a smaller number of longer conditions may often be more appropriate because the function being assessed may be less clearly demarcated in time.

Time series analyses of fMRI data are sometimes employed, but typically these analyses are within the “frequency domain,” i.e., they decompose a time series of fMRI signals into component cycles of various frequencies [Bandettini et al., 1993]. A power spectrum

indicates the relative prominence of the component frequencies. While these methods have value for analyzing physical and biological constituent cycles in the signal, they are not well suited to the analysis of data from studies in which sequences of experimental conditions do not resemble any kind of cycles or periodicity.

Friston et al. [1991, 1994b] have developed methods for addressing the multiple significance test problem and for assessing the spatial extent of activation. These methods, however, have assumptions that may not always be met [Holmes et al., 1996].

Many other methods for analyzing fMRI data exist, each with its particular limitations and problems. For example, use of principal components analysis is computationally problematic because of the large number of pixels relative to images, and cluster analysis methods do not have universally accepted statistical significance tests associated with them [Aldenderfer and Blashfield, 1984].

A METHOD FOR ASSESSING THE STATISTICAL SIGNIFICANCE OF TASK-RELATED SIGNAL CHANGES

Our fMRI data analysis strategy has two major components. The first assesses experimental condition effects at the individual pixel level employing a model based on autoregressive-moving average (ARMA) time series analysis to overcome temporal problems [Box et al., 1994; Box and Jenkins, 1976; McCleary and Hay, 1980; McDowall et al., 1980; Gottman, 1981]. (ARMA methods were developed largely for econometrics and the social sciences, but are applicable to fMRI.) The second component employs computationally intensive resampling methods to adjust for multiple statistical significance tests across pixels [Westfall and Young, 1993; Good, 1994; Efron, 1982; Mooney and Duval, 1993]. Variations of ARMA and related time series methods have been developed and tested for fMRI analysis by Bullmore et al. [1996], Tagaris et al. [1995], Friston et al. [1994a, 1995a], and Worsley and Friston [1995] with some success. Resampling methods are also beginning to be employed for adjustment of multiple significance tests in functional brain imaging data (L.J. Wei, statistician, Harvard School of Public Health, personal communication, 1995) [Holmes et al., 1995, 1996]. Bullmore et al. [1996] have experimented with a method of fMRI data analysis that is similar to what we are proposing (see Discussion). Although variations of components of our method have been developed independently and have recently been introduced into the functional neuroimage literature, to our

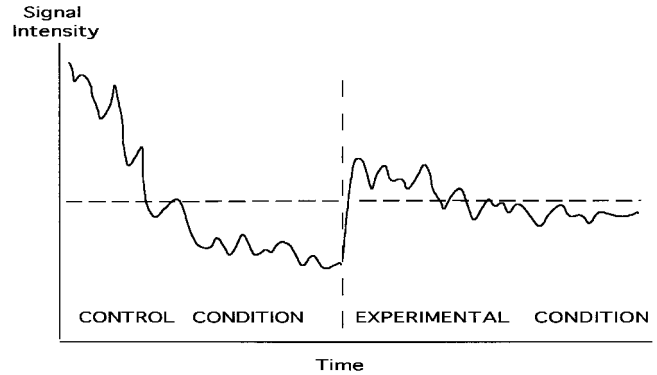


Figure 1.

A temporal trend obscuring task-related effects within a pixel. The comparison of the mean signal intensity (dashed horizontal lines) in the experimental versus control condition is confounded by the presence of a curvilinear time trend.

knowledge, no one has employed the specific variations and combination of linear/nonlinear models, ARMA, and resampling methods that we are proposing for analysis of fMRI data.

Problems

In attempting to perform statistical significance tests of task-associated activation at each pixel of a brain image, two major problem areas need to be addressed: (1) Within-pixel temporal confounds and correlations need to be estimated and removed, and (2) the significance test for the task effect at each pixel needs to be adjusted for the multiplicity of these tests being conducted across all pixels, especially taking into account correlations between pixels, which the Bonferroni method incorrectly assumes to be nonexistent.

Temporal related problems within pixels

Figure 1 simulates a typical time trend of fMRI signal intensity values for an individual pixel. The first half of the images were taken during a control condition, and the second half during an experimental condition. The approximate mean signal values for the control condition and for the experimental condition are indicated, respectively, by the horizontal dashed lines. We need to determine whether the mean signals for the two conditions are different beyond what is likely to be due to chance, and apart from confounds extraneous to true condition effects. The mean for the experimental condition can be seen to be approximately the same as that for the control. A decelerating temporal trend is also evident within conditions, how-

ever. This temporal trend may be independent of condition effects and, if so, is creating an artificially high mean for the control condition. It appears that a positive effect of the experimental task is superimposed on the temporal trend, but this task effect is swamped by the time trend resulting in a misleading equality of the means for the two conditions. Confounding temporal trends can be linear or curvilinear (Fig. 2A), and they can hide condition effects that are there or masquerade as spurious condition effects. They may arise for a multitude of reasons (e.g., subject motion) independent of task effects of interest. Their presence needs to be assessed and taken into account in fMRI data analysis.

Another temporal related problem is *autocorrelation*, the correlation of temporally proximate fMRI signal values. Figure 2B shows a simulated series of fMRI signal values that follows a “white noise” pattern. White noise values are independent from one time point assessment to the next; the value of the signal at time point t has no correlation with the value at time point $t + 1$. Such a data structure for error variability would permit testing experimental condition effects with conventional significance tests such as t-tests and ANOVA (other assumptions being met) or the KS or WRS tests. However, there is likely to be autocorrelation, which violates the assumption of independence of observations of these and other parametric and nonparametric tests. Positive autocorrelation is said to occur if values in close temporal proximity are positively correlated; e.g., if a value is relatively high (low) at time point t , it is likely to be high (low) at time point $t + 1$. Positive autocorrelation is sometimes apparent as a kind of “snaking” process in the data (Fig. 2C); high values remain so for a while as do low ones. Positive autocorrelation may be due to carryover effects from one time point to the next or to choosing time intervals that are smaller than actual temporal changes. When values are negatively autocorrelated, a high (low) value tends to be followed by a low (high) value (the “bouncing” pattern in Fig. 2D). (But it should be noted that the presence of autocorrelation and its nature are usually not visually discernible in a plot of the data against time). The problems caused by autocorrelation are in addition to those due to the nonstochastically based temporal trends discussed above.

The problem of multiple significance tests and between-pixel correlations

Conventionally employed statistical significance tests assume that a test is performed in isolation. When a

“family” of tests is performed, however, P values must be adjusted “familywise” [Hochberg and Tamhane, 1987; Toothaker, 1993]. The probability of drawing an ace of spades from a deck of shuffled playing cards is less than .02, but if one draws a card from each of 100 separate shuffled decks of cards, the probability is about .86 that the ace of spades will be chosen at least once across all the decks. One cannot point to, e.g., the two decks where this card turned up and claim that the probability of that happening was less than .02. In performing a statistical significance test for activation at a pixel in an fMRI image, some adjustment must be made for the multitude of these tests run across all the pixels in the image. The decks of cards are metaphors for the pixels. In performing n significance tests each with a type I error rate of α , the probability of finding one or more significant results, under a global null hypothesis, is greater than α (or equal to α only if all the tests are perfectly positively correlated). If all tests are independent, the probability is $1 - (1 - \alpha)^n$, but lower to the degree that the tests are positively correlated. An adjustment made on this basis assuming independence was developed by Sidak [1967]. A simpler, slightly more conservative approximation of the Sidak correction, called the Bonferroni method [Myers, 1979], is to multiply all obtained P values by n (or equivalently *divide* the significance cutoff by n). If these methods are applied to significance tests run on each of many pixels, they can generally be expected to be conservatively biased because pixels can usually be assumed to be positively intercorrelated (to varying degrees for different sets of pixels within an image) even under null task effects. Correlations may be the result of physiologically based associations, close spatial proximity, smoothing, or image resolution that is finer than areas of (non-task-related) activation.

Proposed solutions

ARMA based approach to within-pixel temporal problems

ARMA time series analysis [McCleary and Hay, 1980; McDowall et al., 1980; Pankratz, 1983; Gottman, 1981], often referred to as Box-Jenkins analysis [Box and Jenkins, 1976; Box et al., 1994; Box and Tiao, 1975], analyzes data for a variable collected at regular intervals of time by assessing and attempting to model the nature of correlations across error terms from time point to time point, i.e., the autocorrelation. These correlated components are then removed, permitting valid significance tests of the effects of “exogenous” variables on the dependent variable of interest, in this

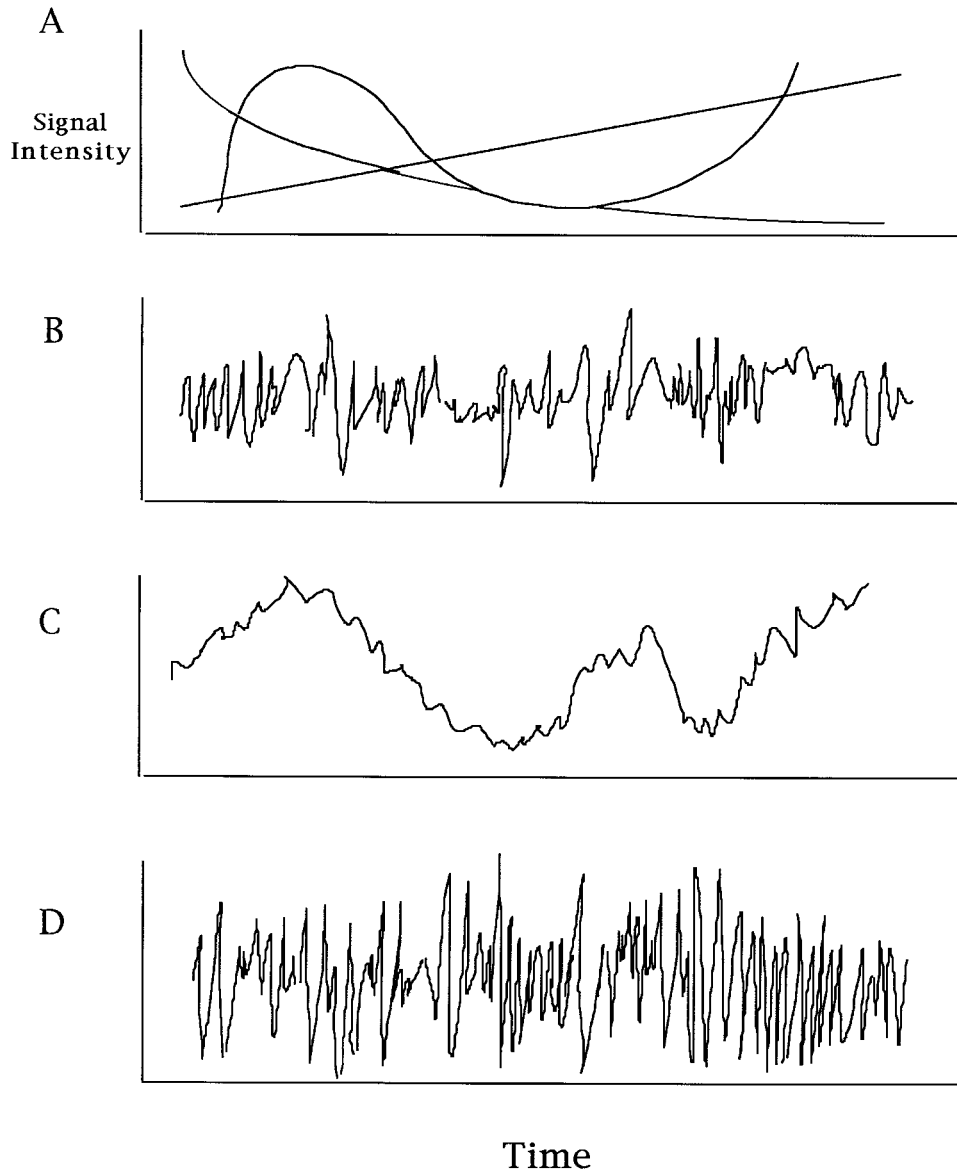


Figure 2.

Time trends, white noise, and autocorrelation. **A:** Nonstochastic linear, curvilinear, and cyclic trends may obscure or bias effects of interest. **B:** Conventional significance tests assume error variability is white noise, i.e., values are independent from time point to time point. However, there may be positive (**C**) or negative (**D**) autocorrelation, i.e., correlation of temporally proximate signals.

case, fMRI signal intensity. The exogenous variables are independent or predictor variables and can be nonstochastically based linear or curvilinear functions of time, indicator variables indexing experimental conditions, or random variables of interest co-measured with the dependent variable across time (e.g., a continuous numeric index of stimulus intensity or of a biological variable like blood pressure). In removing

the autocorrelated components of errors, residuals (for the model at large) become “white noise.” The variance of this white noise is used as the base estimate of random error variability against which relations of exogenous and dependent variables are tested for statistical significance. Uncorrected error variance estimates based on the original autocorrelated data are biased, and significance tests employing them are invalid.

ARMA and related methods are sometimes referred to as time series analyses in the *time domain* as opposed to methods such as “spectral analysis,” which decompose the time series into component cycles of varying frequency, which are referred to as time series analyses in the *frequency domain* [Gottman, 1981; SAS/ETS (Statistical Analysis System/Econometric Time Series) Users’ Guide, 1993, the Spectra Procedure, pp. 749–770]. Analyses in these respective domains have complementary strengths and weaknesses (although the two approaches are in one sense the same because parameters derived from analyses in one domain can be mathematically related to those from the other [Gottman, 1981]). Methods in the frequency domain are better suited to searching for physically or biologically based periodicities in the fMRI time series, although ARMA also has a limited capability to detect periodicities. ARMA-based methods are better suited to assessing effects associated with noncyclical experimental paradigms while removing autocorrelation and confounding temporal trends unrelated to experimental conditions.

In an ARMA model, the nature of autocorrelation is assumed to be of two possible types: “autoregressive” (AR) or “moving average” (MA). In the autoregressive model, an observed value of the dependent variable at a given time is considered to be the sum of fractions of the values of the same variable or residual terms at various past times, as well as a current perturbation of random white noise (aside from the effects of exogenous variables if any). The number of immediately previous values of which the dependent variable is a function is the “order” of the autoregression. Autoregression of order p is usually denoted AR p . Thus, the model for a first-order autoregression process or AR1 is

$$Y_t = \phi_1 Y_{t-1} + e_t$$

where Y_t is the fMRI signal intensity at time “ t ”; ϕ_1 is the first-order autoregression coefficient ($-1 < \phi_1 < 1$); e_t is a random “shock” at time “ t ”, i.e., white noise; $E(e_t) = 0$ (the “expected value” or mean of the e_t is zero); $\text{corr}(e_t, e_{t-k}) = 0$ ($k \neq 0$) (temporal independence); $\sigma_{e_t}^2$ is constant (add more $\phi_k Y_{t-k}$ terms for higher orders). In a moving average process, each observed value of the dependent variable is equal to a current white noise random perturbation plus the sum of fractions of such perturbations from past time points (aside from effects of exogenous variables). The number of contributing previous time points is the order of the moving average process. A moving aver-

age process of order q is indicated MA q . Thus, a first-order moving average or MA1 process is

$$Y_t = e_t + \theta_1 e_{t-1}$$

where θ_1 is the first-order moving average coefficient ($-1 < \theta_1 < 1$); other terms as above (add more $\theta_k e_{t-k}$ terms for higher orders).

The nature of the autocorrelation and its order is assessed by examining correlations of values of the dependent variable at various lags. For example, the autocorrelation at lag 1 is computed as a Pearson correlation coefficient of all values vs. their immediate respective predecessors. The one for lag 2 is computed for all values vs. their respective predecessors two time points back, and so on. The autocorrelations for a number of different lags (typically up to 20 or more) are displayed together in a “correlogram.” Similarly, a partial autocorrelation correlogram indicates the correlation of observations a given number of lags apart with all intervening time point values held constant statistically (partial correlation). Autoregression and moving average processes of various orders leave characteristic patterns of correlations in correlograms, and the empirical patterns are used as an indication of the process that generated the data and of its order. (Spectral analysis can also be used as an adjunct in identifying ARMA models, e.g., a positive AR1 model may show up as a low-frequency peak in a power spectrum [Gottman, 1981].) Likely models for the autocorrelation are tested for fit by seeing whether coefficients for the hypothesized autoregressive or moving average process are statistically significant and whether the residuals with the modeled autocorrelation components removed are not significantly different from white noise using a test developed by Ljung and Box [1978]. Specifically, the value from the formula below has an approximate chi-square distribution if the residuals are white noise:

$$n(n+2) \sum_{k=1}^m [(r_k)^2 / (n-k)] \quad \text{with} \quad \text{df} = m - p - q$$

where r_k is the autocorrelation of the residuals at lag k ; n is the number of time points; m is the number of lags across which the test is made; p and q are the order of AR and MA in the model, respectively. The correlograms for the residuals should also show no significant autocorrelations or partial autocorrelations if the residuals are white noise. Usually the most parsimonious model with acceptable fit is retained. Occasionally,

time series are best fit by models that combine AR and MA terms (but see “parameter redundancy” in Box and Jenkins [1976], pp. 248–250, and McCleary and Hay [1980], pp. 64–66). Also, periodic variation can be modeled with high-order AR or MA terms (or “seasonal” ARMA models) [McCleary and Hay, 1980; McDowall et al., 1980; Pankratz, 1983, 1991].

Our ARMA-based model is as follows:

$$Y_t = \text{intercept} + \sum a_i C_{it} + b_1 \text{time} + b_2(\text{time})^2 + (\theta(B)/\phi(B))e_t$$

where $\sum a_i C_{it}$ is, at time “t”, the sum (across contrasts “i”) of the product of each indicator variable (“ C_i ”) and its coefficient (“ a_i ”), for a full set of orthogonal contrasts of the experimental and control conditions; time is a simple ordinal indicator of the successive images, i.e., 1, 2, 3, . . . (assumes the images are equally spaced in time); b_1 , b_2 are coefficients for the nonstochastic linear and quadratic effects of time (higher-order polynomial functions of time can be added); B is a backshift operator, i.e., $BX_t = X_{t-1}$; $\theta(B)$ is the moving average operator, represented as a polynomial in the backshift operator: $\theta(B) = 1 - \theta_1(B) - \theta_2(B)^2 - \dots - \theta_q(B)^q$ where there are q moving average terms for MAq; $\phi(B)$ is the autoregressive operator, represented as a polynomial in the backshift operator: $\phi(B) = 1 - \phi_1(B) - \phi_2(B)^2 - \dots - \phi_p(B)^p$ where there are p autoregressive terms for ARp; e_t is white noise at time “t”; intercept is a constant scaling term. Contrasts are comparisons among means of conditions or combinations of conditions [Myers, 1979]. There must be $n - 1$ contrast indicator variables for n experimental and control conditions [Cohen and Cohen, 1983] where repetitions of conditions are considered separate conditions. (The indicator variables have values of negative or positive 1, fractions, or 0, that provide a “contrast” of interest.) At least one of the contrasts among conditions should be of substantive interest, with the rest completing the full set of orthogonal contrasts. This full set reproduces all the respective condition means as predicted values (adjusted for any modeled polynomial time trends and autocorrelation). If the set of orthogonal contrasts is incomplete (less than the number of conditions minus 1), some condition effects may be lumped into the residual. These effects may then masquerade as autocorrelation (Elkan Halpern, Center for Imaging and Pharmaceutical Research, Massachusetts General Hospital, MGH, Charlestown, MA, personal communication, 1995). The statistical significance of the coefficient of a given contrast indicator demonstrates the significance of the corresponding

contrast of conditions, corrected for any modeled nonstochastic time trends or autocorrelation that may exist. The significance of the coefficients for one or more of the “time” variables indicates a nonstochastic linear or higher-order polynomial effect of time, independent of any experimental condition effects or autocorrelation. The AR or MA terms are assessed on the basis of residuals from the rest of the model. We will henceforth designate the above model as CARMA (Contrasts and ARMA).

Because the backshift operator may be unfamiliar to those new to Box-Jenkins time series methods, variations of the above formula are shown below after carrying out the backshift. A CARMA model with a quadratic time trend and MA1 autocorrelation would be

$$Y_t = \text{intercept} + \sum a_i C_{it} + b_1 \text{time} + b_2(\text{time})^2 + e_t - \theta_1 e_{t-1}$$

The same with AR1 instead of MA1 would be

$$Y_t = \text{intercept} + \sum a_i C_{it} + b_1 \text{time} + b_2(\text{time})^2 + e_t + \phi_1 [Y_{t-1} - (\text{intercept} + \sum a_i C_{it-1} + b_1(\text{time} - 1) + b_2(\text{time} - 1)^2)].$$

The autoregressive component is ϕ_1 multiplied by the part of the previous signal value that is not due to the nonstochastic predictors in the model (intercept, contrasts, time trends).

Different algorithms for applying the CARMA model can be used. We have used SAS statistical software [SAS/ETS User’s Guide: Version 6, 1993; the ARIMA Procedure, pp. 99–182] with “conditional least-squares” solutions for coefficients (conditional on assumed values prior to onset of the time series). (Other major statistical software packages also have ARMA programs, e.g., BMDP [BMDP Statistical Software Manual, 1990a,b].) SAS also provides unconditional least-squares and maximum-likelihood estimation methods. An iterative nonlinear optimization algorithm called “Marquardt’s method” is used by SAS to estimate parameters of the intrinsically nonlinear general CARMA model; this algorithm usually shows rapid convergence for ARMA models. Estimates of error variance and of standard errors of parameters, as well as t-tests, are corrected for autocorrelation. Details on nonlinear optimization methods, as well as estimation of parameters and their standard errors for CARMA, are complex and can be found in Marquardt [1963], Box et al. [1976, 1994], Pankratz [1983, 1991], and McCleary and Hay [1980]. Further, in addition to SAS

and BMDP, there are a wide variety of software subroutines, stand-alone packages, and comprehensive statistical packages that do ARMA time series analyses or components of ARMA. Much of this software is discussed and compared in Cromwell et al. [1994a], McDowall et al. [1980], McCleary and Hay [1980], Kim and Trivedi [1994], and is cited in the Appendix below (some of it was developed for econometrics and the social sciences but is applicable to fMRI). If one makes use of available software, ARMA analyses can be performed without the necessity of understanding all the technical details of parameter estimation.

In running CARMA separately for each of many pixels, the above-mentioned procedure of checking correlograms to determine the best-fitting autocorrelation model during repeated CARMA runs is not practical. Instead, we have used an algorithm in which a model is run that contains as many autoregressive or, separately, as many moving average terms as are likely to be operative (we have used up to the third order for each for tests of many pixels). Nonsignificant higher-order terms are then progressively removed until only significant autocorrelative terms remain, if any, analogous to a stepwise regression method (see Box and Jenkins [1976; pp. 220–224] regarding the use of “overfitting” in estimating parameters). The residuals from the final model must also pass the white noise test, indicating that they are not significantly different from white noise. We have required the test be passed 3 times at each pixel, at up to $k = 6, 12,$ and 18 lags. Typically, we have discarded a small proportion of pixels (less than 5%) that do not pass the white noise test with or without any set of the autocorrelative terms being considered. (Periodic effects may be occurring for these pixels. If there are not many such pixels, it may be possible to assess and model them with high-order ARMA terms.) In the uncommon event that AR and MA models of the same order meet fit criteria, we have retained the AR for its greater conceptual and computational simplicity. In order to increase processing speed, we have sometimes used an algorithm that retains a set of autoregressive or moving average terms in the model, regardless of which or how many are statistically significant (the white noise test must also be passed). The assumption here is that nonsignificant coefficients will not greatly affect estimates of other effects. Reduction in power due to lowered degrees of freedom resulting from inclusion of the nonsignificant terms should be minimal in a long term series. Regardless of which algorithm is used, it is important that all the experimental condition contrasts, the time trend effects, and the autocorrelative components be estimated simultaneously so that their respective effects

are not incorrectly absorbed into each other. For example, nonstochastic polynomial temporal trends in the data that are not separately modeled can produce erroneous estimates of autocorrelative coefficients (coefficients are said to be outside the bounds of “stationarity” or “invertibility”). A method of subtracting temporally adjacent scores from each other called “differencing” is sometimes employed to meet the ARMA assumption of “stationarity” [see McCleary and Hay, 1980; Gottman, 1981; McDowall et al., 1980], but including terms for linear and curvilinear effects of time should obviate the need or lessen the problem. We feel that modeling is a more analytical and flexible way of dealing with temporal trends than is differencing. (The term “ARIMA” is often used to denote ARMA models with differencing. The “i” stands for “integrated” extraneous terms that are presumably removed by differencing.)

Although we generally view autocorrelative components and nonstochastic temporal trends as just nuisance factors to be removed so as to be able to clearly assess task-related effects, they could have substantive importance in their own right. For example, estimated autoregressive or moving average terms may indicate carryover and persistence of neuronal activation, cyclical events, inappropriateness of time intervals between images, rebounding phenomena (for negative coefficients), or possibly characteristics or artifacts of the measurement process. Time trends may indicate head motion artifacts. Brain maps displaying temporal trends, i.e., showing at which pixels linear time trends are occurring, at which there are quadratic effects, etc., can be produced as byproducts of assessing task-related effects, and similarly for maps of autocorrelative effects (showing where the error is AR1, AR2, MA1, etc.).

The formulas above assume that task-related effects will be modeled as simple step functions, i.e., a task will uniformly elevate or lower the fMRI signal value throughout the task condition epoch. However, CARMA can model other kinds of effects of task performance or stimuli, for example, effects where a condition causes a gradual increase in activation to an asymptote, or a sharp increase with gradual decline, a short pulse, a steady “ramp” effect, and so on. Thus, hemodynamically mediated delays between onset of a task or stimulus condition and change in fMRI signal can be explored and modeled. Also, interactions between task conditions and time trends can reveal effects of tasks on temporal change in the signal. Residuals from some models may deviate from white noise because complex condition effects were inappropriately modeled with simple step functions. Or com-

plex condition effects could masquerade as independent temporal trends. It may be necessary to assess the best fit among alternative models, each with different combinations of condition, temporal, and autocorrelative specifications, where all the models meet acceptable fit criteria [McCleary and Hay, 1980, pp. 100–101].

McCleary and Hay [1980] and McDowall et al. [1980] suggest some ARMA-based models that may be suitable for complex activation effects [see also Pankratz, 1991]. The models are presented in the context of a simple design with a control condition followed by an experimental condition but can be adapted to repetitions of conditions or multiple distinct conditions. To model an activation effect that *gradually increases to an asymptote*, replace the condition contrast terms in the CARMA equation with

$$[\omega/(1 - \delta B)]X_t$$

where X_t is the value at time t of a dummy coded variable that = 1 during the experimental condition epoch and 0 otherwise; B is the backshift operator (see above);

$$\omega > 0; \quad 0 < \delta < +1;$$

Working out the backshift operator shows that at the n th timepoint into the experimental condition, the above equals

$$\sum_{i=0}^{n-1} \delta^i \omega$$

a value that increases by smaller and smaller amounts as one goes further and further into the experimental condition, to an asymptotic change of $\omega/(1-\delta)$. If a point is reached where the control condition resumes again, the value begins to decrease gradually.

The same formula can be used to model activation that has an *abrupt increase that gradually declines*, by setting the dummy variable X equal to 1 only at the inception of the experimental condition epoch and 0 otherwise. A *transient pulse* of activation would be modeled as a simple linear function of this same dummy variable (ωX_t). For all these models, ω indexes the immediate effect of a condition, if any, and δ is a rate parameter, reflecting increase or slowness of decrease of effects over time, if any.

The above three models can be tested sequentially to see which shows significant parameters and the best fit. Models that are higher order in the backshift operator, as well as combinations of the above, can be

fit [McCleary and Hay, 1980]. Additional models can be developed tailored to other kinds of activation effects considered possible. Estimated parameter values for the models could provide information on the nature and length of any hemodynamic delay. Even if the same model is run at each pixel, spatially variable hemodynamic delay could be indicated by parameter estimates that vary across pixels.

Resampling approach to multiple test correction

The method we describe below is similar to that presented by Holmes et al. [1996], although their discussion is primarily in the context of PET (positron emission tomography) studies, whereas our application is fMRI.

Conventional tables of inferential statistics (e.g., t or F) and their associated P values assume that only an individual significance test is performed. In fMRI studies, multiple significance tests are performed, one at each of many pixels. Thus, what is needed is a table of the distribution of the statistic value indicating the strongest effect (e.g., maximum t or F) or its corresponding P value, the *minimum P* value, across *many* tests under a global null hypothesis, where these tests have a correlational structure like that of the pixels being analyzed. Such a table is appropriately suited to the researcher's search for the area of greatest activation in a brain image, i.e., the statistic value indicating the strongest effect or the lowest P across all pixels. (Dealing with P values instead of other statistics as the index of activation simplifies algorithms for two-tailed tests; very negative or positive effects both have low P s.) Effectively, an empirically based estimate of the necessary table can be derived through computer-intensive resampling methods [SAS Technical Report P-229, 1992, the Multtest Procedure, pp. 369–405; Westfall and Young, 1993]. The algorithm for this method is as follows: The post-CARMA white noise residual at each of all pixels at the first time point are reassigned in unison randomly to an experimental or control condition. The same is done for the second time point, third, and so on, until all time points are randomly reshuffled (the same number of time points is assigned to each condition as occurred originally). The same appropriate parametric significance test for the same condition effect of interest is then computed at each pixel (e.g., a t -test for a relevant partial regression coefficient. The temporal dimension of the data has been lost and only condition contrasts are tested.) The minimum P value across all the pixels is recorded. This entire process is repeated many times, typically 10,000 times, and a distribution of these

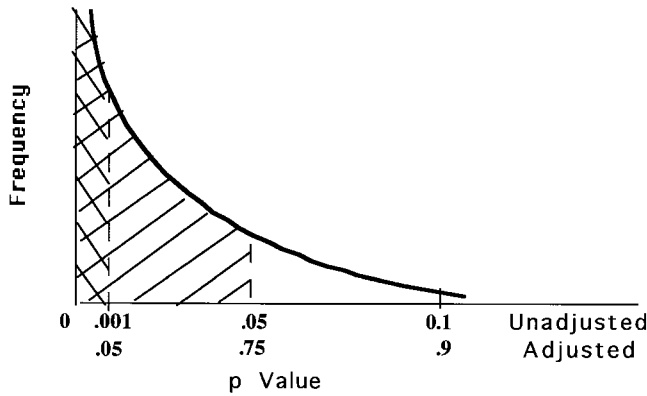


Figure 3.

Distribution of the minimum P value across all pixels under the null hypothesis, estimated with resampling methods. The empirical minimum P value is adjusted to the proportion of times it or one lower in value occurs in this distribution. For example, an obtained minimum P of .05 is adjusted to .75 because 75% of the time, the minimum P value can be expected to be .05 or lower under the null hypothesis.

minimum P values is compiled. Because values from all pixels at a time point are yoked together during reshuffling, the between pixel correlational structure is retained, i.e., the distribution of minimum P values compiled is based on between-test correlations equal to the between-pixel correlations. If a relatively low P value occurred by chance at one pixel at one of the 10,000 samplings, another pixel with which it is correlated would have also tended to have a low value at that sampling. (Even negative correlations among pixels are taken into account by this procedure should they exist.) This distribution of minimum P values is one for which the null hypothesis is true at every pixel because values (which were white noise to begin with) were randomly assigned to experimental and control conditions.

Once this distribution of minimum P values is produced, the smallest P value across pixels from the CARMA analyses of the original raw data is adjusted to the proportion of times it (or a lower value) occurs in the minimum P value distribution (Fig. 3). In order to adjust the other P values from the original CARMA tests, a *step-down* method is applied. The second smallest P value is tested against an analogous resampling derived minimum P value distribution that is based on all pixels, except the one that had the smallest raw value. Similarly, the third smallest P value is adjusted with a distribution based on all pixels but the two most significant, and so on. This step-down method provides a more powerful test than using the same minimum P value distribution to adjust all P

values from the CARMA runs on the raw data. In a sense, pixels already verified by the resampling method as having significant (therefore presumably real) effects are progressively excluded from the pool for which the global null hypothesis is assumed. An efficient computer algorithm can be used that makes one pass through all the original raw P values for each resampling set of P values so that only one sample of 10,000 needs to be drawn. (Monotonicity in the P values must also be maintained; see Holmes et al. [1996], and Westfall and Young [1993], for detailed algorithms for these methods.)

Variations of this resampling method use a bootstrap algorithm, whereby time points are randomly selected with replacement during the random reassignment process, or a permutation algorithm, where each time point is randomly reassigned once and only once, the method described above. We have found results with these two variations to be nearly identical with the permutation showing slightly greater power. Figure 4 displays the results of employing these methods on a set of 50 pixels with simulated signal values. The raw P values are liberal, whereas the Bonferroni method is conservative, producing a high proportion of adjusted P values that are equal to 1. The step-down bootstrap and permutation tests produce more realistic intermediate values.

EXAMPLES

Three real data applications of our data analysis methods follow. Analysis is emphasized and substantive information is minimal. (In each of the three studies, echo-planar MR images were collected using a 1.5 Tesla GE Signa scanner with a receive-only RF quadrature head volume coil and an asymmetric spin echo sequence.)

CARMA analysis of one time series (motor sequencing activates the supplementary motor area)

In this example, CARMA was used to analyze an individual time series that was computed as the average of fMRI signals from 15 pixels covering part of the supplementary motor area (SMA) of an individual subject [Jennings et al., 1996b]. In this case, CARMA was run to confirm findings using the more commonly employed KS test at each pixel. Briefly, a subject was presented with visual stimuli and asked to perform a sequence of finger keypresses dependent on the stimuli. There were eight experimental and control condition epochs with roughly equal numbers of images in each

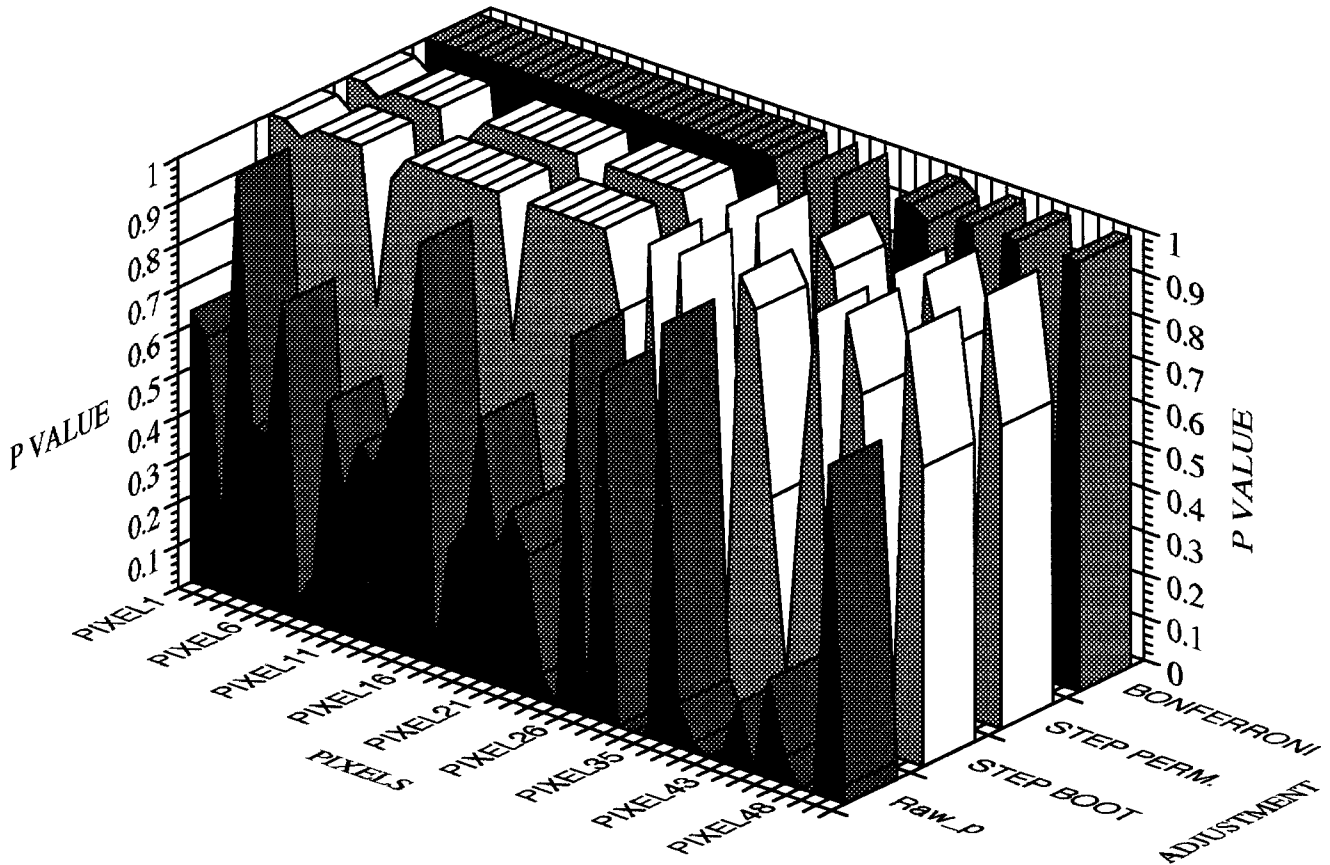


Figure 4.

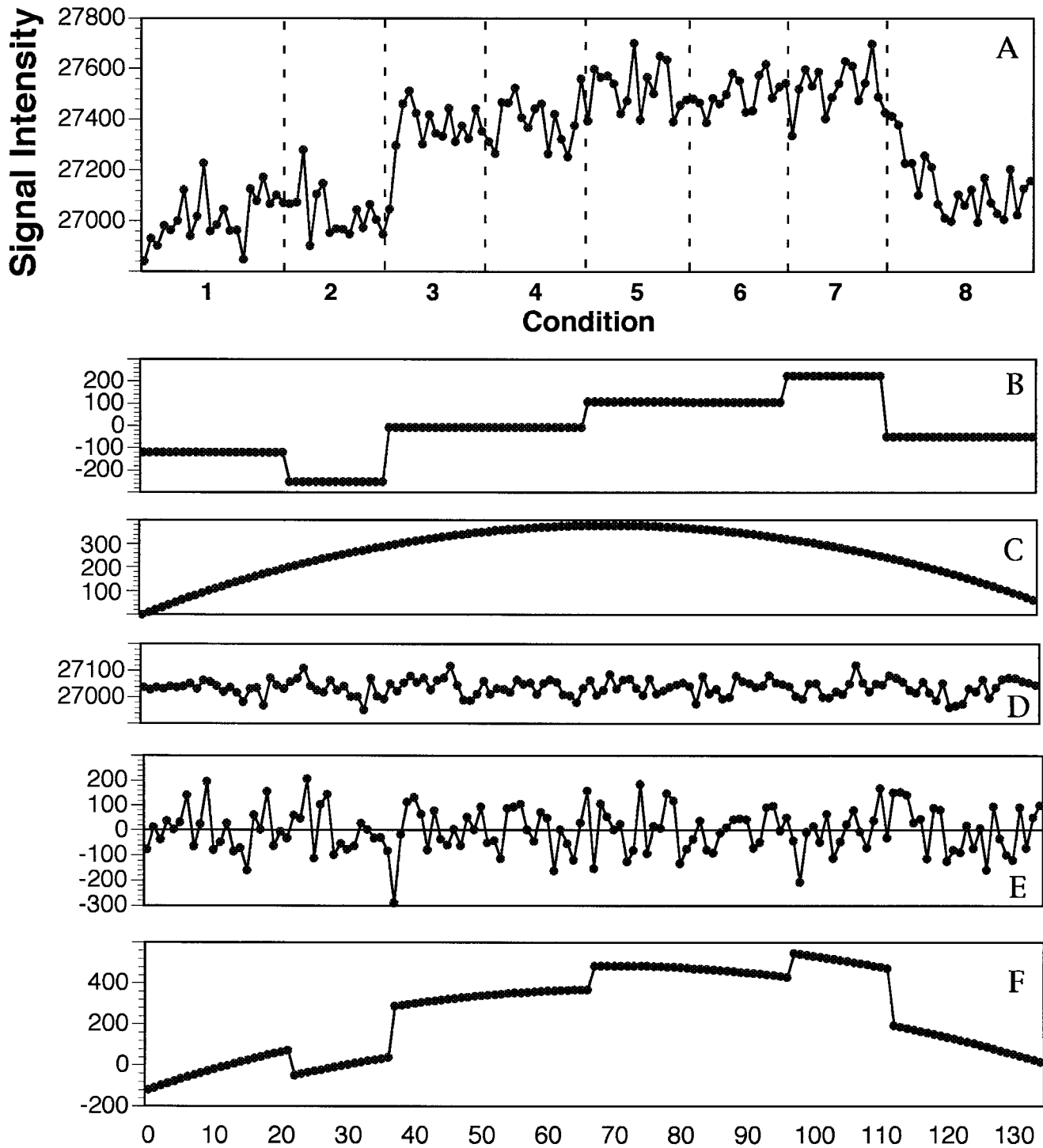
Example of adjustment of P values across pixels. Adjustment of P values from a two-tailed significance test (t-test) of the effect of an experimental condition vs. control on fMRI signal intensity at 50 pixels (135 timepoints; simulated data with varying strength of condition effects across pixels). Comparison of the step-down bootstrap and permutation resampling methods using 10,000 resamples, vs. unadjusted (liberal) P values and ordinary Bonferroni (conservative) adjustment. When P values were unadjusted, 17

pixels showed significant or marginally significant ($P = \text{approx. } .1$ or less) condition effects; there were seven after correction with the resampling methods and five after the Bonferroni adjustment. (Interpixel correlations were predominantly positive. For about $\frac{2}{3}$ of the pixels, the mean correlation was about zero with a maximum of about .2 to .3; for $\frac{1}{3}$, the mean correlation was about .25 with maximum .6 to .7.)

epoch. The experimental conditions examined learning and memory. Some epochs repeated an earlier condition in order to verify effects and help rule out temporal confounds.

Figure 5 displays the results of the CARMA analysis. The raw signal in A was separated into four component time series (in panels B, C, D, and E, respectively) that sum at each time point to the value at the corresponding time point in the raw series. Figure 5B shows the estimated step function effects of the eight experimental and control conditions. Various contrasts of interest for these conditions were statistically significant ($P < .05$; e.g., the mean signal value for conditions requiring memory was contrasted with the mean of nonmemory control conditions and found to be signifi-

cantly higher). Figure 5C displays a significant curvilinear (quadratic) temporal trend found in the data. The data in Figure 5A have already been corrected for subject motion; however, the correction is known to be imperfect, and the curvilinear trend may be wholly or partly residual motion artifact (which CARMA has removed, as it should). A composite of autocorrelative components is shown in Figure 5D. Figure 5E shows the noise series removed by the CARMA analysis, which was not significantly different from temporally independent white noise. We coded contrast indicators for the condition effects so that the effects centered at zero, whereas the temporal component was generally all positive (as it is here) or all negative. The intercept term is added into one of the panels (arbitrarily D), to



Images Across Time

Figure 5.

fMRI signal values for study of supplementary motor area (A) decomposed by CARMA into component time series due to task-related effects (B), temporal trend (C), autocorrelation (D), and white noise (E). F is the sum of B and C. (Intercept is added to D; average of 15 pixels, TR = 2,000 msec, 135 images/scan.)

bring the components up to the same absolute level as the raw series.

Because only one time series was analyzed, a relatively thorough assessment of autocorrelation was conducted. In this case, a significant lag 9 negative autoregressive component was found and removed, thus demonstrating that ARMA can detect some kinds of periodicities. A marginal positive first-order autoregressive effect was also removed. Figure 5F is the sum of the condition effects and the curvilinear time component, which together constitute the *deterministic* as opposed to the *stochastic* component (white noise and autocorrelation) of the raw data.

Testing two contrasts of conditions (motor behavior and the basal ganglia)

In this study of the basal ganglia, a young normal subject saw letters and performed keypress sequences dependent on the letter presented [Jennings et al., 1996a]. Six experimental and control condition epochs (four different conditions and repetitions of two) of equal length varied according to whether one or two letters were seen and whether keypresses were performed or not. Signals in a rectangular grid of 208 pixels covering the head of the caudate nucleus and neighboring areas were examined. Two contrasts among the conditions were of primary interest. One contrast compared the mean signal of conditions requiring a choice among keypresses with the mean of their control conditions (choice contrast). The other contrasted the mean of conditions requiring any kind of movement vs. that of those where there was no movement (motion contrast). CARMA and the resampling adjustment for multiple tests were employed to analyze the data (although, as a whole, interpixel correlations were not large for this example; mean $r = .024$; sig. > 0 , $P < .0001$; range, $-.48$ to $.76$).

Figure 6 shows the area of the brain slice examined and the results for the choice and motion contrasts. CARMA- and resampling-adjusted significant positive and negative effects are shown. For example, a positive choice contrast effect meant that the conditions requiring choice had a higher mean signal level than those not requiring choice, whereas a negative effect meant the nonchoice conditions had the higher mean. Although only a few pixels for each contrast are still significant after adjustment with the resampling method, this result should not be interpreted as meaning that effects occurred only within the spatial boundaries of those pixels. Rather, we feel it is reasonable to generalize effects to CARMA-adjusted significant pixels contiguous to the resampling adjusted significant

pixels, with the latter serving the role of verifying that the former are sensitive to what is real, and not just chance activation in that locus (see Discussion). Figure 7 shows a decomposition analogous to that of Figure 5, in this case for signals at one pixel, the one showing the significant positive choice effect and motion effect after the CARMA and resampling adjustment (row 28, column 61 in Fig. 6). For this pixel, the autocorrelative component was a significant, negative second-order autoregressive coefficient. In Figure 7, the estimated condition effects and curvilinear temporal trend are fairly opposite in shape and, therefore, hide each other in the raw series.

Testing a complex contrast (hippocampal activation during novel picture encoding)

In this study [Stern et al., 1996], subjects viewed a series of pictures while in the MRI scanner. There were two conditions, each lasting 1 minute: one in which 20 unique pictures were presented sequentially, and another in which the same picture was presented 20 times. In the condition with unique pictures, subjects were told to look at the pictures and remember them so that they could recognize them later. Activation in the hippocampal region related to picture encoding was of interest. The conditions were repeated once in an alternating sequence, so that there were four condition epochs including the repeated conditions.

We structured the analysis in terms of an ANOVA paradigm. Variability among the four conditions was parceled into a main effect of novelty (in which the two conditions with unique pictures were contrasted with the two conditions with a repeating picture), a main effect of repetition (in which the second set of unique and repeated picture conditions were contrasted with the first set), and an interaction of these two main effect factors. The interaction assessed whether the difference in mean signal level between the unique and repeated picture conditions was greater in the second repetition of these conditions than in the first, or vice versa. This interaction was not of primary substantive interest in the study, but it provided an exercise of the application of CARMA to testing an effect that is more complex than a simple contrast of two conditions.

Figure 8 displays the results of tests of the significance of the interaction at each pixel in a coronal brain slice for one subject. The analysis at each pixel was corrected by CARMA for linear and quadratic temporal trends and autocorrelation. A total of over 1,400 pixels was analyzed; about 5% did not meet the fit criteria of our CARMA algorithm and were excluded. A positive interaction denotes a situation in which the

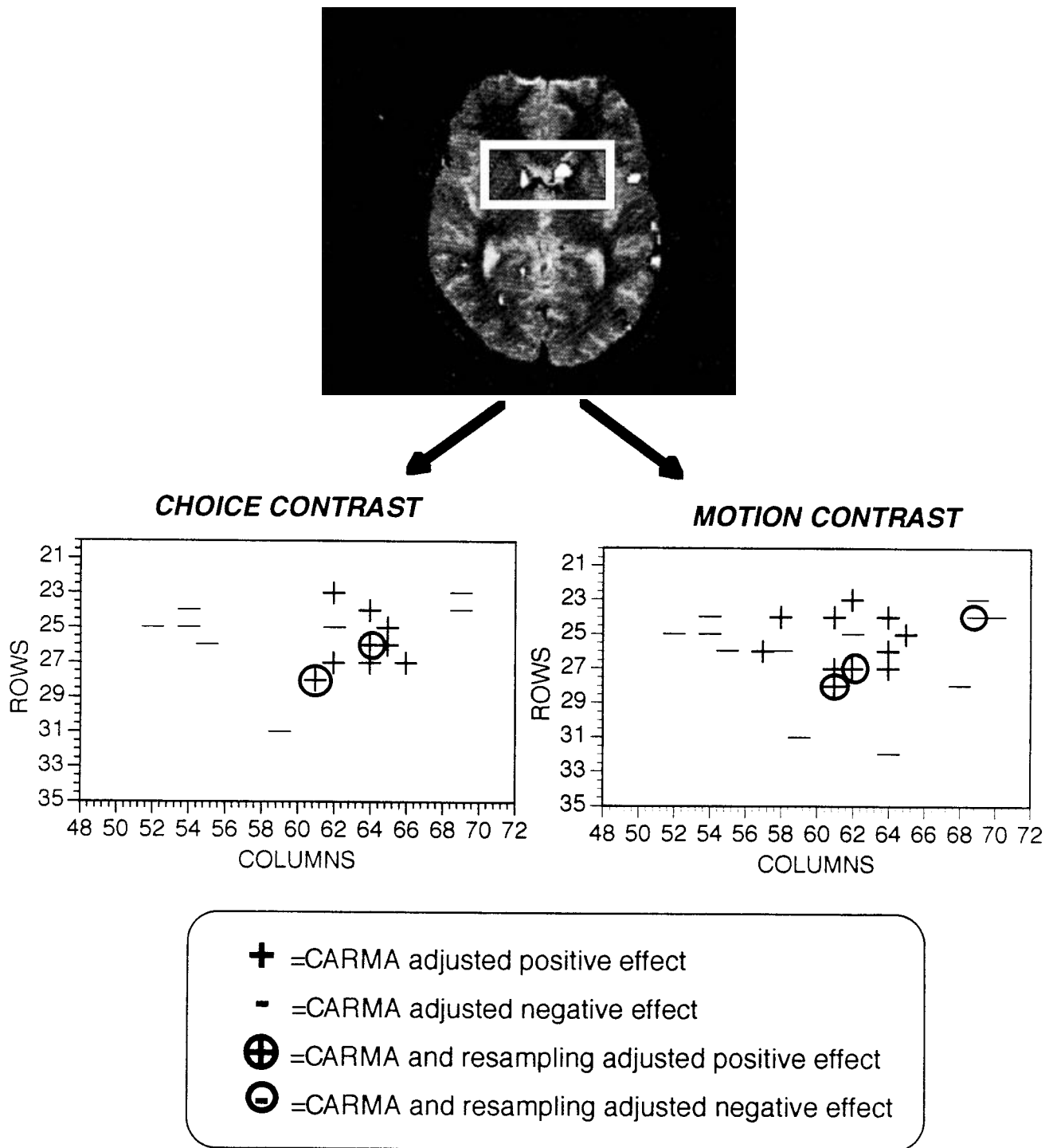
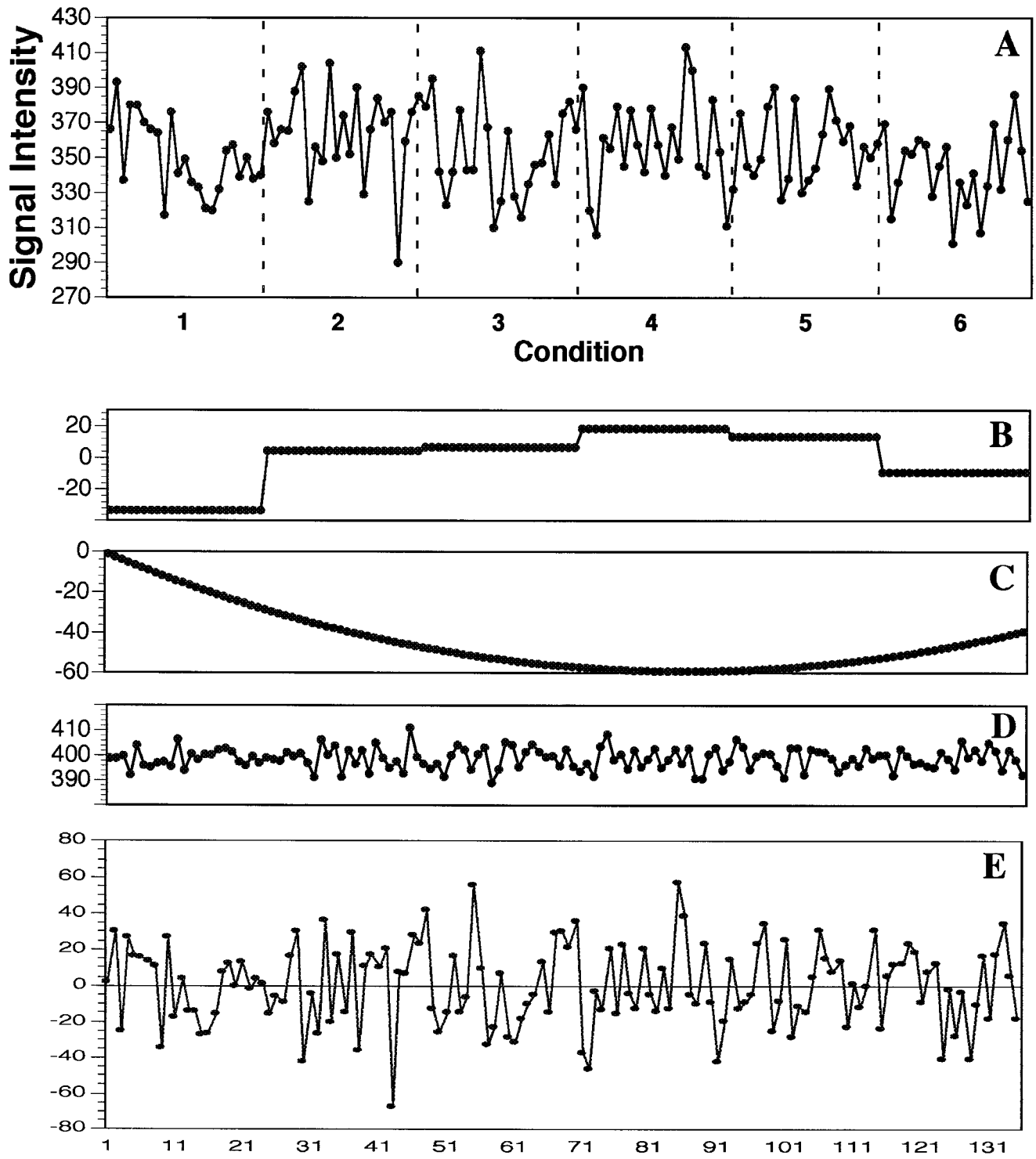


Figure 6.

Pixel locations of significant choice minus choice baseline contrast and motion vs. no motion contrast for study of basal ganglia. Rectangles in lower part of the figure correspond to the area of the horizontal brain slice bound by the rectangle in the brain image at top. (TE = 70 msec, TR = 1,750 msec, 135 images/scan.)



Images Across Time

Figure 7.

Separation of fMRI signal intensities (A) for the pixel corresponding to row 28 and column 61 in Figure 6 into component parts due to task-related effects (B), temporal trend (C), autocorrelation (D), and white noise (E); intercept is added to D.

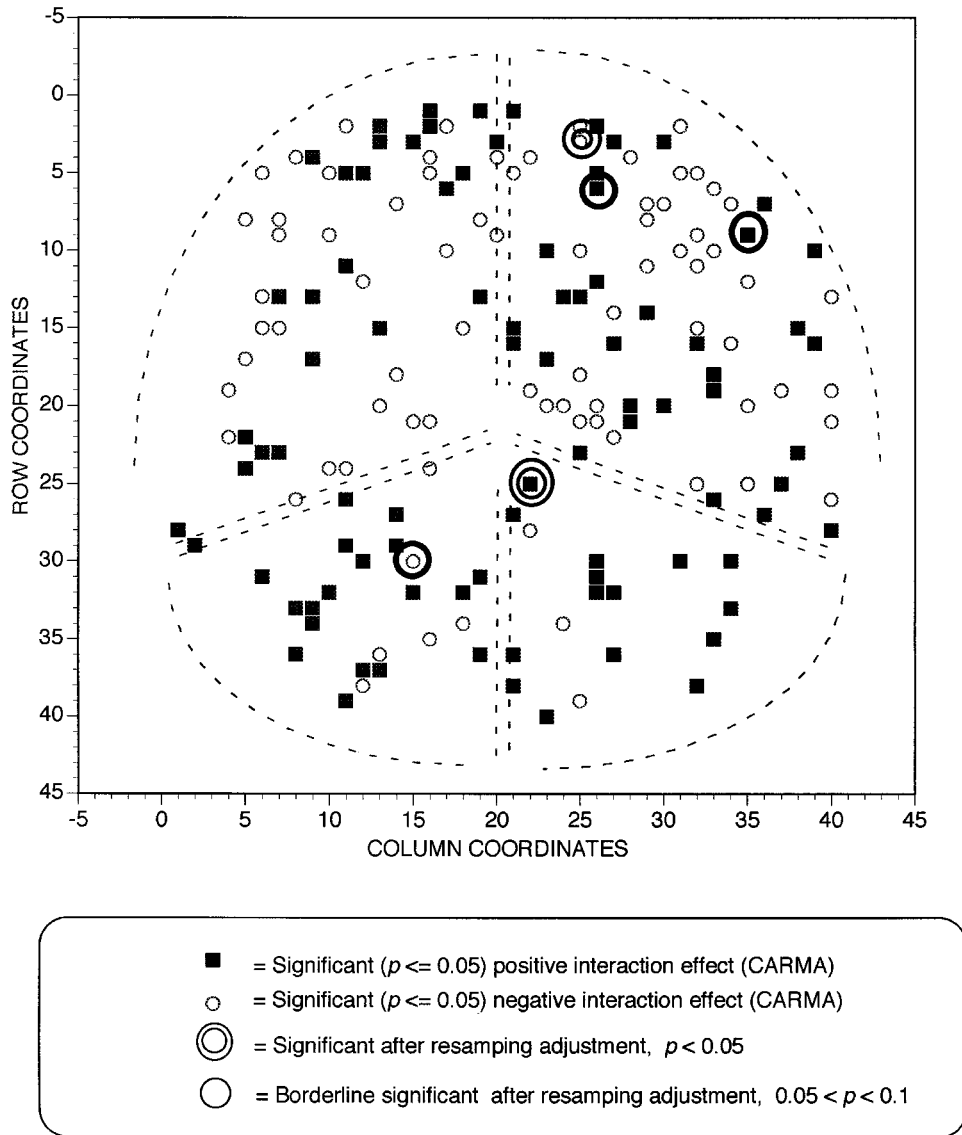


Figure 8.

Significant interactions of picture novelty and condition repetition in a study of hippocampal activation during novel picture encoding [Stern et al., 1996], for a coronal brain slice for one subject. (TE = 50 msec, TR = 2,500 msec, 96 images/scan.)

difference in mean for the novel picture condition minus that for the repeated picture condition is greater during the second than the first repetition of the conditions. A negative interaction denotes the converse. (There is a counterintuitive result in that some of the pixels found significant with the resampling methods are in the vicinity of clusters of pixels whose effects are of opposite sign. It is unclear yet whether this finding is due to “vascular steal,” i.e., a decrease in blood in a region neighboring an activated region

because of temporary displacement from one to the other, to local neuronal inhibition near the locus of excitation, or to an unresolved artifact of the data analysis method.) (Bruce R. Rosen, Nuclear Magnetic Resonance Center, MGH, personal communication, 1996).

As an aside, the data for this example were deliberately chosen because subject motion was known to have corrupted signal readings for many pixels near the bottom of the slice, resulting in some spurious KS

indexed activation. CARMA appears to have modeled most of these artifactual effects as linear and curvilinear time trends, and removed them (as it should).

SIMULATION STUDIES

Two important questions about the use of CARMA in the analysis of fMRI time series are, (a) Does CARMA correctly find components that are actually in a time series (sensitivity), and (b) does it *not* find components when none exists (specificity)? With regard to the second question, one may ask whether CARMA will pull apart phantom oppositional processes in what is really just flat white noise. Suppose one tests a time series model in which task-related effects and a quadratic time trend are both allowed for, but in reality neither is generating the series. Will CARMA tend to separate out what is essentially a level time trend into, e.g., a bell-shaped quadratic time trend combined with task-related effects that are lowest at the point of highest elevation in the quadratic trend, thus canceling each other out in the raw data? Or conversely, will it find a U-shaped quadratic trend combined with elevated task-related effects at the point of depression in the quadratic trend? (For example, are the effects in Figure 7B and C artifacts?)

In order to answer these questions, we performed simulation studies. To address the sensitivity question, we created a time series with simulated effects. The time series contained 96 signal values (for 96 sequential images) and was affected by step functions of four experimental/control conditions with equal numbers of images in each. The conditions were to be modeled with an ANOVA design as two main effects and an interaction. A quadratic time component was also added. The pattern of condition effects and the temporal trend were chosen to be oppositional and hide each other in the raw data. Last, randomly generated, normally distributed white noise with mean zero and constant variance was added, and first- and second-order positive autoregression components overlaid on it. All these condition-related, temporal-based, autocorrelative, and white noise components were summed into a composite time series. We constructed 50 such time series, each with the same components added, except that each had a new, independently generated set of white noise. For each of the 50 series (trials), the results of three analyses were compared: (1) an ordinary regression analysis that modeled the signal values against the experimental/control condition parameters only (Main Effect 1, Main Effect 2, and the Interaction), which corresponds to a simple t-test for each effect because each has 1 degree of freedom

(regression/condition or RC method); (2) an ordinary regression analysis that modeled the quadratic effect of time (the square of time as well as linear time) in addition to the condition effects, i.e., an analysis of covariance of condition effects adjusting for quadratic time trends (regression/condition/time or RCT); and (3) CARMA, which modeled the condition effects, quadratic time trend, and first- and second-order autoregression terms (using the SAS ARIMA Procedure with a conditional least-squares method of estimation [SAS/ETS Users' Guide, 1993]).

Figure 9 shows the results of the analyses. The estimated value for each parameter as well as its statistical significance is indicated for each of the 50 trials and for each of the analysis methods. The horizontal lines mark the pre-set actual values of the parameters. Only the interaction effect was set to zero (see figure legend for the other values). The RC method is biased for all parameters because it does not take into account the confounding effect of the quadratic time trend. The two methods that do take into account this time trend show fairly unbiased estimates for condition and time parameters; however, the variances of the estimates are less for the CARMA method than for RCT, i.e., CARMA is more efficient (in the statistical sense) than RCT. The greater spread of values for RCT causes it occasionally to show a statistically significant ($P < .05$) parameter estimate that is opposite in sign to the true value. The greater efficiency of CARMA is due to its recognition and removal of autocorrelation from the error variance. This phenomenon is especially evident in the graphs for estimates of the error variance. The percentage of significant parameters found by CARMA was 86% for Main Effect 1, 44% for Main Effect 2, 10% for the nonexistent interaction, 56% for the linear time coefficient, 20% for the smaller quadratic coefficient, 96% for the first-order autoregression coefficient, and 46% for the second-order coefficient. The RCT method found 40% of estimates for the null interaction to be significant, consistent with the known adverse effect of ignoring autocorrelation on statistical inference in ordinary regression analysis. For CARMA, the estimates of the autoregression coefficients appear to have a slight bias in underestimating the true values. The reason for this bias is unclear, and the effects of different methods of estimation need to be tested (e.g., maximum-likelihood vs. least squares, as well as different combinations of coefficients of different signs). Tests to determine whether residuals were significantly different ($P < .05$) from white noise were also run for the CARMA analyses. Residuals for only 14% of the trials were different from white noise, which is

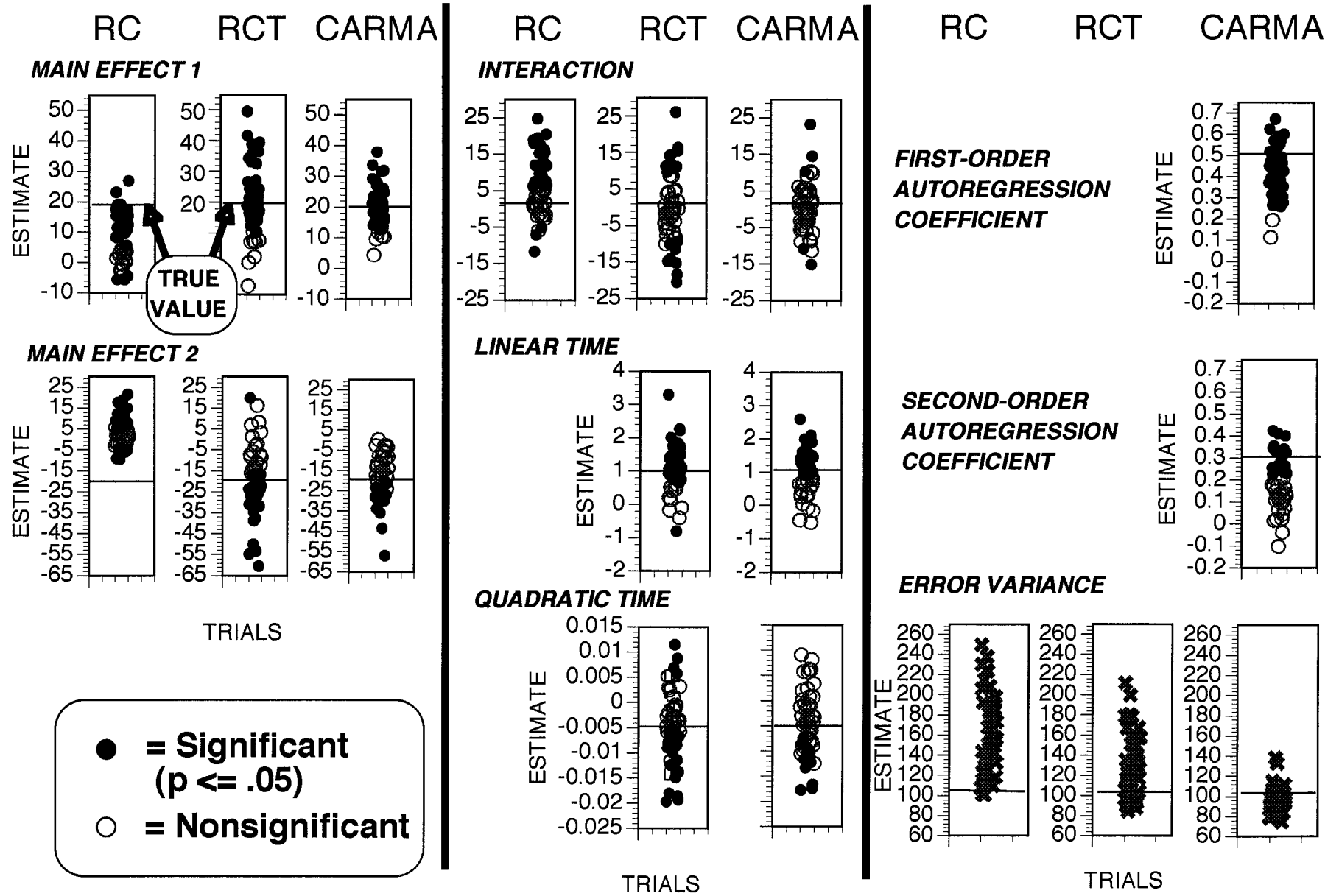


Figure 9.

Parameter estimates obtained from analyses of simulated time series data. Data created with the following parameter values: main effect 1 = 20; main effect 2 = -20; interaction = 0; quadratic time = -0.005; linear time = +1; first-order autoregression = .5; second-order autoregression = .3; white noise error variance = 100. Fifty

analyses (trials) done with (a) ordinary regression modeling experimental condition effects only (RC), (b) regression with condition effects plus time trend (RCT), and (c) CARMA modeling condition effects, time trend, and first- and second-order autoregression coefficients. Horizontal lines indicate true values.

roughly within range of what would be expected by chance. (The test result for each trial was the most significant of three tests for varying lags [up to 6, 12, 18] of autocorrelation. One would expect the three tests to be positively correlated so the chance probability is between .05 and .143; for independent tests, it is $1 - [0.95 \times 0.95 \times 0.95] = 0.143$.)

Additional simulation tests identical to that above, but with different combinations of presence and absence, or negative and positive condition effects, showed results parallel to those above.

In order to determine whether CARMA would extract spurious components from white noise, time series comprised of only white noise were analyzed by the three methods used above. We conducted 100 trial runs, each analyzing an independently sampled set of 96 temporally independent values chosen randomly from a normal distribution with a mean of zero. For each method, the estimates for each condition or time parameter correctly centered at zero with only between 1% and 8% significant results at the $P < .05$ level, about what would be expected by chance. CARMA showed no tendency to extract statistically significant phantom oppositional components. CARMA again showed a slight underestimate bias of the autoregression coefficients with 10% and 7% significant estimates (negative) for the first- and second-order coefficients, respectively.

Across all simulation tests, especially those for the white noise, patterns of correlations among parameter estimates did indicate a tendency for CARMA (and RCT) to find oppositional condition and time effects. For example, if an increasing time component was found, estimated condition effects tended to show an oppositional declining step function across time. As noted above, however, this tendency did not affect the sensitivity of CARMA to find components that were present and its specificity to not extract significant spurious components. In situations in which two effects were constructed to have the same sign but their estimates were negatively correlated, a fair proportion of significant findings for those effects occurred when they were simultaneously significant and had that same sign. A related point is that CARMA's power to detect condition effects that were correlated with time trends appeared to be lower than that for condition effects that were more orthogonal to time. For example, for the simulations corresponding to Figure 9, Main Effects 1 and 2 were of the same absolute size. Main Effect 1, however, which contrasted the first and third quarter of images with the remainder (a contrast relatively orthogonal to time) showed 86% of its estimates as significant, whereas Main Effect 2, which

contrasted the first half of images with the second half (more correlated with time) had only 44% significant estimates. Nevertheless, Main Effect 2 was "found" by CARMA 44% of the time, compared to 6% significant results from analysis of white noise. These findings suggest that like all statistical adjustments (e.g., partial regression coefficients), there are limitations to CARMA's ability to pull apart correlated effects.

There is need for more simulation testing of CARMA, with varying sets of effects and predictors, different experimental designs, autocorrelative components, time trends, different methods of estimation (e.g., maximum likelihood and least squares), different degrees of correlation of parameters, and nonorthogonality of predictors. Statistical power and bias in estimates need to be studied further. Also, more tests under null conditions, but with varying autocorrelative structures, should be conducted.

DISCUSSION

CARMA time series methods, coupled with resampling methods to adjust significance levels from multiple tests across pixels, hold promise as a useful method of assessing the statistical significance of task-related effects in fMRI studies. For the individual subject, the end result can be a two- or three-dimensional brain map of P values that indicates whether a given experimental condition changed MRI signal intensity at each pixel. These P values do not have to be treated as relative indices only of strength of effects, but can be taken to have absolute meaning, i.e., each estimates the true probability of obtaining, at that pixel, the observed effect by chance assuming the truth of the null hypothesis.

Variations of individual components of our proposed method have been presented recently by Bullmore et al. [1996] and Holmes et al. [1996]. Friston et al. [1994a, 1995a, 1995b] and Worsley and Friston [1995] have discussed related time series methods, and Friston et al. [1991, 1994b] have proposed methods for dealing with spatial extent of activation and multiple test problems in fMRI.

The methods of Friston et al. [1994a, 1995a] and Worsley and Friston [1995], which deal with temporal autocorrelation by adjusting degrees of freedom, produce less efficient estimates of parameters of interest than our method, assuming the validity of our ARMA models. We increase the likelihood of approximating the correct model for each pixel by our procedure of sequentially testing the significance of many AR and MA components of different orders for each pixel, requiring in each case that residuals pass a test that

they are not different from white noise, and allowing for a different model for each pixel. We increase power by including no ARMA terms for those minority of pixels for which no such terms are significant and residuals from the rest of the model pass the white noise test. Further, Friston et al.'s [1991, 1994b] spatial extent and multiple test methods have a number of assumptions, e.g., involving Gaussian fields (normal distributions), that can be questioned in some applications [Holmes et al., 1996], whereas our resampling methods are nonparametric techniques that are essentially assumption-free, although more computationally intensive. Friston et al [1995b] provide an excellent discussion of the utility of general linear models in functional imaging experiments, which would subsume as specific instances our use of condition contrasts, adjustment for time-related covariates, and interactions in factorial ANOVA designs. (Their discussion is, however, primarily oriented to less temporally intensive PET studies, rather than fMRI. Autocorrelation is additionally dealt with in Friston et al. [1994a, 1995a] and Worsley and Friston [1995].)

The approach of Bullmore et al. [1996] is similar to ours. They employ Box-Jenkins time series methods to remove temporal correlation, but they report the use of only a first-order AR model, which is applied uniformly to all pixels. We have flexibly tested AR and MA models of up to the third-order (and suggest amalgams of AR and MA, if necessary), with the model individualized for each pixel. In our tests so far, we have given priority to AR models over MA for the sake of simplicity when they both adequately model a time series, but we have found that only the MA models provide an acceptable fit for many pixels. The substantive underpinnings of the AR vs. MA distinction are still unclear, although it is known that high orders of one can approximate low orders of the other. In addition, Bullmore et al. present a more specialized application to a repeating on-off paradigm that may be most suited to sensory and motor studies, whereas we propose a more general approach independent of experimental design and substantive area of investigation. Last, Bullmore et al.'s proposed method for correcting multiple tests while taking into account spatial correlation structure requires extra data acquisition under null conditions, whereas our method does not.

Holmes et al. [1996] discuss nonparametric randomization tests for dealing with the multiple comparison problem in PET studies, but essentially the same method can be applied to fMRI data within a person, as we have done above. Randomization of "categories" across persons, is replaced by random reshuffling

of fMRI signal values across time to different nominal conditions within a person. Because of the large number of time points in fMRI data, we use a random reassignment algorithm, instead of working out all possible permutations. Holmes says results from this kind of algorithm "are still (almost) exact."

Finally, all the literature cited above is highly technical and may not be accessible to researchers with a background in conventional statistical methods only. Although technicality is by no means a criticism, the approach of these other reports contrasts with (and complements) our presentation targeting nonstatisticians as well as statisticians.

Software for doing CARMA is available in some of the large statistical packages [e.g., SAS/ETS Users' Guide, 1993; BMDP Statistical Software Manual, 1990a,b]. Also, SAS has a program for adjusting multiple significance tests with resampling methods [SAS Technical Report P-229, The Multtest Procedure, pp. 369-405, 1992], which we used for some of our tests. The Appendix below lists some software for ARMA and resampling methods. Analysis of fMRI data may require more efficient custom written algorithms, but subroutines for ARMA are available also (Appendix), and SAS provides a "macro" for bootstrap resampling. There is a need for software for time series analysis in the time domain for fMRI data specifically.

This report deals with the analysis of data from single subjects. Once CARMA and resampling methods have been run for multiple pixels for each of many subjects, however, it would be possible to conduct between-subject analyses using conventional multivariate approaches, where pixels in Talairach space are the multiple variables, and CARMA/resampling produced P (or t) values are the "scores" analyzed. The CARMA/resampling "scores" are valid effect indicators purged of confounding temporal trends, autocorrelation, and multiple test effects across pixels. Effectively, CARMA has distilled the relevant feature of the temporal dimension of the fMRI data into a static activation map for each subject that is now amenable to the same kinds of analyses applicable to less temporally intensive imaging data (e.g., from PET).

Specific examples of multivariate techniques would be: (a) a single sample multivariate t -test to determine whether activation occurs for a given group (using, e.g., t values from CARMA as the scores analyzed), (b) multivariate analysis of variance, discriminant analysis, or resampling methods to compare levels of activation among groups, or (c) multiple regression to relate activation to continuous variables, such as demographics or behavioral measures. These analyses, however, require that the number of variables (pixels here) be

less than the number of subjects (approximately, depending on the specific technique), which will not be the case when analyzing large brain areas. Here, a more moderate number of a priori delimited regions of interest (ROIs) could be used with the average of CARMA/resampling $P(t)$ values across pixels within each ROI for each subject as the dependent variable. These ROIs could be, for example a mosaic of anatomically defined areas of varying sizes and shapes, or a grid of uniform squares or cubes covering the relevant brain areas. CARMA can also be run on a time series that is the average of the series for all pixels within an anatomical ROI, i.e., the signal for each time point is the mean signal at that time point across all pixels in the ROI. Either way, each subject has one score for each ROI indexing activation at the respective ROIs for the subject. Multivariate between-subject tests would then be run with the ROIs as the multiple variables. Although the ROIs provide a low resolution analysis, these tests are only meant to provide inferential confirmation of the finer resolution original findings. The tests at the ROI-level merely rule out chance; that being done, the specific nature of effects are explored at the original pixel resolution.

In addition, raters can apply a consistent rule to classify each ROI in terms of whether it has been activated or not in each subject, based on statistics from the intrasubject analyses. (The raters would be blind to group status or other relevant between-subject variables). This dichotomy can then be used to compare groups with a chi-square test of association or Fisher's Exact Test, or else to assess relations with continuous variables with a t-test or logistic regression. The tests across ROIs could be corrected for multiple test chance effects with resampling methods. An ordinal variable indexing extent of activation could be used instead of the dichotomy with analyses appropriate to ordinal variables.

Other benefits of the methods presented in this report include the following:

1. Within reason, CARMA can disentangle confounded processes, e.g., condition effects and unrelated temporal trends. Any kind of processes can be searched for by including appropriate predictor variables. CARMA can pull apart effects with oppositional patterns, which hide each other in the raw data.
2. Brain maps showing the location of linear and curvilinear time trends and specific autocorrelative effects may provide useful information in their own right.
3. A number of different contrasts among experimental/control conditions can be assessed. Sometimes multiple contrasts of interest can be tested in the same run (if orthogonality among them is maintained). Assessment of the autocorrelative structure needs to be conducted only once, and this information can be re-used during subsequent analyses of contrasts. Complex contrasts can be tested; e.g., in the case of an ANOVA paradigm, a brain map of the significance of an interaction among conditions can be produced. Contrasts among repetitions of a behavioral condition can assess reliability of an effect or change in the effect (independent of extraneous temporal trends in the signal).
4. An objective, quantitative method of adjusting P values for multiple tests is employed. One cannot argue that a large cluster of pixels each showing strong activation coincident with a task is likely to have been produced by the task by virtue of the pixels being relatively contiguous and numerous. If, under null conditions, the signals in these pixels are relatively correlated, contiguity and cluster size may have a bearing on whether a real physiological event has occurred, but not on the likelihood of its being task produced. In adjusting P values, the resampling method takes into account the between-pixel correlational structure, and not necessarily that just tied to spatial proximity. Any complex, spatially varying patterns of interpixel correlations are accounted for, even negative correlations, should they exist. Given the possibility that a task could produce a network of spatially distributed activation, we feel that, in adjusting P values for multiple tests, the interpixel correlational structure is a more critical issue than contiguity and spatial extent of effects. More work is needed to determine whether and how correlational and spatial factors should be weighed in assessing task-produced activation.

Further, the resampling method for adjusting P values is a nonparametric test that is essentially assumption-free, and it provides "strong control" of type I error; i.e., it localizes significant activation and is not an omnibus test that only indicates activation is occurring "somewhere" [Holmes et al., 1996]. Incidentally, to avoid parametric test assumptions of CARMA (see below), we could have coupled CARMA with a resampling method *within* each pixel to obtain the original P values. However, this would increase computational time substantially with little benefit; we are not interpreting the original P values inferentially but are

only using them as indices of strength of activation to be submitted to the between-pixel resampling stage to get the true inferential P values. (Conceivably, statistics from CARMA other than the P could be used as the original indices of activation strength if they correct for or are not affected by confounding temporal trends and autocorrelation. Two-tailed effects would have to be taken into account in the resampling algorithm as needed.)

5. CARMA is fairly independent of experimental design; within reason, data from any design can be analyzed post hoc. Thus, researchers can devise study paradigms based primarily on substantive issues and practical considerations, and not be constricted by the needs of data analysis. For example, designs consisting of only a few lengthy conditions, such as may be appropriate for cognitive studies, can be analyzed, and not just multiple on/off paradigms as are more typical in sensory and motor studies. Theoretically, a design with only one experimental condition epoch and one control epoch can be analyzed by CARMA, although at least one repetition of each is recommended (see below). Repeated on/off paradigms may be best analyzed by using a numeric indicator for the conditions and examining effects at various time lags via cross-correlograms [McCleary and Hay, 1980; see also “dynamic regression models” in Pankratz, 1991] or with the methods of Bullmore et al. [1996].
6. Task-related effects more complex than simple step functions can be assessed, e.g., gradual effects, abrupt temporary activation, transient pulses. CARMA is flexible in modeling the temporal nature of condition effects. Hemodynamically mediated delays between the onset of a task or stimulus condition and change in fMRI signal can be explored and modeled.
7. For a single time series, e.g., after averaging, at each time point, signals across pixels in an ROI (see SMA example above), a thorough and in depth CARMA analysis can be conducted by itself (e.g., employing correlograms). Adjustments for multiple significance tests with resampling methods or other techniques are not applicable. A CARMA test on an averaged time series for an ROI is likely to have more power than that for an individual pixel. In averaging across pixels, noise would tend to cancel out, whereas task-related effects may not (increased signal-to-noise ratio).

8. CARMA can be used to investigate the relation of a changing continuous, numeric exogenous variable (e.g., stimuli varying in intensity) to signal intensity in an fMRI time series. The time lag between change in the exogenous variable and signal change can be assessed with cross-correlograms [McCleary and Hay, 1980; Pankratz, 1991].

Notwithstanding these desirable features of our methods, some cautions and reservations should be noted:

1. The power of the resampling method needs further study. We have often found only a small percentage of pixels showing significant task-related effects after the adjustment with the resampling methods. Stepdown methods, however, increase power. Also, power could be augmented at the CARMA stage by filtering out extraneous variability on the basis of preliminary spectral analysis, or by including in the CARMA model additional predictors known to account for nontrivial variance. Perhaps nonstochastic temporal components that are not significant might be dropped (but see “stationarity” below). Further, although individual pixels are judged to show statistically significant task-related effects only on the basis of the results of the resampling adjustment, in interpreting these results, activated regions do not have to be considered restricted strictly to the boundaries of those pixels. It is reasonable to generalize effects to CARMA-adjusted significant pixels contiguous to the resampling-adjusted significant pixels because the latter can be considered as serving the role of verifying that the former are being sensitive to what is real, and not just chance activation in that locus. This limited extension of what regions are considered activated also seems sensible given that images of signal effects are not perfectly reflective of neuronal activation. Because resolution is blurred by hemodynamics, spatial smoothing, and subject movement, a pixel should be considered only representative of a locality. Another point related to the power issue is that it is unclear which data should be subjected to the resampling algorithm. We used the white noise residuals from CARMA, but perhaps the estimated linear and quadratic temporal effects should also be retained in the data during resampling to provide a better baseline measure of the between-pixel correlational structure against which to judge the significance of task-related

effects. Within-pixel nonstochastic temporal trends sometimes occurred in clusters spatially. Even though task effects are adjusted for extraneous temporal trends in the original CARMA tests, whatever is the basis of a set of pixels exhibiting the same time trend in unison may cause a common spurious task effect among them under the null hypothesis. Our example analyses may have underestimated the appropriate interpixel correlational structure with resultant loss of power. Our resampling-adjusted P values were not as much an improvement over the Bonferroni adjusted as we had hoped. More work is needed to establish what the appropriate data should be for resampling.

2. Like other statistical adjustments, there is a limit to how much confounding CARMA can pull apart (the general problem of multicollinearity [Cohen and Cohen, 1983]). Temporal and task-related effects may be too correlated in some instances. Our simulation studies above confirm the theoretical expectation that power is inversely related to the extent of such a correlation. In designing fMRI studies, investigators should make experimental conditions as orthogonal to time as is reasonably possible and repeat each at least once. In Figure 1, without repetition of the two conditions, it is less certain whether the sharper decline in the control condition is caused by that condition, or is part of a decelerating trend independent of conditions. The distinction has implications in terms of interpreting task effects. CARMA may have low power to disentangle aliased cardiac and respiratory effects that are highly correlated with task-related effects. However, a study can be designed so that experimental conditions have a pattern that is less likely to coincide with known extraneous cyclical phenomena. Extraneous periodicities that are temporally smoother than condition effects could cause pixels to be removed for not passing the white noise test or could be picked up by high-order polynomial functions of time or high-order ARMA. They could also be assessed with ARMA, spectral analyses, time functions, or sinusoidal modeling during preliminary dry runs and then statistically removed from the data obtained during the experiment proper. Nevertheless, even though an attractive feature of CARMA is its allowance for flexibility in study design, some good sense and caution must still be exercised in this regard.
3. A reasonably large number of images (time points) is necessary for CARMA to estimate autocorrela-

tion and temporal effects reliably, and a reasonably small ratio of unique behavioral conditions to time points is needed to estimate effects of these conditions. A minimum of 50 total time points is recommended by McCleary and Hay [1980] for ARMA analyses in general (SAS suggests at least 30 [SAS/ETS User's Guide, Version 6, 1993, p. 126]), but for fMRI studies when there are multiple experimental conditions, approximately 100 would seem desirable and feasible.

4. A technical issue concerns ARMA assumptions, especially stationarity of the mean and the problem of autocorrelative components being outside the bounds of stationarity or invertibility [McCleary and Hay, 1980]. We always model linear and curvilinear time trends, which should reduce or eliminate these problems, but further research is needed regarding the advantages and disadvantages of modeling temporal trends versus differencing. Further, significance tests for CARMA assume that residuals are normally distributed, and ideally this assumption should be checked; however, parametric tests are fairly robust to violations of normality [Myers, 1979]. Also, as noted above, P values from the CARMA stage of analysis are not interpreted as probabilities, but only as indices of strength of activation (blind to the multiplicity of tests and interpixel correlations); they have to be transformed by the resampling method before they are considered values with absolute inferential meaning. Thus, assumptions of significance tests for CARMA are less critical than they typically would be for parametric tests. Stationarity of variance across time is also an assumption of ARMA, and transformations of the data may be required to obtain homogeneity of variance throughout each time course (homogeneity *across* pixels is not necessary). One could argue that the KS method is sensitive to more kinds of effects than CARMA because changes in variance as well as mean will affect KS indices. The KS, WRS, and t -tests may be useful as quick, preliminary screens, but their assumption of independence of observations precludes them as accurate inferential tools for fMRI. Furthermore, there are tests for changes in variance and correction for it in time domain analyses [Cromwell et al., 1994a; SAS/ETS User's Guide, the Autoreg Procedure, 1993, pp. 183–253].
5. ARMA does not deal with periodicities in time series as well as do methods in the frequency domain, e.g., spectral analysis [Gottman, 1981]. Periodicities can be detected by ARMA with

high-order terms (see SMA example above), but it may not be feasible to include such tests in algorithms for analyses of many pixels. High-order ARMA components may also be confounded with task effects, especially for rapidly alternating conditions. Preliminary assessment and removal of extraneous cyclical components with spectral methods or sinusoidal modeling may be a possibility [but see “seasonal ARMA” in McCleary and Hay, 1980; McDowall et al., 1980; Pankratz, 1983, 1991].

6. The methods presented in this report, like many other fMRI data analysis techniques, are restricted to looking at the brain as comprised of separate spatial modules, each of which does or does not become activated by various stimuli or tasks. In addition to this approach, analyses that assess networks of activation need to be further developed. Such methods are being employed in PET studies [McIntosh and Gonzalez-Lima, 1994]. Also, instead of a strategy that assesses significance of task effects at each pixel individually and then subsequently corrects for the multitude of tests across pixels, a more elegant approach to explore is to assess the within-pixel temporal effects and between-pixel spatial correlations simultaneously with multivariate time series methods [Cromwell et al., 1994b].
7. Our methods are computationally intensive, but not prohibitively so. The SAS programs for the novel picture encoding example above (Fig. 8) took approximately 1 day to run on a VAX Model 4300 computer with 32 megabytes of memory; more efficient programs can almost certainly be written.

We encourage more simulation and real data tests of the methods we present here in further evaluating their usefulness in assessing task-related activation in the brain in fMRI studies.

ACKNOWLEDGMENTS

We are grateful to Elkan Halpern of the Center for Imaging and Pharmaceutical Research, Massachusetts General Hospital (MGH), Charlestown, MA, and Robert Weiskoff, David Kennedy, Bruce R. Rosen, Robert Savoy, and other researchers associated with the MGH Nuclear Magnetic Resonance (NMR) Center in Charlestown, MA, for valuable comments and suggestions. We also thank Chantal Stern and colleagues at the MGH-NMR Center for permitting us to use their

data in testing our analysis methods. We take responsibility for any errors in analysis or interpretation of those data. This study was supported by NIH grants AG06605, AG05134, and RR00088.

REFERENCES

- Aldenderfer MS, Blashfield RK (1984): Cluster Analysis. Quantitative Applications in the Social Sciences Series, No. 44. Thousand Oaks, CA: Sage Publications, Inc.
- Bandettini PA, Jesmanowicz A, Wong EC, Hyde JS (1993): Processing strategies for time-course data sets in functional MRI of the human brain. *Magn Reson Med* 30:161–173.
- Binder JR, Rao SM (1994): Human brain mapping with functional magnetic resonance imaging. In: Kertesz A (ed): *Localization and Neuroimaging in Neuropsychology*. Orlando, FL: Academic Press, pp 185–212.
- BMDP Statistical Software Manual. Volume 1. (1990a): The 2T Program: Box-Jenkins Time Series Analysis. Berkeley, CA: University of California Press, pp 435–488.
- BMDP Statistical Software Manual. Volume 2. (1990b): The 1T Program: Univariate and Bivariate Spectral Analysis. Berkeley, CA: University of California Press, pp 1079–1120.
- Box GEP, Jenkins GM (1976): *Time Series Analysis: Forecasting and Control*. San Francisco: Holden-Day.
- Box GEP, Tiao GC (1975): Intervention analysis with application to economic and environmental problems. *J Am Stat Assoc* 70: 70–79.
- Box GEP, Jenkins GM, Reinsel GC (1994): *Time Series Analysis: Forecasting and Control 3rd Ed*. Englewood Cliffs, NJ: Prentice-Hall, Inc.
- Bullmore E, Brammer M, Williams SCR, Rabe-Hesketh S, Janot N, David A, Mellers J, Howard R, Sham P (1996): Statistical methods of estimation and inference for functional MR image analysis. *Magn Reson Med* 35(2):261–277.
- Cohen J, Cohen P (1983): *Applied Multiple Regression/Correlation Analysis for the Behavioral Sciences*, 2nd Ed. Hillsdale, NJ: Lawrence Erlbaum and Associates.
- Cromwell JB, Labys WC, Terraza M (1994a): Univariate tests for time series models. *Quantitative Applications in the Social Sciences Series*, No. 99. Thousand Oaks, CA: Sage Publications, Inc.
- Cromwell JB, Hannan MJ, Labys WC, Terraza M (1994b): Multivariate tests for time series models. *Quantitative Applications in the Social Sciences Series*, No. 100. Thousand Oaks, CA: Sage Publications, Inc.
- DeYoe EA, Bandettini P, Neitz J, Miller D, Winans P (1994): Functional magnetic resonance imaging (fMRI) of the human brain. *J Neurosci Methods* 54:171–187.
- Disbrow E, Buonocore M, Antognini J, Carstens E, Shumway R (1995): Time series analysis: an alternative method for processing fMRI data. Abstract presented to the Cognitive Neuroscience Society, San Francisco.
- Efron B (1982): *The Jackknife, the Bootstrap and Other Resampling Plans*. Philadelphia, PA: Society for Industrial and Applied Mathematics.
- Friston KJ, Frith CD, Liddle PF, Frackowiak RSJ (1991): Comparing functional (PET) images: the assessment of significant change. *J Cereb Blood Flow Metab* 11:690–699.
- Friston KJ, Jezzard P, Turner R (1994a): The analysis of functional MRI time-series. *Hum Brain Mapping* 1:153–171.

- Friston KJ, Worsley KJ, Frackowiak RSJ, Mazziotta JC, Evans AC (1994b): Assessing the significance of focal activations using their spatial extent. *Hum Brain Mapping* 1:210–220.
- Friston KJ, Holmes AP, Poline J-B, Grasby PJ, Williams SCR, Frackowiak RSJ, Turner R (1995a): Analysis of fMRI time-series revisited. *Neuroimage* 2:45–53.
- Friston KJ, Holmes AP, Worsley KJ, Poline J-P, Frith CD, Frackowiak RSJ (1995b): Statistical parametric maps in functional imaging: a general linear approach. *Hum Brain Mapping* 2:189–210.
- Good P (1994): *Permutation Tests: A Practical Guide to Resampling Methods for Testing Hypotheses*, New York, NY: Springer-Verlag, Inc.
- Gottman JM (1981): *Time-Series Analysis: A Comprehensive Introduction for Social Scientists*, New York, NY: Cambridge University Press.
- Hochberg Y, Tamhane AC (1987): *Multiple Comparisons Procedures*, New York: John Wiley & Sons, Inc.
- Holmes AP, Blair RC, Watson JDG, Ford I (1995): Non-parametric analysis of statistic images from functional mapping experiments. *NeuroImage* 2:S72.
- Holmes AP, Blair RC, Watson JDG, Ford I (1996): Non-parametric analysis of statistic images from functional mapping experiments. *J Cereb Blood Flow Metab* 16:7–22.
- Jennings PJ, Stern CE, Kwong KK, Locascio JJ, Corkin S, Rosen BR, Gonzalez RG (1996a): Basal ganglia function in a sequential movement task. Poster presentation at Second International Conference on Functional Mapping of the Human Brain, Boston, MA, June 17–21. *NeuroImage* 3 Part 2:S385.
- Jennings PJ, Stern CE, Kwong KK, Sugiura RM, Guimaraes AR, Gonzalez RG, Corkin S, Rosen BR (1996b): Supplementary motor area activity in sequential movements from memory: evidence from functional MRI. Poster presentation at Massachusetts Alzheimer's Disease Research Center Ninth Annual Scientific Poster Session, Massachusetts General Hospital, Boston, MA.
- Kim J, Trivedi PK (1994): Econometric time series analysis software: a review. *Am Stat* 48:336–346.
- Kim S-G, Ashe J, Georgopoulos AP, Merkle H, Ellermann JM, Menon RS, Ogawa S, Ugurbil K (1993a): Functional imaging of human motor cortex at high magnetic field. *J Neurophysiol* 69:297–302.
- Kim S-G, Ashe J, Hendrich K, Ellermann JM, Merkle H, Ugurbil K, Georgopoulos AP (1993b): Functional magnetic resonance imaging of motor cortex: hemispheric asymmetry and handedness. *Science* 261:615–617.
- Kim S-G, Ugurbil K, Strick PL (1994): Activation of a cerebellar output nucleus during cognitive processing. *Science* 265:949–951.
- Le Bihan D, Karni A (1995): Applications of magnetic resonance imaging to the study of human brain function. *Curr Opin Neurobiol* 5:231–237.
- Ljung GM, Box GEP (1978): On a measure of lack of fit in time series models. *Biometrika* 65:297–303.
- Marquardt DW (1963): An algorithm for least squares estimation of nonlinear parameters. *J Soc Industrial Appl Mathematics* 2:431–441.
- McCleary R, Hay RA (1980): *Applied Time Series Analysis for the Social Sciences*, Thousand Oaks, CA: Sage Publications, Inc.
- McDowall D, McCleary R, Meidinger EE, Hay RA (1980): Interrupted time series analysis. *Quantitative Applications in the Social Sciences Series*, No. 21. Thousand Oaks, CA: Sage Publications, Inc.
- McIntosh AR, Gonzalez-Lima F (1994): Structural equation modeling and its application to network analysis in functional brain imaging. *Hum Brain Mapping* 2:2–22.
- Mooney CZ, Duval RD (1993): *Bootstrapping: a nonparametric approach to statistical inference. Quantitative Applications in the Social Sciences Series*, No. 95. Thousand Oaks, CA: Sage Publications, Inc.
- Myers JL (1979): *Fundamentals of Experimental Design*, 3rd Ed. Boston: Allyn and Bacon, Inc.
- Pankratz A (1983): *Forecasting With Univariate Box-Jenkins Models: Concepts and Cases*. New York: John Wiley & Sons, Inc.
- Pankratz A (1991): *Forecasting with Dynamic Regression Models*, New York: John Wiley & Sons, Inc.
- Resampling Stats (1996): 612 N. Jackson St., Arlington, VA 22201. (703) 522-2713. Internet: resample@cais.com.
- SAS/ETS (Econometric Time Series) User's Guide: Version 6, Second Edition. (1993): The Arima Procedure. pp 99–182. The Autoreg Procedure, pp 183–253. The Spectra Procedure. pp 749–770. Cary, NC: SAS Institute Inc.
- SAS/STAT User's Guide (1990): Vol 1 and 2. Version 6, Fourth Edition. Cary, NC: SAS Institute Inc.
- SAS Technical Report P-229. SAS/STAT Software: Changes and Enhancements. Release 6.07. (1992): The Multtest Procedure. pp 369–405. Cary, NC: SAS Institute Inc.
- Sidak Z (1967): Rectangular confidence regions for the means of multivariate normal distributions. *J Am Stat Assoc* 62:626–633.
- Siegel S, Castellan NJ (1988): *Nonparametric Statistics for the Behavioral Sciences*, 2nd Ed. New York, NY: McGraw-Hill, Inc.
- Stern CE, Corkin S, Gonzalez RG, Guimaraes AR, Baker JR, Jennings PJ, Carr CA, Sugiura RM, Vedantham V, Rosen BR (1996): The hippocampal formation participates in novel picture encoding: evidence from functional magnetic resonance imaging. *Proc Natl Acad Sci USA* 93:8660–8665.
- Tagaris GA, Kim S-G, Ugurbil K, Georgopoulos AP (1995): Box-Jenkins Intervention Analysis of Functional MRI Data. Presentation at Society of Magnetic Resonance, Third Scientific Meeting, and European Society for Magnetic Resonance in Medicine and Biology, 12th Annual Meeting, Nice, France, August 19–25.
- Toothaker LE (1993): *Multiple Comparison Procedures. Quantitative Applications in the Social Sciences Series*, No. 89. Thousand Oaks, CA: Sage Publications, Inc.
- Tyszka JM, Grafton ST, Chew W, Woods RP, Colletti PM (1994): Parceling of mesial frontal motor areas during ideation and movement using functional magnetic resonance imaging at 1.5 Tesla. *Ann Neurol* 35:746–749.
- Westfall PH, Young SS (1993): *Resampling-Based Multiple Testing: Examples and Methods for p-Value Adjustment*. New York, NY: John Wiley & Sons, Inc.
- Worsley KJ, Friston KJ (1995): Analysis of fMRI time-series revisited—again. *NeuroImage* 2:173–181.

APPENDIX A

Software with ARMA time series programs or subroutines

BMDP, BMDP Statistical Software, Inc., 1440 Sepulveda Blvd., Suite 316, Los Angeles, CA 90025
 IMSL, International Mathematical and Statistical Libraries, Inc., 7500 Bellaire Blvd., Houston, TX 77036
 MicroTSP (Micro Time Series Processor), Quantitative Micro Software, 4521 Campus Drive, Suite 336, Irvine, CA 92715

MINITAB, Minitab, Inc., 3081 Enterprise Drive, State College, PA 16801-3008

PACK, Automatic Forecasting Systems, PO Box 563, Hatboro, PA 19040

RATS (Regression Analysis of Time Series) Estima, P.O. Box 1818, Evanston, IL, 60204-1818

SAS (Econometric Time Series Package: The ARIMA Procedure), SAS Institute, Inc., SAS Campus Drive, Cary, NC 27513

SHAZAM, UBC Economics, No. 997-1873 East Mall, Vancouver, BC, V6T-1Z1, Canada

S-PLUS, StatSci Division, MathSoft, Inc., 1700 Westlake Avenue North, Suite 500, Seattle, WA 98109-9891

SPSS, SPSS, Inc., 444 N. Michigan Ave., Chicago, IL 60611

TSP (Time Series Processor), TSP International, P.O. Box 61015, Station A, Palo Alto, CA 94306

Resampling software

SAS (The Multtest Procedure), SAS Institute, Inc., SAS Campus Drive, Cary, NC 27513 (See SAS Technical Report P-229, SAS/STAT Software: Changes and Enhancements, Release 6.07. (1992), pp. 369-405)

Resampling Stats Software, Resampling Stats, 612 N. Jackson St., Arlington, VA 22201